ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

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



THE UNIVERSITY OF CHICAGO

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

The New Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 2

Knowledge in Depth

FOUNDED 1768
15TH EDITION

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

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

Printed in U.S.A.

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

Arizona

A classic symbol of America's Old West, Arizona is rich in legends of the Indian leaders Geronimo and Cochise and of such towns as Tombstone, where the feud between the Earp and Clanton brothers erupted into the West's most famous shoot-out, the gunfight at the O.K. Corral. The Arizona of the 1970s, with its modern cities, sprawling suburban developments, and nationwide transportation arteries, shows quite a different face, though in its broad expanses of sparsely settled country, its extraordinarily coloured rock walls and mountains rising above dramatic desert plains, and its mud adobe structures and baroque Spanish missions is preserved much of the natural and human landscape that characterized its frontier days.

During many decades of territorial status, Arizona developed strong regional ties with neighbouring California on the west and lesser but important exchanges with Mormon Utah on the north and Mexican Sonora on the south. The long-time bonds with Sonora give Mexicans distinctive status in Arizona. Arizona was the last of the contiguous 48 states to enter the Union, in 1912, and, though for long it was little more than a stretch of country that had to be crossed to get to somewhere else, it has become a stopping place as well. For many years it has been one of the tourist meccas of the nation and one of the fastest growing of the states.

Internal contradictions

Arizona is a deceptive state, even beyond its more obvious contrasts between sophisticated urban life and open ranges with cattle roaming across roadways. It has long been touted as one of the healthiest and most pollution-free areas of the world; yet for its poor, including its more than 90,000 Indians, the rate of infant mortality and of such diseases as tuberculosis and diphtheria is well above the national average. Industries and automobiles and other phenomena of modern living have begun to cloud the once clear skies of the state, especially around Tucson and Phoenix, the capital and largest city in the tier of Mountain States. Irrigated vegetable and cotton farms spread greenery from one horizon to another, but moisture for the arid land remains a critical need. This problem has led to disputes with California over the waters of the Colorado River and to continual draining of underground reserves that were built up over thousands of years. Arizona is attempting, like Israel, to build a modern industrial society in a dry land and, like Alaska, to conserve as much as possible pristine qualities that could so easily be sacrificed in the process. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; ROCKY MOUNTAINS; NORTH AMERICAN DESERT; GRAND CANYON; BASIN AND RANGE PROVINCE; and NORTH AMERICA.)

THE HISTORY OF ARIZONA

The history of the Southwest, both before and after the arrival of the European, paid little heed to boundaries, which, whether state or national, were often vague.

Aboriginal civilizations

Early inhabitants. Human settlement existed throughout the area from about 25,000 BC, whereas the nomadic Apache and Navaho Indians arrived within only a few centuries of the Spanish. These invaders arrived in large numbers after the collapse of the Anazazi and Hohokam civilizations. Cliff dwellings of the Anazazi and remnants of the elaborate Hohokam irrigation systems dot the northern and central sections of present-day Arizona. The dryness has helped preserve artifacts, even occasional bodies.

Spanish territory. Early Spanish exploration was concerned with finding the fabled seven golden cities, and settlement was mainly for mission work well into the 18th century. The entire area was called Nuevo Méjico, and most activity was in present-day New Mexico. Tucson was settled under Spain in 1776 and was the only post in Arizona that remained under the flag of Mexico after Apache warfare virtually emptied the land by the 1840s.

American territory and state. After the Mexican War, Arizona was ceded as part of New Mexico to the United States in 1848, and the Gadsden Purchase, an area south of the Gila River, was added in 1853. California gold brought adventurers and miners into the area in the 1850s, but the Civil War and troop withdrawal delayed settlement and allowed the Apache continued sway. In the 1870s and 1880s ranches and towns finally spread into former Indian lands, though a few lawless whites and Indians kept the bloody frontier tradition alive throughout the latter decade.

Following attainment of statehood in 1912, Arizona became the site of labour disputes involving copper-mine owners and the Industrial Workers of the World (IWW), culminating in 1917 in a deportation of suspected agitators that had repercussions for decades. Throughout the 20th century the state was in constant dispute with California over use of the waters of the Colorado River—the boundary of the two states. The 1963 Supreme Court decision in the main upheld Arizona's claims and has ushered in a new era of greater cooperation among all states of the lower Colorado River Basin.

THE NATURAL AND HUMAN LANDSCAPE

Next to climate, the beauty and variety of Arizona's landscape constitute its greatest natural attraction. The terrain combines generous amounts of sloping plains to provide a foreground for seemingly ever-present mountains, cliffs, and hills. The charm of its relief is enhanced by the widespread exposure of colourful rocks. The brilliant sunlight illuminates vegetative cover that ranges from green-carpeted pine and fir forests to bizarre shrub and cactus deserts, all augmented by brief bursts of colour from spectacular seasonal displays of wild flowers.

Physiographic characteristics. To Arizona's two major physiographic divisions, the Colorado Plateau and the Basin and Range Province, local authorities add the Transition. The northeastern two-fifths of Arizona is part of the scenic Colorado Plateau. Far less rugged than adjacent portions in Utah, these tablelands in Arizona consist mainly of plains interrupted by step-like escarpments. Though labelled mesas and plateaus, their ruggedness and inaccessibility has been exaggerated. The incomparable Grand Canyon of the Colorado River provides the major exception to what has proven to be an area easily traversed. Forest-clad volcanic mountains atop the plateaus provide the state's highest points, Humphrey's Peak, 12,633 feet (3,851 metres), in the San Francisco Mountains, and Baldy Peak, 11,590 feet (3,533 metres), in the White Mountains.

Land forms

Over 200 miles of the southern border of the Plateau is marked by a series of giant escarpments known collectively as the Mogollon Rim. West and south of the rim, a number of streams follow narrow canyons or broad valleys south through the Transition region and into the Basin and Range Province. The Transition region border-

ing the plateaus comprises separated plateau blocks, rugged peaks, and isolated rolling uplands.

The Basin and Range region of the southern and western third of the state, containing the bulk of the population but none of the large canyons and mesas for which Arizona is famous, consists largely of broad, open-ended basins or valleys of gentle slope. Isolated mountains rise like islands in the desert plain.

Contrary to desert stereotypes, sand dunes cover less than 1 percent of the state, and stony desert surfaces are seldom visible. The younger, transported soils of past and present river floodplains provide the more desirable soils for agriculture. Only the Colorado River and several dozen small headwater streams in the well-watered White Mountain–Mogollon Rim–Transition areas have year-round flow. Old records indicate that such major rivers as the Gila, Verde, San Pedro, and Santa Cruz once had stretches of permanent water complete with fish, beaver, and otter.

Climate. About half of Arizona is semi-arid, one-third is arid, and the remainder is humid. The Basin and Range region has the arid and semi-arid subtropical climate that attracts most winter visitors and new residents. Receiving over 80 percent of the possible sunshine, January days have a mean maximum temperature of 65° F (18° C) in Phoenix. Mild frosts can be expected at most locations during a four-month period, and light rains punctuate the winter months, interrupting exceedingly dry falls and springs. Summer daily maximum readings average 104° F (40° C) in Phoenix, and night temperatures drop to an average of 78° F (26° C).

Moisture-laden air from the Gulf of Mexico moves into Arizona during July, bringing more than two months of irregular but sometimes heavy thundershowers that are locally referred to as the "summer monsoon." Phoenix receives more than one inch (25 millimetres) of rain in both June and July, and Tucson receives twice that amount.

The Colorado Plateau has cool to cold winters and a semi-arid climate. Near mile-high elevations and direct exposure to polar air masses can produce January mean highs and lows as divergent as the 46° F (8° C) and 19° F (−7° C) in Winslow. Summer temperatures on the plateau are generally 10° cooler than those of Phoenix. Most of the region receives from 10 to 15 inches (255 to 380 millimetres) of rain, with a winter maximum along the western borders.

Because of the great amount of relief, climatic conditions within the Transition vary widely over small areas. Despite its desert image, 17 percent of Arizona falls into the humid class, much of it lying in the Transition and adjacent high southern edge of the Plateau.

Vegetation and animal life. Considering the great variety in relief and climate, it is not surprising to find similar diversity in the vegetation. About 10 percent of Arizona is forested, 25 percent woodland, 25 percent grass, and 40 percent desert shrub. Elevations above 6,000 to 7,000 feet host magnificent forests of ponderosa pine, topped in the highest areas by Douglas and other firs, spruces, and aspen. Below the forests, piñon pine dominates the plateau woodlands, and evergreen oak or chaparral the Basin and Range. Plains grasses cover about one-third of the plateau, and Sonoran or desert grass carpets the higher elevations of the basins. Mesquite trees are invading many of the former grasslands in the south. Foothills in the Tucson–Phoenix area carry exotic giant saguaro cacti of the Sonoran Desert, matched in areas of the northwest Basin and Range by dramatic stands of Joshua trees. Shrubs dominate the lowest portions of all areas: big sagebrush and saltbush in the plateau, creosote bush in the Basin and Range.

Variety of natural life

The animal life is even more varied, containing representatives of the Rocky Mountain, Great Plains, and Mexican ecological communities. Among the larger animals are black bear, deer, desert bighorn sheep, wapiti, or elk, and antelope. The tropical coati, a racoon-like mammal, has spread northward into Arizona, and the javelina ("wild pig") is a favourite game animal in the south. Among the several cats, the bobcat thrives, and the mountain lion is the most prized. Coyotes, skunks, and porcupine abound, as do cottontail and jackrabbits, and there are several varieties of foxes. Tropical birds include the thick-billed parrot and the brilliantly plumed, coppery tailed trogon, which visit the mountains along the southeastern border. Game birds include turkey and several varieties of quail and dove. Among native fish are the Arizona trout and the six-foot minnow known as the Colorado squawfish. Poisonous animals include the rattlesnake, the scorpion, and the Gila monster.

Patterns of human settlement. Despite Arizona's romantic image as a land of picturesque ghost towns and mining camps, isolated ranches, primitive Indian reservations, and bucolic cotton and citrus farms, more than three-fourths of its people are congregated in modern urban settlements of over 2,500 population. Phoenix and Tucson together account for about half of the state's population. The unincorporated status of mining centres with populations of up to 6,000 and of new towns comprising retirement and resort communities of from 2,000 to 15,000 people further obscures the degree of urbanization of the Arizona population.

Approximately 30,000 people live on the more than 6,000 recognized farm and ranch units. The typical ranch house and headquarters stands at the mouth of a canyon or some other source of dependable water. The irrigated lands of the central and southern regions contrast with the surrounding desert, presenting a flat, green, tree-studded landscape that often resembles cultivated plains of the Middle West and South. Rectangular fields planted in rows of cotton, sometimes paralleled by concrete-lined irrigation ditches with earthen banks, give regularity and linearity to the landscape.

Rural and ranching life

Most towns and cities have low-density populations, the result of large lots and considerable vacant land within built-up areas. Residential building during the 20th century adhered almost exclusively to styles popular in the West, producing relatively uniform neighbourhoods of bungalow, Spanish revival, and ranch-style houses. Smaller towns contain mostly wood-frame dwellings and brick commercial buildings, and structures of mud adobe can be seen in older areas throughout the southern part of the state. Prescott, the first capital of Arizona, reflects its founding by Northerners in its red-brick buildings, central courthouse square, and dwellings more typical of eastern America.

The dominance of the Phoenix trading area, which extends over most of the state, began with its selection as the permanent location for the state capital, its central location, and the extensive agricultural economy that developed after the completion of Roosevelt Dam and Reservoir on the Salt River in 1911. Tucson, the early centre of Mexican and Anglo settlement, continues to be the wholesale, retail, and entertainment focus of southern Arizona. It maintains well-developed commercial and medical ties with Sonora and northern Sinaloa.

THE ARIZONANS

Arizonans traditionally identify themselves as Anglo, Mexican, Indian, black, and Chinese Americans. The numerically dominant Anglos in the cities have a variety of European backgrounds, but few are foreign born, and ethnic identification is less important than association with a prior home state. Traces of the heavy Texas contribution to the rural population can be detected in speech, attitudes, and customs.

Minority ethnic groups. The more than 450,000 Arizonans with Spanish surnames proudly claim to be Mexicans, the term generally used in Arizona for the Mexican-Americans. These Mexicans are affected in varying degrees by the practice of social and residential segregation. Some are found in barrios where Spanish is the common tongue, whereas others disperse throughout the cities and participate fully in the overall business, political, and social life of their communities. Where language and lack of skills are not a barrier, equal employment opportunities generally prevail in southern Arizona's larger cities and companies. Some of the most prominent families of Tucson are of Mexican descent, and intermarriage in

The Mexican and Indian peoples

such border communities as Nogales is not uncommon. Mexican music, food, building styles, furnishings, clothing, and social customs have been widely adopted by all Arizonans.

Arizona has long been associated with the Indian American, for here occurred many of the final, dramatic conflicts between frontiersman and native. Few native Arizona Indian tribes experienced the tragic annihilation or displacement that occurred elsewhere, and today more Indian Americans live in Arizona and comprise a higher percentage of the total population than in any other state. More than 90,000 people group themselves into 14 tribes on 19 reservations, which range in size from the 1,400-acre Yavapai Reservation to the nearly 9,000,000-acre reserve of the Navaho people. The latter tribe, numbering over 50,000 in Arizona, presses vigorously to direct the development of its land and people, and the tribal government assumes partial or complete responsibility in many areas of social and economic life.

Among the remaining tribes, which number more than 30,000, the best known are the once-militant Apache and the highly talented Hopi people, each of whom also pursues aggressive development programs similar to those of the Navaho. Less well-known are the equally numerous Papago and Pima tribes, historic allies of the frontiersman. The peaceful Papago, whose median family income on the reservation is 6 percent of the national Anglo income, may well be America's poorest people. About half of them live off the reservation.

For Arizona's black population, comprising 3 percent of the total, living quarters remain largely segregated, but Arizona schools desegregated voluntarily in the early 1940s.

Demography. The 1970 population of 1,770,900 represented a 36 percent increase over 1960, but the rate of growth was only half that recorded during the previous decade. Conservative estimates project a 1980 population of 2,400,000. Immigration of persons seeking the state's climatic amenities and economic opportunities accounts for half the population gain, the remainder resulting from a birth rate well above the national average. The death rate remained below the national average, in part because the percentage of residents above 65 years of age is below the national average.

THE STATE'S ECONOMY

As Arizona's population grew, its economy shifted from a frontier stage emphasizing primary, natural-resource-oriented industries to secondary and tertiary industries associated with more advanced economies.

Manufacturing and tourism. Manufacturing has become the most important basic industry, notably in electrical, communications, aeronautical, and aluminum-product production. The great increase in manufacturing employment since World War II has given Arizona one of the most dynamic economies in the nation.

Tourist expenditures account in part for the high per capita employment and volume of retail sales. Manufacturing and tourism both depend upon natural resources in an essentially nonconsuming fashion. Manufacturers find a large, multitalented labour pool attracted to the state's climate and scenery. Tourism reflects even more directly the warm winter climate and the physical features of the landscape. Natural and human features of historic, scientific, and recreational value have been protected and developed by the government, probably to a greater extent in Arizona than in any other state. There are 34 federal parks, monuments, memorials, forests, recreation areas, historic sites, and game and wildlife refuges, as well as ten state and tribal parks.

Presence of the federal government

Government and land resource. The magnitude of the federal government's role in the Arizona economy is evident also in its numerous military facilities. The total personal income of federal military and civilian employees exceeds that from agriculture and mining combined, and only manufacturing generates a higher gross income. These expenditures reflect in part the federal government's ownership of 45 percent of Arizona's land and the holding in trust of another 27 percent in the form of Indian reservations. Only about 18 percent of Arizona is private land, but there is no shortage of acreage for urban settlement. Though the economic effects of large public holdings remain controversial, the positive contribution of public lands to the aesthetic quality of the landscape is unquestioned.

Agriculture and water supply. The most stable industry is agriculture. Acreage and employment have been dropping slightly, while income increases modestly. Although Arizona produces 47 percent of America's long-staple cotton, the demand is limited, and this premium variety accounts for only 10 percent of the total cotton acreage. Vegetable acreage, especially in lettuce, has increased so dramatically that winter produce is now the leading source of crop income. Citrus acreage continues to increase in the Yuma area, more than replacing acreage lost to urban expansion in the Salt River Valley. All but a few thousand acres on Indian reservations of the Colorado Plateau are under irrigation, producing yields double the national average for most crops. Farms are larger, more heavily capitalized, and more frequently organized on a corporate basis than in any other state.

The grazing of livestock occupies 30 times more land than crop raising, and its cash value is more than triple that of either cotton or vegetables but less than half that of tourist income. Feeder lots near agricultural areas, much less romantic than the open range, hold half of the more than 1,000,000 cattle within the state.

Arizona agriculture anticipates a water shortage stemming from the expansion of urban areas into surface-water irrigation districts and increased cost of pumping from declining water tables. Farms and ranches account for over 90 percent of the water used annually for all purposes. The annual overdraft of groundwater withdrawals over natural recharge has been between 3,500,000 and 4,000,000 acre-feet annually for some years.

Mining and forestry. Copper remains Arizona's most distinctive contribution to the national economy, accounting for over 50 percent of the nation's total annual production. The conversion of most mines to highly efficient open-pit operations employs a relatively small but growing labour force. Interest in small-scale mining, especially of gold and silver, greatly exceeds the actual production, most of which comes as a by-product of copper mining. Petroleum from the Four Corners area of the Navaho Reservation moves by pipeline to California, and coal from the Black Mesa area of the Hopi reservation will be sent by pipeline to Nevada generating stations.

The lumber and pulp-paper industry deserves notice because it clearly contradicts two widely held illusions about Arizona: that the state is all desert and that a shortage of water has deterred industrialization. Most of the ponderosa-pine lumber comes from the Mogollon Rim–White Mountain area.

Transportation. Arizona's transcontinental routes carry more people and goods through it than to it. The east–west route traversing the relatively level Colorado Plateau provides the nation's major highway and railway link between the Middle West and Southern California. A historically older but currently less used route through the valleys and basins of the south is followed by the Southern Pacific Railroad and a modern interstate highway. Divided, limited-access highways also serve most of the north–south route connecting Flagstaff, Phoenix, Tucson, and Nogales. The large amount of gentle relief and the small amount of rainfall make unpaved roads serviceable for remote areas, except immediately following a rain. Four-wheel-drive vehicles and motorcycles explore the backcountry on weekends—to the increasing concern of cattlemen and environmentalists. Both Phoenix and Tucson have direct air flights to major cities in the United States and to Mexico.

Historic and contemporary routes

ADMINISTRATION AND SOCIAL CONDITIONS

Governmental structure. The constitution of Arizona reflects the strength of the Progressive movement at the time of the constitutional convention in 1910. It provides for maximum citizen participation through initiative and referendum on legislation and recall of all elected of-

ficials, including judges. It also features a broad dispersal of executive power: the governor's office shares administrative authority with numerous other elective offices. The chief functions of the governor are to recommend and veto legislation, perform ceremonial acts, and mold public opinion, but there is a trend toward a stronger executive branch. The attorney general, who succeeds to the governorship in case of a vacancy, holds the second most contested elective office.

The legislature comprises a 60-member House of Representatives and a 30-member Senate. The massive growth of Phoenix and Tucson, combined with re-appointment, has given urbanized Maricopa and Pima counties three-fourths of the seats in both houses.

A constitutional amendment in 1960 restructured the judicial branch into a Supreme Court, Appellate Court, Superior Court, and local justice and appeals courts; there are no special courts. Despite repeated attempts to make judicial offices appointive, judges campaign for election on a nonpartisan basis after running in the primary with party identification. The state constitutes one entire federal judicial district.

The 14 counties, acting as agents of the state, constitute the basic medium of local government. Elected county supervisors are relatively free of legislative direction other than recently imposed restraints upon taxation and budgeting. There are no township governments. State law prescribes the type of town government available for settlements under 3,000 and the city form to be used by larger communities, but metropolitan centres have considerably more freedom in organization and operation.

Politics. Arizona has changed from a traditional one-party state dominated by the Democrats to a two-party distribution, in which Republicans repeatedly have won the highest offices since 1950 and gained control of the legislature for the first time in 1966. Liberal Democratic factions continue to receive support in mining communities, among traditionally Democratic Mexican-Americans, and from segments of the metropolitan populations, though conservative "Pinto Democrats" in rural areas often support Republican candidates.

Standards of living. Wages and income vary greatly between the extremes of manufacturing and the part-time seasonal work in agriculture and services, and average figures are meaningless. The large number of employees in high-paying electronic and military-related industries push the average wage in manufacturing to the national figure. The large underemployed Indian-, Mexican-, and Negro-American population, however, coupled with the potential inflow of unskilled labour, keeps wages for unskilled workers low.

Arizona as a health mecca

Health and welfare. Public opinion consistently ranks Arizona among the healthiest areas in the nation, if not the world, though actual data are not adequate to support the belief. Among Indians, infant mortality and a host of maladies occur at high rates, tuberculosis remains a problem among the general population, and San Joaquin Valley fever (coccidioidomycosis) has been almost endemic. Evidence indicates that the low humidity and mild winters have a therapeutic or arresting effect on certain forms of pulmonary and arthritic health problems, but it has not been demonstrated scientifically that the general native-born population or young immigrants can expect to be healthier than residents elsewhere in the United States.

The state Department of Health controls a variety of administrative, aid, and inspection services, including institutions for mental and tubercular patients in Phoenix. Several of the more populous counties maintain public hospitals. Private medical care in the metropolitan areas is excellent, but for the state as a whole, the number of physicians and dentists falls below the national level. Despite its attractiveness to the ill and the aged, Arizona has no more than its per capita share of the nation's hospitals and nursing homes. The recently established College of Medicine at the University of Arizona and two university-affiliated nursing colleges work toward expanding the supply of medical personnel.

A generally modern approach to public welfare by the legislature has been coupled with financial restraints that limit the effectiveness of the programs. The Department of Public Welfare works mainly through county agencies with a variety of programs for children and for the aged, blind, and permanently and totally disabled.

Education. Public education has struggled to meet the rapid increase in students accompanying the population boom. Five years of college work is required for secondary-school teachers, and children must attend school between the ages of 8 and 16. Elementary, secondary, and consolidated districts operate with the assistance of county and state superintendents and an appointed state Board of Education. State support for local districts has been increasing, accompanied by tighter controls over maximum annual budget increases. Public school districts have begun to supplement and replace some reservation schools operated by churches and the federal Bureau of Indian Affairs.

Like most states in the Rocky Mountains, Arizona's higher education is dominated by large public universities. The Arizona Board of Regents assumes responsibility for the University of Arizona (founded, 1885) in Tucson, Arizona State University (1885) in Tempe, and Northern Arizona University (1899) in Flagstaff. To meet the need for higher education throughout the state, seven junior colleges had been established by 1970 under the joint control and support of county boards and the state. There are two private four-year institutions, Grand Canyon College in Phoenix and Prescott College in Prescott, as well as the private Thunderbird Graduate School of International Management in Phoenix.

CULTURAL LIFE AND INSTITUTIONS

The arts. Traditionally a centre for Indian folk arts and crafts, Arizona had no early circles of painters and writers comparable to those of neighbouring New Mexico. Interest in painting, crafts, drama, music, and publishing, however, has increased with population growth. Architecture and the graphic arts are particularly aware of Southwestern regional themes. Writers of fiction and nonfiction alike have focussed their attention upon Arizona's 19th-century frontier era. Among the best-known earlier writers of popular fiction were Zane Grey and Harold Bell Wright, both of whom employed local settings and lived and worked in the state during periods of their careers. Later contributions of note were made by Oliver LaFarge, Will Comfort, and Ross Santee.

Indian arts and crafts

Contemporary arts and crafts of the Indian American, executed within the dynamic traditions of the tribes, but with considerable individual creativity, receive enthusiastic support from tribal organizations and the local general public. The Hopi and Navaho people have among them outstanding painters, silver and jewelry craftsmen, weavers, basketmakers, and potters; and Papago women produce a variety of handsome baskets.

No city dominates as an art centre, though Scottsdale, Tucson, Sedona, and Tubac have colonies of working artists. Outstanding collections, mainly paintings, can be viewed in the Phoenix Art Museum and the Museum of Art at the University of Arizona in Tucson. The Arizona State Museum in Tucson, the Heard Museum of Anthropology and Primitive Art in Phoenix, and the Museum of Northern Arizona in Flagstaff feature archaeological and traditional collections of Indian arts and crafts. The Arizona–Sonoran Desert Museum in Tucson has received worldwide attention as a living museum dedicated to the natural world of the Sonoran Desert and the implications of its occupancy by man.

Both Tucson and Phoenix support symphonies. Mexican and ranchero music are heard in the central and southern part of the state. Theatre has been popular since the early mining camps.

More than any other art form, architecture embodies the conflict between the regional traditions of the Southwest and modern international trends. The best known modern structure is Frank Lloyd Wright's auditorium at Arizona State University. Among the many structures in the Spanish idiom, the Heard Museum is outstanding, and the Nogales Public Library synthesizes the Spanish

Southwestern and contemporary styles. The most photographed and beloved building in Arizona remains the Mission San Xavier del Bac, the "White Dove of the Desert," in Tucson, completed by the Franciscans in 1797.

Libraries, publishing, and communications. The University of Arizona library is the largest in the state. In addition to other college and university libraries, there are several specialized collections. County and state regional systems supply library services throughout the state.

The state's leading book publisher, the University of Arizona Press, releases a variety of scholarly and popular titles, most with a Southwestern focus. The state's most widely known publishing venture, *Arizona Highways*, brings varied features of Arizona to a worldwide audience.

Of the 73 newspapers published, 13 are dailies, of which the Phoenix *Arizona Republic* has the largest circulation. Tucson's *Arizona Daily Star* is frequently the only major newspaper that supports Democratic candidates. Both have nearly statewide circulation.

Recreation. A variety of individual and organized sports and recreational activities compete with the arts and communication media to provide diversified entertainment and leisure activity. Varied desert and forest terrains and several dozen manmade lakes atttract thousands of hunters, fishermen, campers, hikers, and amateur prospectors and historians to the open country throughout the year. Rodeos revive the Old West in all the cities and on the larger Indian reservations.

Prospects. The political and economic institutions of Arizona are gradually adjusting to the recent growth in urban population, but problems of providing job opportunities and government services will remain for some time. With growth, local concern increases over the possible loss of the amenities of clean air and undisturbed desert and forest that initially attracted so many residents to the state. The eventual use or disposition of Indian and federal lands is a fundamental consideration in Arizona's future.

BIBLIOGRAPHY. UNIVERSITY OF ARIZONA FACULTY, *Arizona: Its People and Resources* (1971), the most comprehensive and reliable source available; ROGER DUNBIER, *The Sonoran Desert* (1968), a highly readable and competent treatment of southern Arizona and northwest Mexico, dealing with man and his use of the physical setting through time; BYRD H. GRANGER (ed.), *Arizona Place Names*, rev. ed. (1960), an excellent travelling companion as well as an armchair introduction to much of Arizona's geography and history; DONALD W. MEINIG, *Southwest: Three Peoples in Geographic Change, 1600–1970* (1971), a study of the manner in which the Indian, Hispano, and Anglo have given character to a region which includes west Texas, New Mexico, and Arizona; JOSEPH MILLER and HENRY G. ALSBERG (eds.), *Arizona, the Grand Canyon State: A State Guide*, 4th ed. (1966), though still not completely revised, this later edition of one of the best of the W.P.A. Writer's Guides is the most complete guide available; ODIE B. FAULK, *Arizona: A Short History* (1970), a well-written, factual account; *Arizona Highways* (monthly), excellent colour photographs of natural and human landscape features accompanied by sometimes lyrical, sometimes solid texts; ELDRED D. WILSON, "A Résumé of the Geology of Arizona," *Bull. Bur. Mines Univ. Ariz. 171* (1962), written for the curious, intelligent layman; WILLIAM SELLERS and CHRISTINE GREEN (eds.), *Arizona Climate*, rev. ed. (1964), all the reliable data available plus a discussion of the climate of the state of all stations with reasonably complete records; TOM BAHTI, *Southwestern Indian Tribes* (1968), one or two pages of accurate historical and contemporary information about the people, their reservations, and their crafts, plus sound suggestions for more detailed reading (well illustrated); BRUCE B. MASON and HEINZ R. HINK, *Constitutional Government in Arizona*, 2nd ed. (1965), an evaluation as well as a description of state government; JOSEPH WOOD KRUTCH, *The Voice of the Desert* (1955), along with *The Desert Year* (1952), Krutch's work represents the finest literary interpretation of the desert to appear in recent years.

(M.E.H.)

Arkansas

Ever since Arkansas was admitted as the 25th member of the United States in 1836, its people have maintained a remarkable homogeneity, and today most of them, whether white or black, are native to the state. Striking cultural contrasts exist within the state, however, with the long-isolated mountain peoples who eked out subsistence livings in the north and west counterposed to the peoples to the east and south who created a Southern environment in which cotton growing and sharecropping long were dominant modes of economic life. Between the two regions lies Little Rock, the capital and the urban and economic centre of the state. Its location and increasingly cosmopolitan character are symbolic of Arkansas's growing unification and urbanization, reflecting the broader patterns of American life in the 1970s.

Few of the more than 1,900,000 Arkansans counted in the 1970 census any longer feel complacent about the state's relative poverty and backwardness. Although Arkansas remains among the lowest ranking states in per capita income and other economic indicators, the 1960s saw this income rise by nearly 50 percent; value added by manufacture rose some threefold; the overall economy gained faster than that of the national average; and the population increased, reversing a long decline. The decade saw also the beginning of programs to increase these upward trends and to continue the process of equalizing the educational, economic, and social opportunities of black and white citizens.

Arkansas's 53,104 square miles (137,539 square kilometres) make it 27th in area among the states, but, except for Hawaii, it is the smallest state west of the Mississippi River. Its neighbours are Missouri to the north, Tennessee and Mississippi to the east, Louisiana to the south, Texas to the southwest, and Oklahoma to the west. A map of Arkansas shows not only the high Ozark and Ouachita Mountains in the north and west but also a heavy tracery of rivers cutting through rich agricultural lands. Nearly all of them flow from northwest to southeast and empty directly into the Mississippi, which forms the entire eastern boundary. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; CIVIL WAR, U.S.; NORTH AMERICA; and MISSISSIPPI RIVER.)

THE HISTORY OF ARKANSAS

Exploration and settlement. Arkansas's earliest inhabitants were bluff dwellers whose farming and hunting culture flourished *c.* AD 500. Later mound-building cultures left sepulchral mounds and other remains along the Mississippi.

Following Spanish and French explorations of the trans-Mississippi regions in the 16th and 17th centuries, the Frenchman Henri de Tonty founded the Arkansas Post on the Lower Arkansas River in 1686. The first permanent white settlement in the future state, it served as a fur-trading centre and a way station for travellers between the Gulf of Mexico and the Great Lakes.

Following the Louisiana Purchase by the United States in 1803, Arkansas lay within the territories of Louisiana until 1812 and of Missouri until 1819, when it became a separate territory. Its northern boundary, latitude 36° 30′ N, was the famous line of the Missouri Compromise in 1820 that was to separate the slave and free states in the newly opening West.

Statehood and Civil War. By the time of statehood in 1836, all land titles of the Quapaw, Osage, Caddo, Cherokee, and Choctaw Indians had been withdrawn by Congress, and the tribes were forced westward into the Indian Territory, the future Oklahoma. Violence broke out along the state's western border until near the end of the 19th century, when the frontier atmosphere disappeared with the white settlement of the Indian Territory.

War and reconstruction: conflicting sentiments

Although a slave state, Arkansas did not secede from the Union until May 1861—five months after South Carolina did so. Arkansas took this action only after the Confederate capture of Ft. Sumter and President Lincoln's call for volunteers. Union sentiment was strong in northern Arkansas; about 6,000 Arkansans joined the Federal forces. About 58,000, however, fought for the Confederacy. Little Rock fell to Federal troops in 1863,

and the state was a legislative battleground between secessionist supporters and the imposed Republican government for a decade. Arkansas was readmitted to the Union in 1868, but internal strife approached open warfare. In 1874 the state returned to the Democratic Party, and remained with it until Winthrop Rockefeller, a Republican, was elected governor in 1966.

The chief long-range effects of the war on Arkansas, in common with most of the former Confederate States, were a crop-lien sharecropping system, a race problem of formidable and new dimensions, a one-party political system, and widespread poverty. Economic development in Arkansas was severely handicapped by the collapse of state credit following repudiation in 1885 of bonded indebtedness including interest of nearly $14,000,000.

Recent decades. Until World War II, Arkansas experienced slow economic development, remained predominantly rural, and was tied to a single cash crop—cotton. The Depression of the 1930s was worsened by years of drought that turned many farm families into itinerant labourers across the nation. In 1957, federal troops entered Little Rock to maintain order there for a year, after the state militia had been ordered to prevent the desegregation of Central High School; the confrontation focussed international attention on the state. In the 1960s the state's attempts to improve its economic status seemed at last to be bearing fruit.

THE NATURAL AND HUMAN LANDSCAPE

The physical environment. *Geographical regions.* A line drawn from the southwest corner to the northeast corner of the state approximates the division between the highlands lying west and north and the lowlands lying south and east. The highlands are divided by the Arkansas River Valley into the Ouachita Province on the south and the Ozark Plateau on the north. The lowlands include the Mississippi Alluvial Plain in the east and the West Gulf Coastal Plain in the south and extreme southwest. The highlands are covered by the dense pine and hardwood forests of the Ouachita and Ozark national forests.

Mountains and plains of Arkansas

The Ozark Plateau is broken by broad, flat-topped ridges and steep valleys with fast moving streams. The more rugged southern edge, known as the Boston Mountains, contains the highest elevations. Excellent farmland, producing a wide variety of crops, lies in the northern part. The Arkansas River Valley contains the highest point in the state, Mt. Magazine, at 2,823 feet (860 metres). The western section has extensive coal and natural-gas deposits. Several peaks in the Ouachita Province reach 2,500 feet. The mountains are eroded, exposing faulted rock, and the ridges extend west and east. The famous Hot Springs National Park is in this area.

The West Gulf Coastal Plain has gentle hills suitable for livestock grazing as well as general farming. Much of this area consists of pine and white-oak forests, which sustain extensive lumbering industries. Petroleum and natural-gas deposits have been developed in the region around Smackover and El Dorado. The Mississippi Alluvial Plain, much of which was once a vast swamp, is now well-drained and protected against flooding. It contains the state's richest and most fertile farmland. Rice and soybeans have replaced cotton as the major crops. A long, narrow chain of hills, Crowley's Ridge, runs north and south through the centre of the plain.

Climate. The climate generally is mild in winter and hot in summer. Normal high–low temperatures in Little Rock in January are 51°–31° F (11°– −1° C); in July, 93°–71° F (34°–22° C). The normal annual precipitation of 48 inches (1,220 millimetres) is distributed about equally during the year, though summers tend to be drier.

Vegetation and animal life. A great variation in soil types and elevations in Arkansas provides environments for a large number of plant species. There are more than 200 species of trees, of which the pine, oak, hickory, maple, gum, ash, cypress, and elm are most important. The woods in fall and spring are colourful with dogwood, flowering fruit trees, redbud, innumerable wild flowers, and azaleas.

Arkansas is situated on the Mississippi flyway; migratory water birds and some 300 native species attract hunters to the rice fields and reservoirs of eastern Arkansas. Deer, quail, opossums, turkeys, squirrels, and rabbits are among the more abundant game animals. Bobcats and wolves are not uncommon in the hill country. The lakes and streams of the state offer an abundance of fish—including crappie, bass, drum, catfish, buffalo, gar, and trout.

The landscape under human settlement. *Regional diversities.* The inhabitants of the Ozarks and Ouachitas lived until recently in rural isolation, which bred an independence of spirit and a suspicion of strangers. They long lived off the land: hunting and fishing were essential to supplement the limited produce of their farms. Since a plantation economy was impracticable in the uplands, few slaves were brought into the region. In eastern Arkansas the plantation economy produced a vast gulf between the sharecroppers and tenants on one end of the social scale and the managers and landlords on the other. Small farmers and merchants constituted another class. The croppers lived a bare and meagre existence. Handicapped by lack of economic resources and education, they accomplished remarkable results through the Southern Farm Tenants Union, which they organized in eastern Arkansas in the 1930s; this organization has influenced the national farm policy of presidents from Franklin D. Roosevelt onward.

Contrast of cultures

Although changes in the economy of both regions were evident earlier, the rate of change since World War II has been dramatic. The Ozarks are no longer isolated. A network of paved highways brings tourists to enjoy the region's scenic beauty and varied recreational activities. Numerous "retirement villages" attract visitors and buyers from across the country. The tourist industry will probably remain the economic mainstay, though small industrial plants have taken advantage of the climate and the ample labour supply.

Mechanization of farming in eastern Arkansas and the shift from cotton farming to rice and soybeans has virtually eliminated the sharecropper—though not the rural poor. As the pace of mechanization increased, so did the exodus of the former tenant farmers to cities in the North and East. Farming is increasingly a corporate venture rather than a family affair. Eastern Arkansas is still, however, more Southern in character than the mountainous region. Many of the rural shacks have gone and the towns have grown, but mile after mile of cultivated fields are broken only by small woodlands and pastures. Here and in central Arkansas reside the majority of the blacks, many of whom still work the land as did their ancestors. The people still take pride in maintaining traditional "Southern hospitality."

Urban Arkansas. Much of the residential area of Little Rock lies in the low hills and valleys of the easternmost foothills of the Ouachita Mountains. Its downtown is being redesigned as part of an urban renewal project. The city is the major port on the Arkansas River, a marketing centre, and a fast-growing manufacturing centre. Fort Smith, the second largest urban centre on the Arkansas at the western boundary, is the most industrialized city in Arkansas and a regional centre for northwestern Arkansas and northeastern Oklahoma. The economy of Pine Bluff, some 50 miles downriver from Little Rock, depends primarily upon the surrounding agricultural area. Texarkana, contiguous with the Texas city of the same name, is important as a railway regional trade centre.

THE PEOPLE OF ARKANSAS

Composition and distribution. Prior to the Civil War, Arkansas's population was largely from Kentucky and Tennessee, a part of the westward movement of Scottish, Scots-Irish, and English stock from Virginia and the Carolinas since early Colonial times. The black population in 1860 was about 110,000, or 25 percent of the total; in 1970, it was nearly 360,000 but only about 18 percent. Three counties in eastern Arkansas were more than 50 percent black. The heaviest concentration of

population is in the fertile eastern alluvial plain, in the river valleys, and on the plateaus in the northwest.

Religion and social life. The largest religious denominations are the Baptist, Methodist, Presbyterian, Church of Christ, and Roman Catholic. The general religious atmosphere is that of conservative fundamentalism, and Arkansas is considered to be a part of the Middle American "Bible Belt." This fundamentalism underlies many characteristic attitudes of Arkansans. The state's law forbidding the teaching of the theory of human evolution never has been repealed, though federal Supreme Court decisions have rendered it unconstitutional. The sale of alcoholic beverages is subject to local option; many counties and cities prohibit their sale or permit it only in private clubs and certain other establishments in major cities. The right-to-work amendment to the state constitution in 1944, prohibiting compulsory union membership, was sponsored by the Christian Association, among others. Harding College in Searcy is the site of an annual Freedom Forum, which advocates a blend of religious fundamentalism, extreme patriotism, and free-enterprise capitalism.

Demography. Following a peak in 1940, Arkansas's population experienced two decades of sharp decline. Although it was close to the national average in natural increase (*i.e.*, in the greater number of births than deaths), Arkansas in each decade lost through emigration about one-third more blacks and about one-sixth more whites than it gained through immigration. The loss was checked after 1965 to produce a 7.7 percent gain during the decade, though the black population continued to decline, both relatively and absolutely.

Patterns of migration

The 50 percent urbanization of Arkansans recorded in 1970 reflects both a real growth in the cities and a real decline in the countryside, as well as a major change in the state's economy. Between 1950 and 1965, the number of persons engaged in agriculture declined by nearly 75 percent, to about 60,000. The loss of these former farmers to other states appears to have been checked by greater economic opportunity within the state.

THE STATE'S ECONOMY

Cotton is no longer king in Arkansas, and the state is no longer primarily agricultural. Industrialization and urbanization are major factors behind Arkansas' record of economic progress during the 1960s. Labour unions are strong in transportation, utilities, construction, and heavy industry, but most of the state's labour force is unorganized. In both political and economic policy-making, labour has been less influential than business.

Natural resources. Oil production began in 1921 in southern Arkansas, but conservation controls have restricted production. Natural gas is drawn from the oil fields and the Arkansas River Valley. Coal of near smokeless quality also is found in the valley. The bauxite (an aluminum-bearing ore) that is obtained by strip-mining to the south and southwest of Little Rock represents more than 90 percent of America's supply. Magnet Cove near Hot Springs contains more than 40 different minerals in one small valley; barite and titanium are the most important. Arkansas whetstones made from novaculite are regarded as among the finest in the world. Near Murfreesboro in southwest Arkansas is the only diamond mine in the nation, now operated only as a tourist attraction. Almost one-half of Arkansas is covered by forests, notably extensive stands of pine and white oak.

Mineral wealth

Hydroelectric power is produced at most of the dams erected by the U.S. Army Corps of Engineers and at one privately built dam, but steam-generated power plants produce most of the energy. Two nuclear-powered generating stations are under construction near Dardanelle.

Agriculture and manufacturing. As recently as 1960 cotton was still the major source of agricultural income, but by the end of the decade the major crops were, in order, soybeans, poultry, cotton, and rice. Commercial fish farming has begun to take advantage of the extensive rice paddies of eastern Arkansas. Farms have followed the national trend of increasing in size while decreasing in number.

Manufacturing is chiefly of the consumer-goods type. Industries that increased substantially in numbers of employees in the 1960s were food processing and manufacturing of wearing apparel, furniture, and electrical and nonelectrical machinery.

Tourism. Arkansas devotes considerable effort to its attractions for out-of-state vacationers, who annually contribute millions of dollars to its economy. State and national agencies stock lakes and streams with fish, and the state's preserves and conservation practices assure ample game in hunting seasons. The largest single attraction is Hot Springs National Park, which offers both out-of-doors and luxury-hotel recreation throughout the year. Blanchard Springs Caverns, which appears to rival New Mexico's Carlsbad Caverns in size, is being developed by the National Park Service; it is expected to be another major tourist attraction.

State finances. Although state revenue has been increasing, mounting needs for additional revenue led to tax increases in 1971, the first since 1957. The state hopes to generate future revenue by raising per capita incomes—increasing the number of employment opportunities and developing the human resources to their utmost.

Transportation. Five major railroads provide freight service on some 3,000 miles of track within Arkansas, as well as to major cities in the central United States. Airline service is provided by three national and two regional carriers to any point in the nation from a growing number of airports. Some 60 common truck carriers operate over more than 600 miles of interstate expressways, and 50 percent of the nation's population is within a two-day driving radius of Arkansas. The Arkansas River Project in navigation and flood control is the largest civil works project ever undertaken by the U.S. Army Corps of Engineers. The project has cost over $1,300,000,000 and provides access to more than one-half of the nation's navigable inland waterways.

The Arkansas River Project

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. Adopted in 1874, Arkansas's constitution has been amended more than 50 times. In 1970 the voters rejected a new constitution. The governor is elected to a two-year term; he has the authority to pardon convicts, to summon the legislature into special session, and to veto acts, though he may be overridden by a two-thirds vote in each legislative house. In 1971 the governor gained greater control over 13 previously independent executive agencies through the creation of "cabinet-rank" departments. The Senate has 35 members with four-year terms; the House of Representatives, 100 members with two-year terms. The judicial branch contains a Supreme Court of seven popularly elected members who serve eight-year terms, 18 circuit court districts, and 13 chancery districts. Where established, municipal courts have jurisdiction throughout the county.

Arkansas's penal system attracted nationwide attention during the 1960s, with allegations of brutality and murder at the prison work farms. Shocked, the state set about to remedy the worst abuses. The elected officials of the 75 counties include judge (the chief executive), clerk, treasurer, sheriff, and collector or surveyor or both. Elected justices of the peace compose a Quorum Court, which serves as an advisory body to the judge and exercises some legislative functions. There are numerous local improvement districts, and, of course, school districts. Although a number of incorporated cities have a city-manager form of government, the traditional mayor–council form is most common.

The Democratic Party has dominated political activity since Reconstruction, with the notable exception of the Republican Winthrop Rockefeller's terms as governor from 1967 to 1971. In recent years the Republicans have begun holding nominating primaries. Unless a candidate receives a majority of votes cast in a preferential primary, a runoff is required. Permanent voter registration replaced the poll tax in 1965.

Political life

The social milieu. Wage levels of Arkansas's workers are among the lowest in the nation, and living costs ap-

proximate those of the south central region. Blacks continue to live at distinctly lower economic and social levels despite an improved climate in the 1960s, and they emigrate in substantial numbers.

Status of race relations

Civil rights movements. Public-school desegregation has been the primary target of civil rights movements in recent years. Attempts to alter discrimination in housing and employment have been less dramatic. The demonstration against school integration in Little Rock in 1957 was a momentary setback; school integration has subsequently gone ahead with little difficulty in most areas. Opposition has been strongest in eastern and southern Arkansas, where blacks make up a large percentage of some communities. Many private schools have been established to circumvent integration.

Legalized discrimination through the Jim Crow laws no longer exists. Under Governor Rockefeller many blacks received appointive office or were sought out for employment by state agencies. Blacks seek election to the state and local legislatures in increasing numbers. A major problem of the black Arkansans, like that of blacks nationwide, is an economic one: to obtain and hold the better-paying jobs that require both education and skill.

Education. The public school system functions under a State Department of Education and district school boards. Specialized institutions include schools for the deaf and the blind. Two children's colonies for the retarded have attracted nationwide acclaim in the field.

In addition to its main campus in Fayetteville, the University of Arkansas operates in Little Rock a separate university as well as medical, technological, legal, and social-work centres. Several other state colleges and a university developed from institutions for the agricultural and mechanical sciences and for teacher training. The national trend toward establishing a statewide system of two-year colleges has not been strong in Arkansas.

Health and welfare. Problems of malaria, pellagra, and pinworm that once plagued the region have been virtually wiped out by widespread efforts of state and local health authorities. The state health department and welfare commission administer many programs funded in part by the federal government. Emigration of young people over the past two or three decades has aggravated health and welfare programs, especially in declining rural areas. Welfare payments are among the lowest in the nation. The mild climate and attractive scenery has fostered the establishment of retirement villages in the Ozarks.

CULTURAL LIFE AND INSTITUTIONS

Eastern Arkansas is typically Southern in speech pattern and customs, and a leisurely pace prevails. Central Arkansas reflects still its Southern heritage, but in speech and manners it is also Middle Western. The rural areas of the Ouachitas and Ozarks have retained to the fullest degree an unchanged culture.

The arts. The several symphony orchestras and community theatres and the civic ballet in Little Rock are not of a professional level. Most colleges and universities offer training and performance in these arts. A four-state opera workshop is held each summer in the Ozarks.

Folk arts and culture

Arkansas's richest contributions are in the folk arts of the Ozarks. A major folk-art centre in Mountain View has been designed to provide a showcase for local and visiting performers in dance and music, to preserve traditional skills in ceramics, jewelry, wood carving, hooked rugs, and basketry, and to offer instruction in the native folk arts. Other aspects of folk culture include the gospel singing of rural areas. Black spirituals and "soul music" flourished in Arkansas long before they became popular nationally.

Preservations. The University of Arkansas has a fine collection of archaeological and mineral artifacts. A collection of Early American glassware is featured in the Museum of Science and Natural History, housed in the old federal arsenal in Little Rock. Historic sites include Arkansas Post, the first European settlement in French Louisiana; Washington, the Confederate state capital during the Civil War; and the Territorial Restoration and First State Capitol in Little Rock.

Communications. Little Rock has the oldest newspaper west of the Mississippi, The *Arkansas Gazette*, founded in 1819; like the *Arkansas Democrat*, it serves the entire state. Central Arkansas is served by radio and television affiliates of the major networks. The entire state is blanketed by commercial radio stations, and all major urban centres are able to receive one or more television stations. Cable serves the more isolated communities. Educational television is received only in central Arkansas.

Prospects. Arkansas has important natural and human resources that have yet to be utilized effectively. Progressive and enlightened leadership often has been lacking. The late 1960s and early 1970s saw some improvements in prisons, governmental planning, tax structures, race relations, and educational and welfare services. Such projects as the Arkansas River navigation program may well attract new industries, expand existing ones, and offer all Arkansans brighter economic prospects.

BIBLIOGRAPHY. ARKANSAS HISTORICAL ASSOCIATION, *Arkansas Historical Quarterly* (since 1942); JOHN GOULD FLETCHER, *Arkansas* (1947), one of the best single volumes about the state; WRITERS' PROGRAM, ARKANSAS, *Arkansas: A Guide to the State* (1941), although dated, still a valuable compendium; HENRY M. ALEXANDER, *Government in Arkansas*, 5th ed. (1963), a comprehensive treatise; FRED W. ALLSOPP, *Folklore of Romantic Arkansas* (1931); MARGARET ROSS, *Arkansas Gazette: The Early Years, 1819–1866* (1969), an excellent account of an old and respected newspaper; ORVILLE W. TAYLOR, *Negro Slavery in Arkansas* (1958); DAVID Y. THOMAS, *Arkansas in War and Reconstruction, 1861–1874* (1926); UNIVERSITY OF ARKANSAS, *Arkansas Business and Economic Review*, a quarterly publication containing economic data.

(B.A.D.)

Armed Forces

In 1970 approximately 23,000,000 men were under arms throughout the world, and it was estimated that almost 60,000,000 civilians were engaged in economic activity related to military purposes (in Europe, the figure was about 10 percent of the economically active population).

The total cost of military preparations exceeded $200,000,000,000 in 1970. The significance of the armed forces may be measured in a different way: armies and navies are the largest organizations in the contemporary world, as well as the most technically developed. In many countries they form the strongest political force, establishing governments and controlling their policies.

Over the centuries the armed forces have developed a large intellectual body of military science and history. To this have been added more recently the studies of the military by psychologists, who have examined the internal problems of military organizations and the relations between the organization and its members. Sociologists have studied the military as a social organization, and political scientists have examined the relations between military and civil authorities. This article, which covers all of these areas, is divided into the following sections:

I. General characteristics of the armed forces
- The military system
- The military bureaucracy
- The professional soldier
- Patterns of organization
- The military community

II. Organizational problems
- Manpower and expenditures
- Recruitment and social control
- Function and organization

III. Historical trends
- Historical parallels
- Medieval soldiers
- The armies of the kings
- The national army and the professional soldier

IV. Armed forces and society
- Civil–military relations
- Military intervention in politics
- The future of the armed forces

I. General characteristics of the armed forces

Although armed forces are as a rule national institutions, the forces of different nations are very similar in struc-

Common characteristics of armed forces

ture and function. They all tend to have the same values and norms of behaviour; they also resemble each other in rank system and organization pattern, technology and training programs, and rituals and life-styles.

One reason for this lies in the similarity of their function. In principle, all armed forces are in competition with each other, either in preparation for actual combat or in that modern form of military struggle known as an arms race. The competition extends to more than weapons; personnel and morale are also involved. Continual competition tends to make the armed forces of different countries increasingly alike.

Another reason for their similarity lies in their historical development. During the period following the Middle Ages, a type of military and naval organization came into being and has since become the standard model. While numerous variants have been developed and many innovations have risen from the technicalities of weapons (the most important of them being air forces), the basic structure of the military has remained essentially unchanged. The model was exported throughout the world by Europeans through their colonial armies and later was adopted by the armed forces of new nations.

Thus the armed forces are not only national instruments but an international type of social institution and may be studied on a comparative basis more readily than other social institutions.

THE MILITARY SYSTEM

The management of violence. The principal purpose of armed forces is, of course, the management and carrying out of violence; their effectiveness depends almost entirely upon their success in doing this. But modern armed forces have become complex and include not only combat groups but various support and service units with tasks that range from planning to transportation and from the improvement of morale to the cultivation of relations with the public. Perhaps an extreme has been reached in the U.S. Army, in which no more than 15 percent of the uniformed personnel have ground combat functions and the other 85 percent consist of mechanics and repairmen, craftsmen, administrative and clerical workers, electronic specialists, and various service personnel. The percentage of combat troops is even smaller in the U.S. Air Force. This is the culmination of a trend that began early in the century.

During the Civil War in the 1860s, about 90 percent of the lower army positions in the U.S. could be called strictly military, but the proportion diminished to 35 or 40 percent during World War I. A similar trend occurred in the U.S. Navy: whereas in 1800 about 90 percent of a frigate's deckhands, or almost its complete crew, were used for fighting and operating the ship, after 1880, when the steamship came into use, this percentage fell to 55.

Military forces are distinguished from other armed groups such as brigands or bands of rioters by the fact that their functions are usually carried out in the name of a legal authority. There are obvious exceptions to this. Armed revolts and guerrilla warfare represent illegal forms of military activity. If the rebels are successful, however, they are increasingly recognized as legitimate, while their opponents will come to be called counterrevolutionaries and terrorists.

Armed forces differ from police in that their operations are directed against the armed forces of other political communities, usually countries, but here, again, the lines cannot always be sharply drawn. In many states the police have paramilitary units that function together with the army in suppressing rebels and pacifying restive areas. These operations sometimes take on a semimilitary character. On the other hand, in a number of countries with military regimes, the army is not primarily an apparatus to be used against other countries but an internal factor of power employed by politicians and military leaders in the domestic political struggle.

Three criteria of armed forces

In general, however, armed forces are characterized by all three of the foregoing criteria: the collective exercise of violence, in the name of a legal authority, directed against other political societies.

Exceptional role of the military. Practically all countries have armed forces. In one case only (Tanzania after its proclamation of independence in the early 1960s) has a modern state hesitated for a moment in providing for military units. A modern state must have an army just as it has a post and telegraph service.

In modern times, nevertheless, armies are much less active than they have been in other historical periods when conflicts were more frequent and the military virtues more highly regarded. The military are now used only as a last resort, after other means have failed. War is felt to be a catastrophe for civilization.

The gap between social and military morale has seldom been wider than at present, and this has two far-reaching consequences for the armed forces and their social role. In the first place, they have an exceptional status in comparison with other social organizations, existing in a state of readiness but seldom being used. Their primary task is to keep themselves up-to-date with developments in military technology, strategy, and tactics. From the standpoint of psychology and morale, this means that in time of peace the armed forces are not much more than an enormous training and maintenance system oriented toward events that may not occur for generations. The prestige of military leaders, the morale of their troops, and the financial resources of the military apparatus may be difficult to maintain under these circumstances.

An allied problem arises in the transition from a peacetime army to a wartime army, which may be accompanied by enormous losses. Defense preparations carried out over many years and at great cost may fail completely, as in the classic case of France's Maginot Line in 1940.

A second consequence of the exceptional nature of military action is that the military institution becomes isolated within society. During war the isolation is lessened as civil society adopts and supports the values and aims of the armed forces. When peace returns the military system again becomes a painful symbol of collective violence. Since World War II, however, the role of the military has been less exceptional than it was in previous generations. Because of their use in international politics, the armed forces have become in some respects an essential part of the political order.

The armed forces as a formal organization. For all of the romance, tradition, and ceremony associated with military life, military institutions are rational and artificial constructions, quite consciously designed and developed for their purposes. In this respect they differ from other social organizations such as families or business enterprises. Their artificial character makes possible an extreme formalization and standardization of structure and function. Even early forms of military organization such as ancient and primitive armies have placed an emphasis on habit and routine. In large armed forces this formalizing has led to something that seems at first sight contradictory to the heroic and danger-loving tradition of the military craft: bureaucracy. Large armies were the earliest large social organizations to adopt bureaucratic ways, with their striving for calculability and predictability as norms of human behaviour. This formalizing of behaviour, subsequently adopted in many other large organizations such as government and industrial enterprises, goes further in the armed forces than elsewhere. All behaviour is exactly prescribed and standardized in manuals, from the writing of a letter to the greeting of comrades, from methods of bayonet fighting to the conduct of funeral ceremonies. In time of peace the ritualism and ceremonialism are often given an exaggerated and superfluous emphasis. The justification of such extreme regulation of behaviour lies in the necessity of working under stress and of being able to maintain a functioning organization even if the personnel are suddenly replaced. At the same time, there is a danger that those in command may lose the capacity to be flexible and to innovate. To prevent this, a military elite has to be placed above the managerial-bureaucratic level to control policy.

The military is more than a specialized organization: it contains within it almost all of the functions of civil so-

The military as a total organization

ciety, with the difference that in the military the separate, relatively autonomous institutions of civil life are combined into one large institution, a "total institution." A large military camp or base is for thousands of persons at the same time a restaurant, hotel, school, place of work, hospital, transportation system, repair station, and post office. The military establishment as a whole also comprises research institutions, administrative staffs, training grounds, warehouses, and even factories. The military world forms a kind of subsociety, making itself as independent as possible within the total society. In addition to being self-sufficient, it must also be mobile; it has to be able to cross oceans and to populate deserts. This comprehensive character of the armed forces is symbolically accentuated by the soldier's uniform, by the geographic isolation of military installations, and by the norm that members are technically in service 24 hours a day.

In recent decades, in some countries, the institutional claims of the military apparatus have been weakened. Relationships have become closer with the surrounding society, enlisted men more often take their furloughs at home, and terms of service have been shortened. The professional soldier of the military academy is less inclined to regard his career as a lifelong commitment, the more so as the military profession itself has come to resemble certain civil professions. But these and other changes have not gone so far as to alter the fundamental characteristics described above.

THE MILITARY BUREAUCRACY

Plural authority. The structure of power and authority in military organizations resembles that in other large organizations but with marked differences of degree. As in any bureaucracy, the hierarchic principle prevails: the office guarantees the authority of the man who, once appointed, enjoys all the authority connected with his position without regard to his personal characteristics. These offices are mutually related through quite detailed organization charts and managed with the aid of equally minute and elaborate organization manuals. The structuring of authority according to function is accompanied by very pronounced differentiations in rank. "An order is an order" means that any superior can give orders to any subordinate, even if it runs counter to or outside of the functional division of labour. The two sources of authority—function and rank—sometimes conflict with each other, in that men may be forced to choose between obeying regulations and following orders. Although this conflict exists in all large organizations, it is extremely severe in the armed forces. Business organizations usually give precedence to rank. Nonmilitary government organizations generally stick to the regulations. In military bureaucracies, however, rank and regulations are held to be equally important.

Authority in the military environment has, however, a third basis: the attained result. In time of war, particularly, the aims have to be attained at all costs. Consequently it is seldom useful to observe all the rules, and neither is it always best to follow only the formal lines of command. Informal hierarchies arise, determined by their success in action. Success may be rewarded later with citations, promotions, and important assignments.

Success, however, is not always easy to judge. Lower field officers have to take into account the expectations of their subordinates, who tend to measure successful leadership by the extent to which it reduces the casualty rate. Higher commanders have other criteria, such as occupying strategic points or wiping out enemy positions.

Modern peacetime armies tend to be bureaucratic, with an emphasis on regulations. Rank is emphasized by combat units in peacetime but by noncombat sectors only in wartime. The third basis of authority, results attained, becomes the ultimate one under conditions of wartime combat.

The armed forces of different countries vary in the types of authority to which they give most emphasis. Complex modern military forces tend to stress regulations. Less complex ones emphasize rank. For elite formations and for critical operations, a combination of rank and performance will prevail. But the social environment also plays a part. Armed forces that are very strong on rank will try to reduce the influence of the social environment by staffing themselves with volunteers or professionals rather than conscripts, unless they are functioning within a social order that accepts extreme emphasis on rank (as in Prussia). Armies with a citizen tradition (Switzerland, Canada, and Australia) will emphasize the ultimate result of the operations as a first standard for good military leadership.

Ascription and achievement. The distinction often made between social positions that are acquired and those that are taken without effort (for instance, through right of birth) can be applied to the sources of military authority.

The decline of class differences

In European countries membership in the officer corps was traditionally reserved to sons of the nobility. When commoners became officers, they were made peers. The tradition survives in the expressions "an officer is a gentleman," and *das Portepee adelt* ("the Portepee ennobles"—the *Portepee* is an ornamental tassel on the sabre). The rank and file, on the other hand, were recruited from the lowest and poorest categories of the population. Soldiers and sailors, particularly in the time before national armies and navies, had a very low occupational status.

By the latter part of the 20th century, the ascriptive base of military authority had been greatly diminished. The contemporary national army draws its officers as well as its enlisted men from all strata of society. The profession of officer has tended to become a middle class occupation, while the social and educational level of the troops has gone up. Class differences still persist: in the United States around half the officers come from white-collar families and a third from blue-collar families, while less than a fifth of the enlisted men come from white-collar backgrounds and over half from blue-collar families. But this is of secondary importance in the authority structure. Other important ascriptive sources of rank authority remain; for example, in racially divided societies in which black people are less frequently admitted to the higher ranks.

The caste system in the armed forces has not disappeared with the decline of class differences. The traditionally great distance between officers and men has become more self-sustaining; it is fixed in social intercourse, ritual, and language. Even in countries in which the social prestige of the officer is relatively low, as in many west European countries, his authority within the rank system has been little affected. Even the noncommissioned officer does not participate in this authority. The caste barrier makes it extremely difficult for noncommissioned officers to achieve promotion to commissioned officer rank. Even promotions on the battlefield are exceptional. It is quite common for very young, inexperienced officers to receive command over much older and experienced noncommissioned officers.

Generalists and specialists. A more important distinction in contemporary armed forces is that between generalists and specialists.

Specialization had its start at the periphery of the military system. Engineering and artillery were the first to develop special bodies of knowledge requiring formal training. The older sections of the military establishment resisted the idea of academies for officers until well into the 19th century, partly because officers came from the educated classes and partly because of the belief that leaders are born and not made.

The notion that military men should be generalists rather than specialists is still strong, although most officers attend a variety of schools and training courses during their professional careers. The practice of rotating officers among jobs is more prevalent in military than in civilian life. Table 1 compares the lengths of stay of officers at a U.S. air base and of the managers of a very large U.S. business corporation. Almost half of the officers had been stationed for one year or less on their base, while only 2 percent of the business managers had

Table 1: Length of Stay of Officers and Business Executives Within a Single Organization

length of time in the organization	military installation (percentage)	business firm (percentage)
1 year or less	45.5	2.1
2 years	18.8	2.0
3 years	15.7	3.9
4 years	14.4	4.6
5 years or more	5.6	87.5
Total	100.0	100.1
Number of cases	554*	1,219†

*Two cases were not ascertained. †21 cases were not ascertained.
Source: Oscar Grusky, "The Effects of Succession; A Comparative Study of Military and Business Organization," in Morris Janowitz, *The New Military; Changing Patterns of Organization* (1964).

been with their company for such a short period. Nine out of ten business managers had been at least five years with their company, while only one out of 20 officers had been on their base for that length of time. The same pattern prevails in other armed forces. While there are important differences between an air base and a business firm, the foregoing comparison shows the emphasis placed on mobility of personnel.

The general staff

The military solution to the problem of specialization was the development of the concept of the military staff. A staff consists of advisers and planners who assist the commander in issuing orders but do not have line or command authority; the commander alone is responsible and cannot delegate his responsibility. A distinction is often made between the general staff that is concerned with such areas as personnel, intelligence, operations and training, or logistics, and other staff groups specialized according to weapons and services. The latter include many more specialists than does the general staff, which is often populated with generalists who happen to be expert in particular problems. Naval staffs are of more recent origin than army staffs and have not acquired the size and importance of the former.

As weapons technology has grown more complex, the staff system has extended downward to lower levels of command. More and more specialists have been attached to line organizations, with the result that the old pyramidal structure of the armed forces has changed to a diamond-shaped structure, particularly in technical and maintenance units. Before World War II, 50 percent of the enlisted men in the U.S. Army and Air Force were in the lowest of the seven grades; after 1945, the percentage was less than 15.

Another result of technological change is a growth in the number of officers and noncommissioned officers who owe their rank to their technical skills rather than to the number of subordinates they command. The growth in size and complexity of the armed forces has produced a continuously expanding military bureaucracy. More and more officers work in offices at great distances from the troops. The moment when a professional soldier gives up active command of troops and takes a desk position in the higher echelons comes earlier and earlier in his career. In the ground forces particularly, the majority of the career officers work in high command and staff positions during wartime, while in many cases the combat forces are led predominantly by reserve officers and noncoms or even by men who have been conscripted.

Commanders and leaders

A distinction may be made between command and leadership—or between executive leadership and operational leadership. The act of commanding has an upward orientation; it is the carrying out of a general plan. The act of leading, on the other hand, points downward toward the people by whom the plan is to be carried out. The commander is essentially a coordinator who has to think in abstract terms while working with the aid of numbers and charts, while the leader must deal directly with people and induce them to act successfully in combat situations. The commander, though shielded from the hazards of battle, is exposed to types of uncertainty quite beyond the personal experiences of the leader. The tensions to which the leader is exposed are those of men in danger, facing immediate physical loss.

At some levels in the organization both roles are equally important. But in most cases the accent falls on one or the other, and, generally speaking, the senior officer acts as commander while the junior fulfills the role of leader. The military leader participates in the life and dangers of his unit as the person responsible for the welfare of his people as well as for the execution of assignments from the top.

The task of the leader is especially aggravated by two circumstances. Modern war has become, much more than the combat of earlier times, a matter of improvisation requiring initiative at the lowest level; there is no longer the blind obedience to orders of earlier times or troop movements directed en masse from the top; the junior officer has to make decisions himself. In the second place, authority in small combat units is no longer simply a matter of the leader dominating his subordinates but requires an ability to deal with and motivate men, to inspire confidence and loyalty.

Leadership is a difficult undertaking for young officers and noncommissioned officers, the more so when their advantages in age, education, and social class are often quite small. Modern military schools accordingly emphasize courses in leadership training.

THE PROFESSIONAL SOLDIER

The military mind. Much has been written about "the military mind," especially by critics of the military's performance. Although some of this is merely captious, one can discern among military men certain characteristic habits of thought. The soldier's occupation tends to lead to a less idealistic and more pessimistic view of man; he comes to be skeptical of words and promises and to place certainty only in power and the means of power. Power is for him the ultimate guarantee of the social and political order. In this respect the military man may not, as a rule, share the optimistic social philosophy of liberalism; his views may differ from those of liberalism in their emphasis on discipline and leadership. He may have less confidence in the development of the individual, seeing man as weak and irrational and in need of organization. His concern with order, however, may not make the soldier a natural supporter of extreme right-wing philosophies. He tends to distrust ideology and romanticism and notions of racial or national superiority.

The soldier's realism and pessimism

The realistic and conservative character of the military mind can be explained in several ways. One reason for it may be the hazardous character of the military enterprise, in which naked power is the main guarantee of success. In a world of international strife, the soldier is forced to see every neighbour as a potential enemy. His enormous responsibilities make him a distrustful and realistic thinker. Because it is difficult to estimate the character and weight of future risks, the soldier will always overestimate them. His desire for ever stronger forces and more potent weapons is perhaps given stimulus by the effect these have on his own prestige and chances of promotion.

A second cause of military conservatism is the soldier's traditional position among the social elite. Although these bonds have weakened in recent decades, the professional soldier tends to identify with the social order whose sword he wields. In many new countries in which rapid social and economic changes offer opportunities for careers, soldiers may side with progressives who aim to accelerate their countries' development. But they will usually reject extreme leftist tendencies, as has been shown, for example, by the military regimes of the Arab countries despite strong support from the Soviet Union.

A third explanation of the soldier's viewpoint is that he tends to have an authoritarian personality. It has been shown that this type of personality is more common in armed forces and police organizations than in other professions, probably because it is selected in the recruiting

process. The pattern may change in the future, however, given the gradual change in the style of military leadership and in the general requirements for a military career. It is also notable that there are important differences in attitude and outlook among the various services (the infantry, for example, being more traditional in outlook than the technologically oriented air force).

Comparison of the military profession with others

Military professionalism. The phrase professional soldier has traditionally been applied to soldiers who pursue a lifetime career with the armed forces. In the 19th century, particularly among officers, the professional soldier developed into a technician in the management of violence. Professionals came to form the backbone of all the armed services. But the phrase professional soldier is also used to mean that the career soldier has characteristics in common with those of the doctor, the teacher, the architect, and the engineer. The characteristics of such professionalism include (1) the monopoly of a well-integrated body of knowledge and skill imparted through formal training, (2) the capacity to apply this knowledge and skill according to ethical standards as a socially valued service, and (3) the legal establishment of the profession, including the right to control the admission, selection, and training of new members. While military officers share these characteristics in some degree, the officer as a professional figure almost always finds himself being used as an instrument of power by his government; he is also lodged in a hierarchy of command that prevents him from making independent decisions.

The degree to which the military services have been professionalized varies among countries. In Communist countries the officer is not only a military specialist but also part of the political elite. In some of the new countries, nonprofessional considerations such as ethnic origin, tribal relationship, class background, and political adherence are often the main factors in a military career. There is a parallel here with several European countries of the 19th century, where aristocratic origin and family ties were important; not until the end of that century did the profession of officer in Germany and Great Britain become accessible on the basis of competitive examinations open to the majority of the population.

Crisis of the military profession

Since World War II, military professionalism has been confronted with challenges from several directions. Technological change has created a number of new fields of specialization that differ radically from the traditional military skills. These technicians, logisticians, comptrollers, management specialists, and personnel administrators are not trained in the way that combat infantry officers and aircraft pilots are, and they tend to have more in common with their civilian counterparts than with their military colleagues. This means that the professional structure is becoming a congeries of diverse functions without much unity. The development of new weapons, particularly long-range ones, has undermined the old professional norms of military behaviour. The aerial bombardment of civilian populations and some of the methods of counterintelligence units are difficult to square with the traditional code of military conduct.

Another influence at work upon the profession is the politicization of the military through its involvement in the Cold War. The creation of multinational forces such as those of NATO in western Europe and the Warsaw Pact in eastern Europe requires political orientation among their officers; patriotism as the basis of morale tends to be replaced by a common ideology.

Along with this there has been an increase in civilian influence in military affairs, partly because of the new technical personnel referred to above and partly because of the so-called think tanks and strategic research institutes that have sprung up all over the world, manned by civilian experts. Important developments in military- and political-strategic thinking no longer originate with the professional soldier.

The increase in the proportion of noncombat troops in contemporary armed forces has had a subtle effect on the outlook of the officer corps: many officers are inclined to put intellectual standards above fighting spirit or to prefer managerial qualities to traditional military qualities. Some experts have even suggested that the military profession model itself upon civilian service institutions. This is not usually taken to be a viable solution, however, because it is felt that, while it might help to offset the unpopularity of the modern military apparatus by minimizing its destructive and violent aspects, it would at the same time be likely to increase the frustration of many career soldiers. Although some marginal functions of the military can very well be taken over by civilian authorities, the main task of the armed forces remains unique.

Middle class status

Social position of the professional soldier. The military profession is a middle class one. Its status has declined since the 19th century, when in some countries officers belonged to the social elite; today, in the developed countries, career officers are regarded as standing socially below the academic professions and quite often below public school teachers. In many developing countries, on the other hand, officers are highly esteemed because of their superior education and their competence in modern forms of activity; the profession in these countries attracts some of the most capable men. In Communist countries the position of the officer corresponds to what it was in western Europe in the 19th century: aloof from civil society, respected, seldom criticized, and sharing in the prestige of the regime.

The decline in status of the military in Western countries is reflected in the social class from which the majority of officers are drawn. Since World War II, in western Europe, officers have been mainly of middle class or even lower class origin. In the United States in 1910–20, one-quarter of the general officers were from the upper class, and two-thirds were from the upper middle class; in 1950, only 3 percent came from the upper class, while half came from the lower middle and lower classes.

The change has been greatest in the Communist countries, where the old aristocratic officers corps were eliminated and replaced by officers of very predominantly proletarian origin. In the U.S.S.R. after the 1917 Revolution, many former tsarist officers continued for some time to hold important positions. In the German Democratic Republic, Poland, Czechoslovakia, and other east European countries, however, where the Communists took power after World War II, the transition was so abrupt that nontrained proletarian officers were chosen above others who were thought to be less reliable. According to official data, in 1965 more than 80 percent of the officers in the German Democratic Republic, 60 percent in Czechoslovakia, and about 50 percent in Poland came from the working class. In the developing nations the majority of the officers are usually of the middle and lower middle class.

Many officers come from military families, although it is not clear that this tendency to follow the father's occupation is more pronounced among military men than in civilian professions. The proportion of officers coming from officers' families has ranged in recent decades from 10 to 50 percent, the highest percentages occurring in France, Spain, and Brazil. Prestigious military schools such as West Point, Sandhurst, and Saint-Cyr attract many officers' sons. Twenty-five percent of the 1960 class at West Point was composed of officers' sons.

PATTERNS OF ORGANIZATION

Armies vary greatly in the way they are organized and structured. Many are so loosely organized as to resemble social movements, as did the crusaders, the Teutonic Knights, and many revolutionary armies. Some armies combine mass enthusiasm for an ideal with a high degree of formal organization. Others show considerable egalitarianism—for example, the European mercenary armies of the 16th and 17th centuries, the guilds of the early artillerists, and the soldiers' councils of the 20th-century Russian and Hungarian revolutions.

The military emphasis on organization

As a rule, however, armed forces are the most strictly organized groups in existence, drawing much of their strength from the thoroughness with which they regulate the behaviour of their members. In this respect the military organization has served as a model for other fields.

Many of the early industrial principles or organization, such as standardization, line and staff, and training of personnel, were derived from the military.

Functions of organization. The main functions of organization are to reduce complexity and to cope with changing reality. The need for organization in the military field is particularly evident, since the degree of uncertainty confronting the soldier is so great.

Military organization has passed through two stages in its development. The first was one of simple organization, in which the emphasis was put on reducing everything to fundamentals. By standardizing actions (drill, routine, ritual), homogenizing elements (weapons, formations), and directing movements from a central point (through a hierarchy), a social machine was constructed that was well integrated internally and capable of absorbing external shocks. This was the system applied in the Roman legions, the European armies of the 16th century, the early naval formations, and many other armed forces before the 19th century. The drawback of the system was its rigidity.

The second stage emphasized flexibility. Military formations were decentralized, and authority was delegated to a larger number of officers who were expected to demonstrate more initiative. The new model had less internal cohesion but was more flexible in response to external challenges. This development has continued up to the present. The decentralization that began at the top and was carried down through the senior commanders to junior officers has now reached the smallest combat units (guerrilla warfare).

In peacetime the armed forces tend to fall back upon the simple system of organization. With the coming of war this structure collapses, and the organization must with great effort be transformed into one capable of responding to the exigencies of wartime. In early times this was done by concentrating the forces in a relatively small area, such as a battlefield or fleet formation, where they could easily be commanded; the early battlefield formations were compact and could readily be integrated as fighting forces. In modern times the task is more complex. Greatly extended formations are now the rule, and they require a network of long-distance communications to handle the continuous flow of information among their subdivisions. This also applies to the interaction between the armed forces as an organization and the social environment: the allocation of resources to the military has become largely a problem of communication; the carrying out of such functions as supply and transportation have become less physical and more mental than in former days.

But more than formal organization and coordination are required to keep modern armed forces in readiness. It is no longer sufficient to allot weapons and functions to military personnel and place them under a leader; the men must also be made to realize the meaning of their task as part of the larger whole.

Levels of responsibility

Organization and operation. In principle, every armed force has three levels of responsibility and control: political, professional, and technical. The political level is generally the uppermost, concerned with overall political and strategic matters. The technical level is at the bottom, where actual operations are carried on. In between is the professional level, in which the majority of officers do their work of management.

The centre of gravity in this organizational pyramid will vary, depending on circumstances. In matters involving nuclear weapons and strategic air operations, political as well as military men participate, and sometimes the head of state himself assumes responsibility. In guerrilla warfare the centre of gravity is very low, with important responsibility borne even below the level of the junior officers. This responsibility encompasses not only combat but also supply and intelligence tasks; the junior commander may even have to exercise political authority in areas in which his unit moves. In conventional warfare the centre of gravity lies in the middle of the organization, with the professional soldier; most of the preparation, planning, and organization, as well as control over the results, is in the hands of those at the managerial level.

Differences in outlook naturally arise among those engaged at different levels of the military operation. Those concerned with problems of nuclear weapons have scientific habits of mind, placing education and rational calculation above arguments from historical experience. Those in charge of special operations and counterinsurgent warfare think in terms of physical courage, group loyalty, and personal leadership. Many officers find themselves caught between these two types, disliking the new technocrats who differ so much from the traditional military type and at the same time feeling threatened by those on the lower level who regard the career soldier as obsolete. Tensions of this kind were visible in the revolt of the French officers in Algeria in 1958 and 1961.

Informal organization

Considerable research has been done on the primary (face-to-face) group and its functions in military life. It would appear that men tend to identify with their platoon, crew, or squad rather than with the larger organization; it is the world in which they are at home and within which almost all their interests are rooted; they find security in the small group, a sense of solidarity vis-à-vis the military hierarchy, a chance for prestige among their fellows, and norms of behaviour that are difficult to ignore. Under conditions of combat, these informal relations are greatly enhanced. Studies of German and U.S. units have shown that soldiers are motivated not by ideology or political purpose but by a sense of personal commitment to their comrades: "You can't let the guys down." The character of these bonds depends partly upon circumstances. Ground combat units draw their cohesion from the peril to the group in battle, while submarine or bomber crews relate to each other much more as specialists who must execute a mission together.

The existence of these primary groups often poses difficult problems for the commanders—particularly the noncommissioned officers, who are most exposed to pressure from a group and have the least prestige at their disposal. In practice they often follow the wishes of the group: their so-called leadership is in part a product of the group they lead. But this group behaviour has its advantages from the viewpoint of the top command. The small group at the base serves as a self-regulating system that can adapt to unforeseen circumstances. For this reason it appears to pay to keep the informal system intact, even to the extent of carrying out personnel replacements in a way that will not destroy particular groups.

Informal organization is not a phenomenon of the base level alone but runs through the whole military organization. Social and political groups and cliques play a decisive role at the highest levels. Their existence, partly concealed in many countries, is obvious in new countries, where a top military position is clearly a political acquisition. Military coups in such countries show the political leanings among the top leaders of the military apparatus, which vary from the right (Greece in 1967) to the left (The Sudan in 1971) and may sometimes divide the military leaders themselves (Indonesia in 1965).

THE MILITARY COMMUNITY

Styles of military life

The armed forces are much more than an organization; they form a community that encompasses the work and life of its members much more completely than do most other social organizations. The life of the professional soldier, however, is quite different from that of the enlisted man. The first, particularly if he is a member of the officers corps, experiences military service as his career; his manner and style of life are often set for him by tradition. The enlisted man, on the other hand, knows the military environment as a male community full of deprivations and rough behaviour.

The elitist style of life. The officers' milieu is characterized to some extent by a ceremonialism that has become passé among other social classes, a heritage of officers corps of the 18th and 19th centuries. This formal "old-fashioned" style is more than the pretension of belonging to an aristocracy: it has certain functional uses in that it helps to facilitate the integration of the members

into military life. The professional soldier is, after all, highly mobile both geographically and socially, continually forced to adjust to new milieus. His adjustment and assimilation are made easier if these milieus all observe the same formal patterns of behaviour. It is the more mobile and isolated parts of the armed forces that have most strongly developed their own styles of life, notably the navy and the old colonial armies.

The officer's wife plays a more important part in his professional life than is the case in most other professions. To some extent this belongs to the traditional style, which makes high demands upon etiquette and social contact. It is also a consequence of the isolation of military life in small communities in which work and residence coincide.

The enlisted man's style. The community life of enlisted men is very different from that of the officers. As an exclusively male community, separated from their families and often unmarried, their social traits resemble those of colonists, woodcutters, sailors, and dock-workers. While officers have an elitist subculture, that of the enlisted men reflects their position at the bottom of society. While the officer attaches an almost exaggerated importance to form, the private most likely will reject all form. The cool distance that characterizes the relationships of officers contrasts with the close ties that prevail among enlisted men.

The life of the enlisted man is reflected in his language —earthy, masculine, and aggressive, frequently with a heavy emphasis on sexuality. Much of his slang expresses the repugnance he feels toward the military hierarchy, for to him it seems that the officers make the mistakes and reap the rewards, while it is he who does the dirty work and wins the battles.

Enlisted men stationed in foreign countries are notoriously promiscuous in their sexual behaviour. In 1955 it was found that 59 percent of the U.S. enlisted men stationed in Japan had regular intercourse with prostitutes; in Korea in 1965 the percentage was 88.

II. Organizational problems

MANPOWER AND EXPENDITURES

The weight of the military effort

The power potential. The size and power of any military force are ultimately limited by a country's ability to sustain them. In the last analysis this depends on the number of men who can be mustered in time of war and on the country's economic strength. In France after 1871 there was much concern over the country's falling birth rate: while population size had not been important in the era of small professional armies, it became so low that France was competing with a more populous Germany in a day of total mobilization. At the outbreak of World War I in 1914, France had no more than 8,000,000 men of military age, while Germany had 15,000,000.

Military power in the contemporary world has to be defined as including potential power; it is the total ability of a nation or a coalition of nations to coerce other nations through the employment of military means or to resist such coercion by other nations. This means not only the ability to mobilize forces for immediate action but the ability to produce additional forces. Potential military power is, of course, relative. An international arms race will drastically influence the relative power of each of the participants, just as disarmament will be more advantageous for some than for others.

In pre-industrial societies, potential military power was very small. Countries with agricultural economies could not support large armies, and prolonged military efforts were carried on at the expense of a functioning economy: the troops sustained themselves by plundering and destroying, which brought impoverishment, depopulation, and famine. Not until after 1800 did the European countries have an agricultural surplus sufficient to maintain large armies. Manpower was also limited by the need for manual workers; the armies of 18th-century Europe were manned largely by nonproductive classes such as nobility, paupers, and foreigners.

A country's potential military power depends on its supply of manpower, its organizational capacity, and its economic resources. Manpower is first of all a function of the size of the population. During World War II only large countries such as the United States and the Soviet Union were able to recruit more than 10,000,000 men. Mass mobilization also requires a large number of trained military reserves. One of the secrets of 19th-century Prussia's success in war was that its system of universal military service produced extensive reserves of men who, with frequent retraining exercises, maintained their military capabilities until the age of 50. During the 20th century, military reserves have grown to be several times the size of the standing armies. In the mid-1930s, when 8,000,000 men were under arms around the world, an additional 2,000,000 were on immediate reserve, and another 30,000,000 had received training and were available for call. Another factor affecting manpower is the psychological readiness known as morale—the willingness to sacrifice safety, income, and personal freedom in the service of the military effort. While this factor cannot be expressed quantitatively, it is of no less importance than numbers in estimating potential power.

A high level of organizational capability is also required for the mobilization and maintenance of armies consisting of millions of men, so much so that only developed and industrial countries are successful at it. The power to mobilize quickly is very important, as the German Army demonstrated impressively in 1914, when it brought 2,000,000 men under arms within 48 hours and sent 1,000,000 soldiers across the Belgian border within a week. By contrast, one of the most important causes of the failure of the Arab countries in their fight against the young state of Israel was their lack of organizational capacity.

Growth rates. The mercenary armies of the 16th century seldom achieved a size greater than 20,000 to 25,000 men. When national armies began to develop in the 17th century, they reached 50,000 or 60,000 men and a century later even 80,000 or 90,000. In the Napoleonic era, armies became much larger, and some battles involved hundreds of thousands of men; at the peak of the French military effort, 1,000,000 men were under arms. Armies and navies continued to increase in size during the 19th century. Around 1850, the total number of men in the European armies was estimated at about 2,500,000; around 1900, at about 3,500,000. In the 1930s, before World War II, the total number of men in the standing armies of the world had risen to almost 9,000,000. The figures for the great powers in 1937 were China, 1,600,000; the U.S.S.R., 1,300,000; Japan, 950,000; France, 825,000; Germany, 760,000; Italy, 760,000; and the British Empire, 645,000.

As armed forces increase in size, they take on different characteristics. Large armies are more "democratic" in the sense that they represent a larger cross section of the population, but they are run more firmly at the top. A paradoxical consequence is that small professional armies are sometimes more dangerous to the established order than are large national armies that are more thoroughly controlled.

Armies in the modern world

Larger armies lead to still larger military reserves. In modern industrial countries, the organized reserves are from six to ten times the size of the active forces, and, when all of the forces are mobilized, the professional personnel become only a fraction of the total. Thus, great wars are fought mainly by reserve officers and conscripts. Germany, in World War I, sent a total of 46,000 professional officers to battle, of whom 11,350 were killed, as against 226,000 reserve officers, of whom 35,500 were killed. Of the Prussian contingent of 22,112 professional officers at the beginning of August 1914, 5,633 had already been slain by November 15.

A large army constitutes a heavy burden upon a country's economy. During the Napoleonic era, as much as 5 percent of the total French population was mobilized, compared with preceding centuries, when the proportion did not rise much above 0.3 percent, or the later 19th century, when approximately 0.5 percent of the population of the leading countries was under arms. The percentage of soldiers to the total labour force was much

higher. The United States had 18.6 percent of its labour force on active service in 1945. In the postwar period the proportion fell to 3 or 4 percent.

Recent developments. In 1964 the total active military manpower of the world was estimated at 20,900,000; in 1968 it was 23,800,000—a growth in the four years of 13.6 percent. The forces of the developing countries grew by 18.4 percent over that period. The fastest growing forces were the relatively small and often new armies of Africa and Oceania (almost 50 percent), followed by Southeast Asia (35.1 percent), the United States (30.1 percent), and the Middle East (26.6 percent). Much of this growth had to do with the existing war situation in various parts of the world, particularly in Southeast Asia and the Middle East. The European countries of the North Atlantic Treaty Organization reduced the size of their armed forces by 1.2 percent in the same period. The military manpower of the Warsaw Pact nations increased by 3.6 percent, the figure for the U.S.S.R. being 6.1 percent. The West and East German armies increased by 6.1 and 11.4 percent, respectively.

Military expenditures between 1964 and 1970 showed contradictory trends. When measured in constant dollars to eliminate the effect of inflation, the expenditures of the NATO countries levelled off between 1967 and 1969 and declined in 1970. In the developing countries the trend was quite different: the level of military expenditures (in constant dollars) rose in the developing countries after 1966 at a faster rate than their gross national products. Table 2 shows that in 1968 the developing countries maintained 12,800,000 men under arms, compared with the 10,900,000 men of the richer countries. On the other hand, the military expenditures of the two groups of countries amounted to $22,000,000,000 and $169,000,000,000, respectively, showing the stress placed by the latter upon technology. In proportion to total population, the developed countries were more substantially mobilized (11 per 1,000 of the population) than the developing countries (five per 1,000).

The burden of military expenditure at the end of the 1960s was heaviest in the countries of the Middle East and Southeast Asia: it was estimated to exceed 10 percent of the gross national product in Israel, Egypt, Jordan, Iraq, Saudi Arabia, North Vietnam, South Vietnam, Laos, North Korea, and the Republic of China (Taiwan). Expenditure was between 5 and 10 percent of the gross national product in the U.S., the U.S.S.R., the People's Republic of China, Great Britain, Poland, Cuba, and Portugal.

RECRUITMENT AND SOCIAL CONTROL

The ways in which a country recruits its military personnel are likely to have important consequences both for the armed forces and for the society from which they are drawn. Soldiers may feel a commitment to patriotic or idealistic aims; they may be coerced; or they may be motivated by pecuniary gain.

Voluntary armies

Moral commitment. The committed soldier is intensely involved in his organization and its success. The organization can draw upon his loyalty and can motivate him with nonmaterial rewards—glory and honour, prestige and status, shame and humiliation. Armed forces recruited in this way take on the characteristics of social and political movements. They have relatively little of the standard military structure with its regulations and its hierarchy; those at the top are accepted as leaders in a common effort toward a common goal, which can be attained only by following and supporting them. A strong ideology helps to maintain the cohesion of the fighting forces. Such voluntary armies have included those of the First Crusade, the religious Hussite armies of the late Middle Ages, the warriors of early Islām, and national revolutionary armies such as those of France after 1789 and Russia after 1917.

The successes of these armies stemmed from their collective élan and not from their training and their technical or tactical superiority. Their lack of structure and professional leadership became a source of weakness at times of internal stress (that is, in a later period when armed struggle continued and more stability and integration was needed). Methods had to be found to give form to the collective mass.

Revolutionary armies often begin with egalitarian institutions, such as the soldiers' councils and militia units of the French, Russian, Hungarian, and Indonesian revolutions. When confronted with the orderly armies of an enemy, they tend to fall back upon traditional hierarchic forms. The initial enthusiasm of the mass diminishes, while at the same time military service becomes obligatory rather than voluntary. The revolutionary armies thus adopt a system of conscription.

Another way of maintaining cohesion is by indoctrination and political training. The young Soviet Army was reorganized along classical lines with the introduction of general conscription, but at the same time a political shadow organization was built up in it through the Communist Party. This approach has been followed in many guerrilla armies, in which the military and political elements are strongly interrelated. In this way they obtain the advantages of the military system of coercion along with those of ideological motivation.

A third method lies in stressing patriotic symbolism such as the monarchy, the fatherland, honour, and glory, supported by a good deal of ritual. The soldier is taught to forget material interests and devote himself to higher ends.

Certain mixed forms exist, two of which are the professional officers corps and the militia army. The modern officers corps is built upon the concept of a military craft, embodying values and norms formerly cultivated by the nobility. These include ideas of honour and status, along with a disdain for the materialistic concerns that dominate the rest of society. Contemporary officers' attitudes, however, are moving away from this toward the attitude of a functionary who seeks remunerative, attractive work and a permanent job.

The other mixed form, the militia, is a citizen army in which the obligation of the soldier is to defend his city or country. Sometimes a militia is raised in defense of group interests, such as those of labourers, farmers, miners, students, or a political party. The militia is a mixed form because in many cases it is combined with conscription. The phrase compulsory militia is not entirely contradictory.

The moral basis of the militia lay traditionally in the defense of home and hearth, an idea embodied in Anglo-American history by such organizations as the Home Guard and the National Guard. The militia is seldom an aggressive, conquering army since citizens in arms cannot undertake long campaigns and military expeditions.

Militias have existed throughout history since the time of the Greek city-states. Examples in modern times are the armies of Switzerland and Israel, in which the professional core is very small and civilians hold all the military positions, including high command. In both countries military service is required of all men and in Israel of women.

The militia form has certain political and economic advantages: it avoids the danger of creating a state within the state and also the high costs of maintaining a standing army. The abilities and occupational skills of the population can be placed directly at the disposal of the military, although this means withdrawing them from the civilian society. The chief objection to the militia form is its inadequacy for the building of a technically advanced and specialized army with modern air forces and artillery. Another objection is that mobilization takes too long. For these reasons, large democratic countries with a citizen-soldier tradition have developed within their military forces extensive cadres of regular soldiers.

Criminals and conscripts

Compulsory enrollment. The opposite of voluntary enlistment is conscription, carried out by the national state. The resulting military force lacks the moral characteristics of a volunteer army; it is essentially a machine requiring severe discipline, its cohesion being maintained by the threat of punishment. Its great problems, desertion and slackness among the troops, can be kept within bounds only by strong organization and leadership. Con-

Table 2: Armed Forces, 1968

	population (000,000)	armed forces (000)	forces (per 000 population)
Summary			
World total	3,509.1	23,819	7
Developed countries	967.5	10,973	11
Developing countries	2,541.6	12,846	5
North America	221.9	3,602	16
Europe	725.6	8,612	12
Developed countries†	610.6	6,986	11
Developing countries‡	115.0	1,626	14
Latin America	261.7	1,153	4
Far East	1,233.8	7,161	6
Developed countries†	101.0	250	2
Developing countries‡	1,132.7	6,911	6
South Asia	689.2	1,752	3
Middle East	91.5	993	11
Africa	270.8	449	2
Developed countries†	19.2	38	2
Developing countries‡	251.6	411	2
Oceania	14.8	97	7
By country			
World total	3,509.1	23,819	7
North America	221.9	3,602	16
United States	201.2	3,500	17
Canada	20.8	102	5
Europe	725.6	8,612	12
NATO, European	301.6	3,026	10
Belgium	9.6	99	10
Denmark	5.0	46	9
France	49.9	505	10
Germany, West and West Berlin	60.2	486	8
Greece	8.8	161	18
Iceland	0.2	—	—
Italy	52.8	440§	8
Luxembourg	0.3	1	3
The Netherlands	12.7	129	10
Norway	3.8	35	9
Portugal	9.5	183	19
Turkey	33.5	514	15
United Kingdom	55.3	427	8
NATO total	523.5	6,628	13
Warsaw Pact	340.2	4,783	14
Bulgaria	8.4	173§	21
Czechoslovakia	14.4	265§	18
Germany, East	17.1	196§	11
Hungary	10.3	137§	13
Poland	32.3	319§	10
Romania	19.7	223§	11
Soviet Union	238.0	3,470§	15
Other European	83.8	803	10
Albania	2.0	51§	26
Austria	7.4	50	7
Finland	4.7	39§	8
Ireland	2.9	10	3
Spain	32.6	305	9
Sweden	7.9	78	10
Switzerland	6.1	31	5
Yugoslavia	20.2	239§	12
Latin America	261.7	1,153	4
Argentina	23.6	144§	6
Bolivia	4.4	18	4
Brazil	80.2	225§	3
Chile	9.2	63§	7
Colombia	19.8	55§	3
Costa Rica	1.6	—	—
Cuba	7.6	394§	52
Dominican Republic	4.0	18	5
Ecuador	5.7	17	3
El Salvador	3.2	6	2
Guatemala	5.0	9	2
Guyana	0.7	1	1
Haiti	5.0	5	1
Honduras	2.5	5	2
Jamaica	1.9	2	1
Mexico	47.3	70	1
Nicaragua	1.9	6§	3
Panama	1.4	—	—
Paraguay	2.2	13	6
Peru	12.8	50	4
By country			
Latin America (cont'd)			
Trinidad and Tobago	1.0	1	1
Uruguay	2.8	13	5
Venezuela	9.7	38§	4
Far East	1,223.8	7,161	6
Burma	26.4	138	5
Cambodia	6.6	84§	13
China, People's Republic of	806.0	3,100§	4
China, Republic of	14.1	528	38
Indonesia	113.7	450§	4
Japan	101.1	250	2
Korea, North	13.4	410§	31
Korea, South	30.5	620	20
Laos	2.8	95	34
Malaysia	10.3	57§	6
Mongolia	1.2	33§	28
Philippines	35.9	47	1
Thailand	35.1	167	5
Vietnam, North	19.3	447	23
Vietnam, South	17.4	735	42
South Asia	689.2	1,752	3
Afghanistan	16.1	70	4
Ceylon (Sri Lanka)	12.0	10	1
India	527.1	1,133§	2
Nepal	10.7	15	1
Pakistan	123.4	524§	4
Middle East	91.5	993	11
Cyprus	0.6	1	2
Iran	27.1	225§	8
Iraq	8.6	92§	11
Israel	2.7	77	28
Jordan	2.1	55	26
Kuwait	0.5	7	13
Lebanon	2.7	12	4
Saudi Arabia	4.6	56§	12
Syria	5.7	69§	12
Yemen	5.0	8	2
United Arab Republic (Egypt)	31.7	391§	12
Africa	270.8	449	2
Algeria	12.9	58	4
Cameroon	5.6	5§	1
Central African Republic	1.5	1	1
Chad	3.5	1	—
Congo (Brazzaville)	0.9	2	2
Dahomey	2.6	2	1
Ethiopia	24.2	45	2
Gabon	0.5	1	2
Ghana	8.4	15	2
Guinea	3.8	5	1
Ivory Coast	4.0	4	1
Kenya	10.2	3	—
Liberia	1.1	4	4
Libya	1.8	8	4
Malagasy Republic	7.0	4	1
Malawi	4.3	1	—
Mali	4.8	4	1
Mauritania	1.1	1	1
Morocco	14.6	62§	4
Niger	3.8	1	—
Nigeria	53.5	70	1
Rhodesia	4.9	5	1
Senegal	3.8	5	1
Sierra Leone	2.5	2	1
Somalia	2.7	8	3
South Africa	19.2	38	2
The Sudan	14.8	24	2
Tanzania	12.1	4	—
Togo	1.8	1	1
Tunisia	4.7	23§	5
Uganda	8.1	6	1
Upper Volta	5.1	2	—
Zaire	17.0	31	2
Zambia	4.1	3	1
Oceania	14.8	97	7
Australia	12.0	84	8
New Zealand	2.8	13	5

*Armed forces refer to military personnel actually on duty, including paramilitary forces where these forces contribute substantially to a country's military capabilities. Reserve forces are not included. †Developed countries, 27 in number, are the countries listed under North America, Oceania, European NATO except Greece, Portugal, and Turkey, and the Warsaw Pact except Bulgaria. They also include Austria, Ireland, Finland, Sweden, Switzerland, Japan, and South Africa. ‡Developing countries, 93 in number, are the countries listed under Latin America, the Far East except Japan, South Asia and the Middle East, and Africa except the Republic of South Africa. They also include Albania, Bulgaria, Greece, Portugal, Spain, Turkey, and Yugoslavia. §Includes paramilitary as well as regular forces.

Source: U.S. Arms Control and Disarmament Agency, *World Military Expenditures, 1970* (1971).

scription in its purest form was used to man some of the armies and fleets of Europe in the 17th and 18th centuries, when condemned criminals were permitted to choose service in lieu of imprisonment. More often, however, conscription is part of a program of universal military service accepted by the public and carried out in cooperation with it.

The system is based on rudimentary training given to all men of a certain age who are deemed physically and mentally fit. After their initial term of service, they are placed in the reserves and recalled periodically for retraining. If complex equipment is involved, the training period must be longer, and for that reason naval and air forces tend to have a high proportion of professional personnel and to rely less upon conscripts. The use of conscription assumes a government that is strong enough to coerce its citizens and a national consciousness that will accept this coercion. It may also find a basis in the democratic idea that all citizens ought to participate in the defense of their country. Conscription has, like the militia system, the advantage of obviating a standing professional army with its dangers of political intervention.

In recent years the system of conscription has been undergoing a crisis in many countries. On the one hand, there is a growing awareness that a complex war machine needs professional soldiers. At the same time, resistance to general military service has been increasing, partly from antimilitarism and partly from a decline in national consciousness in some countries. Many military men would prefer a professionalized army that would free them of the influx of less dedicated and reliable personnel. Against this there is the argument that only a system of conscription can supply the large forces required for a prolonged war. Another argument in favour of conscription springs from the idea that in a democratic country military service ought to be universal.

Soldiers for pay

Mercenaries. The mercenary soldier serves for pay. There is no question either of loyalty or of alienation. Armed forces composed of hired soldiers are therefore quite different social organizations from those discussed above. While the main problem of a revolutionary army is to preserve the enthusiasm for the cause and that of a conscript army is to maintain strict discipline, the main problem for the mercenary army is to provide the soldiers with their pay. When the army cannot be paid, as often happened in the 16th and 17th centuries, it can no longer be commanded.

The mercenary system, like the other systems, is of ancient origin. In Egypt and Mesopotamia, the state itself was based primarily on mercenary forces. The great military successes of Carthage were won by mercenaries, as were those of the Roman Empire in Europe. Mercenary armies reappeared after medieval times. They were employed in European colonial wars, the best known being the French Foreign Legion. Three sets of circumstances are conducive to the adoption of the mercenary system. Large empires cannot be built and maintained with militia armies, and this creates a demand for mercenary armies. Mercenaries, furthermore, are typical of the absolute state in much the way that militias are characteristic of political democracy. Finally, they often appear in commercial nations that are wealthy enough to maintain hired troops, as in the old state of Venice, in the Dutch republic, and in Great Britain.

The hirelings are usually recruited from densely populated countries and from countries with a military tradition. The Swiss and later the Germans served in armies all over Europe. The Roman emperors had their German guard and the French monarchs their Swiss regiments (of which system the Vatican's Swiss Guard is a relic). Small numbers of mercenaries are still involved in the conflicts of the developing countries. The present-day system of voluntary enlistment in such countries as Great Britain, Canada, and Australia is a variation of the mercenary system, since these countries can, in peacetime, define their problems of recruitment primarily in terms of the labour market and the national budget.

The mercenary system seems likely to become more common in the future, as political considerations lead to the abandonment of peacetime conscription and short-term enlistment takes its place. To a considerable degree, the modern military force reflects the business orientation of the surrounding society: many military jobs require the same techniques and skills as civilian jobs, and it has become necessary to allow military personnel to work under the conditions prevailing in the rest of society. Even trade unions and labour contracts are becoming important to the military employer, particularly in times when labour markets are tight.

FUNCTION AND ORGANIZATION

The formal purpose of armed forces is to act against foreign enemies. In reality, however, they have various other functions.

Domestic functions. In many countries, particularly developing countries, the army has great social and political significance. It stands as a symbol of unity and power. Young nations uncertain of their international role, and often internally divided, attach much value to an army as a national institution. This may first require disposing of the military forces inherited from the previous colonial regime, since many colonial armies were recruited from minorities the ruler found reliable—such as the Berber tribes in French Morocco, the Punjabi and Sikhs in British India, and the partially Christian population of Celebes and Amboina in the Dutch East Indies. The new national armies may resemble their colonial predecessors insofar as they continue the task of maintaining order in the country. If they do not succeed in this, the result may be civil war, as in Indonesia, Nigeria, and Pakistan.

The army as a modernizer

But the army in the newer countries is also an agency of modernization, somewhat in contrast to armies in the Western world, where the military has been for centuries a symbol of tradition and conservatism. It is one of the few well-organized and technically equipped modern institutions. The military career often leads to a successful civilian career, as in Israel. The army also brings together members of different ethnic and religious groups, thus becoming, as in the U.S., an institution of socialization for a heterogeneous immigrant population. No military force can escape this socially integrative function within a heterogeneous society. The task of integrating blacks and whites in the U.S. Army has had parallels in the integration of the French-speaking and English-speaking groups in the Canadian Army and of the Flemings and Walloons in the Belgian Army. In the new countries the army may also undertake a variety of social, cultural, and economic tasks. In the 1920s the Turkish Army of Kemal Atatürk set an example as a military modernization movement that still serves as a model for many countries with military regimes. The soldiers may be put to work building roads and bridges, working on irrigation and land development, or even teaching the population to read. Such uses of the military occur under governments of various political hues; even the Marxist revolutionary regimes of China and Cuba, in the early 1970s, were using soldiers for nonmilitary functions in the political as well as the socio-economic sphere.

Conflict management. The modern military force has lost almost all of the old conception of warfare as a heroic struggle, as it was seen in ancient Greece, some periods of the Middle Ages, and in the 18th century. Such attitudes linger on only in elite formations such as special forces and paratroops, in which individual performance is still important.

The prevailing conception of the military is an instrumental one. The military are seen as instruments of national policy, to be used to strike and destroy an enemy with all available means. As wars have become increasingly destructive, expanding to include the whole population, the military have been reduced to one element in a total national effort that involves political, strategic, social, and economic factors. This conception of total war has been tempered since the advent of nuclear weapons by the doctrine of limited warfare—a strategic pluralism in which various military forces are developed simul-

taneously for the purpose of providing appropriate responses in different types of situations. These include the whole array of traditional, or "conventional," forces together with counter-guerrilla forces. The military effort of a great power is now extremely differentiated, each of the subdivisions being equipped with its own philosophy, armament, and organization.

United Nations forces

The new doctrines of flexible response and limited warfare may be seen as a step toward the international management of conflicts. The effort to manage conflict has been made by forces of the United Nations, acting in an international police capacity, in the Middle East, the Congo, Cyprus, and elsewhere. The outlook of such peace forces differs from that of traditional military force in several ways: the goal is not victory but control of the conflict; there is no enemy, only parties whose conflict endangers the international situation; the peace force must act with as little violence as possible; and, unlike traditional military forces, it must work very cautiously and patiently. The task of the United Nations soldier is thus more akin to policing than to military action.

III. Historical trends

At the most primitive level of society, there is no military organization and no warfare in the conventional sense. Skirmishes among tribes take the form of random strikes and personal duels fought at a distance with weapons such as javelins, blowpipes, or bows and arrows.

Organized warfare first makes its appearance among agriculturalists and cattle breeders. Peasant societies are not inclined to war because of their dependence upon harvests and because their manpower is needed to till the land. But a separate warrior caste may develop within a peasant society, perhaps consisting of invading conquerors who constitute a noble or feudal class. The conquerors are likely to be nomadic people, accustomed to moving over large geographic areas with their cattle and inclined to attack and subdue sedentary people who may be more prosperous than they. In European history such conquerors have included the Huns, Magyars, Mongols, and Arabs. Others, on a smaller scale, are the North American Indians, the East African Masai, and the North African Tuaregs. They begin to develop tactical refinements such as ambush and manoeuvre and the use of spies and scouts, along with more destructive weapons.

HISTORICAL PARALLELS

Rise and decline of the aristocratic warrior

The development of military institutions among the great civilizations of history can be studied on a comparative basis. One can find common characteristics among the civilizations of ancient Egypt, Babylonia, Assyria, Persia, Greece, and Rome, as well as those of old India, old China, and pre-Columbian Mexico and Peru. In their earliest period they were governed by aristocracies that monopolized the military craft. The aristocratic class, rising above the peasant masses, was the only group that could afford military paraphernalia such as horses, chariots, and weapons. Sometimes there was also a militia, but it was at best an auxiliary force and militarily inferior.

This was the situation among the Greeks of the Homeric era, in Etruscan Rome, at the time of the Chou dynasty in China, and in India under the Aryan conquerors.

The rise of cities increased the manpower available for military service. In Greece and Rome a middle class that developed broke the dominance of the nobility. Heavy cavalry and chariot formations gradually gave way to foot soldiers and citizen armies, fighting to defend their cities.

The militia armies of this historical phase were gradually replaced by mercenary troops that were capable of more extended campaigns. Rising empires (such as Rome, Macedonia, the Chin and Han dynasties in China, the Mauryas in India, and some periods of the Aztec and Inca empires) relied upon professional soldiers supported by militia who were increasingly recruited from subject nations. Eventually the population and the army came to stand completely apart from each other, the army often becoming a state within the state. At their peak these empires became defensive in their outlook, and great fortifications were built, such as the Roman limes in central Europe, the wall of Hadrian in northern England, and the Chinese wall in northern China.

Military reformers throughout history have borrowed techniques from the parallel phases of other civilizations. The period of the Renaissance was not only an artistic and philosophical rediscovery of antiquity; the military also made certain acquisitions from Roman military science. Machiavelli, attempting to revive the ancient citizen armies in the Florentine militia, drew some of his ideas directly from the old city-states. Prince Maurice of Orange studied classical military writings and introduced Roman formations, drills, and training methods. Since Maurice, like the Romans, employed mercenary soldiers, the experiment succeeded beyond all expectation.

MEDIEVAL SOLDIERS

The Germanic states that arose in Europe after the fall of the Roman Empire represented a return to an earlier form of social organization with small armies and no military class. The kings of the early Middle Ages attempted to keep their empires intact, following the example of the Roman emperors, but they could raise only a few thousand horsemen at most and had no foot soldiers. Consequently they were easy prey for invaders such as the Magyars and the Norsemen. The European states were dismembered into a great number of smaller communities, and the feudal system came into existence. The kings and other rulers rewarded their retainers with land; the weak, on the other hand, became dependent upon their feudal lords, to whom they had to swear allegiance in exchange for protection.

Knights and nobles

This political system was primarily a military system. Landownership became the basis of military power. Only those who had at their disposal large material means (land, tenants, toll money) could afford the costly armour of the knight and the expenses of his vassals. The power of the kings declined, and local populations preferred to entrust themselves to the nearest lord rather than to the distant and powerless king. The knight became the centre of the military structure. All the lords were autonomous, and collective action became very hard to organize. There were no group military exercises. In battle each leader wanted a position in the forefront, since a position in the reserves or in the rear guard was considered dishonourable. The style of battle often assumed the character of a tournament. In the rare instances when large numbers were mobilized, as in the Crusades, there was little cohesion and cooperation; they were in fact conglomerations of small units under autonomous commanders. Naturally such armies produced little in the way of military tactics and techniques. The principal medieval innovations were in the art of siege. The castles and fortified towns were built to withstand assault, and therefore the battles over them were ingeniously fought.

The predominance of the mounted knight in battle ensured him a place in the aristocracy, since superiority in arms guaranteed political superiority. This led in turn to the glorification of warfare. The knight's obligation to avenge insults created a society in which violent battle was an everyday reality, because, when personal fidelity and loyalty were unavailing, a man could maintain himself only by his own resources.

The feudal mechanism gave way in the late Middle Ages to central monarchies. The new rulers allied themselves with the rising middle class of the cities and gradually brought hired knights into their service, thus squeezing out the feudal nobility. Royal standing armies came into being, including both cavalry and infantry (pikemen and musketeers), constituting a force that the haphazard combat groups of the nobility could not defeat. The military revolution was completed when the pistol won out over the lance, since troops with pistols required less training than knights. The development of artillery, a product of the commercial towns, hurried the downfall of the old nobility.

Another factor in the decline of the mounted knights

was the development of the infantry. As early as the 14th century, knights would often descend from their horses and mingle with the foot soldiers, but they were encumbered by their armour. Companies of archers began to win victories in the 14th and 15th centuries, the most famous being the Battle of Crécy in 1346, when English archers and dismounted men-at-arms defeated French knights. The most important postfeudal mercenary forces were the German Landsknechte (men of the plains) recruited by the emperor Maximilian I about the end of the 15th century; these were foot soldiers armed with pikes. The 16th-century infantry lacked discipline and solidarity; they were troublesome both to their commanders and to the populations they pillaged. They represented a transitional phase between the heroic feudal era and the new era of the centrally controlled monarchical army.

THE ARMIES OF THE KINGS

Basis of the modern army

The military revolution. In the period between 1550 and 1650, the basic structure of modern arms was developed. Important steps had already been taken under the Protestant commanders Maurice of Orange in the Netherlands, Gustavus II Adolphus in Sweden, and Cromwell in England, but now the system became common throughout Europe.

The most general element was a striving for rationality and flexibility. The military force was seen as an instrument to be used as efficiently and effectively as possible. Uniforms, regulations, and standardized equipment made their appearance. Training was systematized with exercises in the use of arms and the introduction of a command language. The specialization of forces was further developed, and the first administrative and clerical services appeared, along with staff organs.

Improved discipline of the troops was essential to all this; the excesses of the mercenary armies of the previous period had to be overcome. Although low-calibre mercenaries were still hired, the leaders maintained a firm grip on them through drill and training, severe punishment, and a clear-cut system of ranks. The traditional liberties of the Landsknechte, who had had the privilege of choosing their own immediate officers, were curtailed. The post of platoon commander was established; this officer, appointed from above, was responsible for the discipline of his men.

The new hierarchically guided and disciplined war machine reached its peak in the Prussian Army of the 18th century, created by Frederick II the Great. Its troops were intensively trained in battlefield manoeuvres. Much emphasis was placed on morale; esprit de corps was encouraged by regimental colours and other distinguishing symbols. The core of the system, however, was discipline: Frederick said of his soldiers, "They must be made to fear their officers more than danger" and "The slightest relaxation of discipline would lead to barbarization."

The absolute monarchs of the 17th and 18th centuries were able to support standing armies and navies. Private and municipal armies gave way to royal armies led by professional officers. These were not national armies like those of the 19th and 20th centuries, since most of the populace remained outside them; they were officered by the aristocracy and manned by foreigners, paupers, prisoners, and deserters. The productive categories of the population were not required to do military service. War was a contest among monarchs, carried on over the heads of the population.

The officers corps

The role of the aristocracy. The new armies of the kings opened military careers to the nobility, who had lost much of their political and economic power with the decline of feudalism. The transition from knight to officer was not a simple one, since the new armies also gave opportunity to adventurous commoners. Anybody who could ride a horse and handle a pistol was able to join the cavalry, which underwent a great expansion. But eventually commoners were excluded from officer rank except in engineering and the artillery, which were not attractive to the nobility. The aristocracy was able to maintain its social status and its life-style, while also being favoured in other ways because of its relation to the monarchy. Thus officers were exempted from taxes since they were considered to be paying an *impôt de sang* ("blood tax") on the battlefield.

A similar development took place in the navy, where in the 16th century the officer received his position as "gentleman commander" alongside the traditional post of master mariner, or "seaman commander." Such positions were given to the king's favourites, who did not have to work through the ranks and were not required to have any experience on ships. Access to an officership might also be gained through wealth. The outright sale of military rank became common, particularly in England and France; sometimes only the highest positions were secure from this moneyed incompetence. Sometimes the "boughten" rank was not even filled because the beneficiary was still a child.

The system inevitably led to great overstaffing. The French Army in 1787 had a total of 36,000 officers, of whom only 13,000 were in active service. Where the Prussian Army had 80 generals, the French Army, of equal size, had 1,171. This military elite was not much more than a collection of profiteers and favourites, with only a few capable officers; but it had the traits of an officers' caste, with a strong tendency toward isolation and a cultivation of feudal values and chivalrous ideas of honour.

The aristocracy was a European elite that cultivated relationships across national borders. This was one reason why the wars of the 18th century often resembled games conducted according to strict rules. The spirit of the time would not tolerate great hazards, and in any event there was little to be hazarded because the unwilling troops were not the material of which fighting armies are made. Military science was practically static, and the new military academies were little more than finishing schools for aristocratic young men.

National differences

Within these general tendencies there were important national differences. France pioneered in the development of a standing army. Under Louis XIV, in the second half of the 17th century, the army became the most powerful in Europe; at one time it numbered 400,000 men in a population of 23,000,000. The French war fleet was also a creation of the monarchy, developed in competition with the English Navy; France lacked the large merchant marine of England and the Netherlands that provided trained personnel for their navies.

England, living in bitter memory of Cromwell's militarism, kept its army small. The army and navy elite were less isolated from the populace than that of France; many officers served in Parliament and regarded themselves as members of the political establishment rather than as exclusively military. The Parliament, moreover, retained close control over the armed forces.

The purchase of military commissions was practiced in England on a large scale and regarded as one of the guarantees against military usurpation. The practice continued throughout much of the 19th century, even surviving the catastrophe of the Crimean War; not until 1871, after the Prussian victory over France, was the purchase system abolished in England.

Prussia's military organization, which reached its height in the 18th century, reflected the social, economic, and political backwardness of the country. It was feudalism controlled by an absolute monarch. The nobility, as vassals of the king, held the highest positions. In 1791 the 895 highest officers were from 518 noble families. In 1806 there were only about 700 commoners among the 7,000 to 8,000 Prussian officers, although there were more than 1,000 non-Prussians, particularly French aristocrats.

The ranks of the Prussian Army included not only paupers, criminals, and foreign mercenaries but also conscripts—the latter being tenants who followed their lords to war. Feudal Prussia was the first state to move toward a national army.

THE NATIONAL ARMY AND THE PROFESSIONAL SOLDIER

The 19th century saw the gradual development of national armies led by professional officers. The fullest ex-

pression of the "nation in arms" was in the Franco-German War of 1870–71. The age of mercenary armies came definitely to an end.

The breakdown of the old system. Many writers and military reformers had long held that the ideal army should be a national militia. The practical superiority of the militia had been shown in the American War of Independence (1775–83) and in the successes of the French revolutionary armies of the 1790s. In both cases, without appreciable changes in military tactics, regular forces were defeated by irregular armies of mobilized citizens. The American victories made a great impression throughout Europe. The colonial sharpshooters and the grim ferocity of Washington's forces were a departure from the gentlemanly tradition. From a military point of view, however, the French innovation had much greater consequences. The national *levée en masse* in 1793 raised large, untrained armies that, notwithstanding serious losses, were able to defeat the mercenaries.

Although 70 percent of the French officers were from the nobility, there was a strong new element of revolutionary egalitarianism and nationalism. In the early stages of the Revolution, the officers were often chosen from the ranks by the men themselves. Foreign contingents were abolished, and the French citizen took on the role of defender of his fatherland. The Constitution of 1793 declared: "Tous les Français sont soldats" ("Every Frenchman is a soldier"). This national army did not long maintain its pure form. Napoleon made use of the new enthusiasm, but he conducted his wars predominantly with professional soldiers supplemented by conscripts. The citizen army thenceforward became partially a conscript army and partially a professional army.

Prussian military reforms carried out after the German defeats in the Napoleonic Wars broke only partially with the past. In 1808 officership was opened to all, but the aristocracy retained its dominance. The establishment of a national army was resisted. Prussia's main innovation in 1807–13 was the development of the professional officers corps.

The new officer

The Prussian model. The Prussian reformers emphasized the systematic training of officers along scientific lines. Their efforts were crowned by the establishment of the Kriegs Akademie in Berlin in 1810. German military thinkers, among whom Carl von Clausewitz was the most famous and influential, led the world in developing the science of war. By 1860, it was estimated, half of Europe's military literature had been published in Germany.

A second important Prussian reform was the creation of the general staff, which after 1860 was the core of the professional elite. Preparation for military operations became systematic and minute business, making possible very rapid and efficient action at the start of a war.

Prussia also led in technical innovations such as breech loading for cannon and rifles and the development of the repeating rifle and machine gun. Another insight of the Prussians lay in their stress on transportation and communication. They used railways and postal and telegraph facilities in ways that greatly increased troop mobility, as well as strengthening the central direction of operations.

The Franco-German War of 1870–71 demonstrated Prussian military superiority, and the other German states proceeded to imitate the Prussian model. Other countries also began to copy it, notably defeated France, which built up a mass army with the help of conscription, extended the term of military service, and invested more heavily in military preparations. After 1870, when all of the European imperial powers and Japan introduced general conscription, the concept of the "nation in arms" finally took concrete shape.

An armaments race developed between 1870 and 1914, principally between the German and French armies and the German and British navies. Although it greatly increased the destructive power of the forces through the improvement of the artillery and the expansion of technical personnel such as the engineers and signal troops, it was not a military revolution. Armies grew in size, but no qualitative change occurred. Air power and armoured weapons had not yet become important.

IV. Armed forces and society

Relations between the military and the rest of society have never been easy. With their monopoly of the means of violence, the military inevitably acquire some degree of political power, even when a country's constitution and laws stipulate that the civilian authorities are to be dominant. Military intervention in politics is not as surprising as the fact that in many countries such intervention does not occur.

The following discussion deals first with civil–military relations in countries where military intervention in politics is exceptional; second, with the infiltration of political conceptions into the military; and, finally, with the militarizing of the political and social order in some countries.

CIVIL–MILITARY RELATIONS

The term civil–military relations is of course much too simple to describe the complex network of interests that usually exists between the various branches of the military and the various sectors of civil society. For example, the relation between the professional officers corps and society is of a character completely different from that between enlisted men and society.

The old and the new establishments

Only since the 19th century has it been possible to speak of civil–military relations in a clear-cut sense, for not until then did the military become a specialized institution with a professionalized leadership apart from the rest of society. In the 17th and 18th centuries, the lines of political authority converged upon the person of the monarch, and the military elite was an indissoluble part of the aristocracy or the social elite. The rank and file of the armed forces, on the other hand, were drawn from classes that were not yet part of political society. Consequently a difference between military and social interests was scarcely discernible except as between the aristocracy and the rising bourgeoisie.

In the 19th century the military elite began to dissociate itself from the political–social elite, while the development of conscription brought more civilians into the military apparatus. The technical development of the military led to internal differentiation, so that special military interests and groups arose parallel to similar groups in civil society.

The complex of civil–military relations thus created is often written about as if there were a contrast between military bellicosity and civil peaceableness. As a generalization this is false. Bellicosity may, in fact, be more characteristic of certain political and social movements, of certain intellectual traditions and nationalistic political parties, than of the professional soldier. It is nonetheless true that the military has obvious interests to pursue and that it often attempts to influence important elements in society or at least to maintain close relations with them. Particularly important to the military are the government, industry, and public opinion.

The government is the most important because it makes decisions about international relations and the size of the military budget. In developed countries the relationship of the military with the government is less a matter of exercising power than of bringing to bear the pressure of special military interests. The military is in this respect on a level with representatives of other special interest groups such as industry, education, and public health care. It differs from them in its relation to parliament since it cannot participate in elections as directly as other interests can; soldiers are often disliked by members of parliament and seek to defend themselves by closing their ranks and keeping parliamentary inquisitors as ignorant as possible of the internal problems of the military.

Inter-service rivalry

Another problem arises from rivalry and competition between the services, each of which may seek to enlarge its budget at the expense of the others, as occurs in the United States, where the army, navy, and air force have sometimes aired their rivalries in public. The rapid technical development of arms leads to new sources of friction such as the need to choose between the strategic air force and naval aircraft carriers or the question of giving the army and navy aircraft and missiles. Such interser-

vice rivalry can in theory be solved by a complete integration of the armed forces, which would also permit economies in auxiliary services such as training, research, developments, and medical services. Canada has carried out a far-reaching unification program, but such thoroughgoing centralization is not possible for large countries.

Special relationships between the military and industries that depend on military contracts, known in President Eisenhower's phrase as the "military–industrial complex," have aroused much controversy, but it is difficult to tell how significant the relationships are. To begin with, there is obviously a connection between military spending and general economic prosperity: if some countries were to disarm rapidly and extensively, they would face serious problems of unemployment. It is also true that some business firms in the electronics and aircraft industries receive very large military contracts. Many officers retire from service to take jobs with companies having military contracts; in the U.S. in 1968, according to a study by the Joint Economic Committee's Subcommittee on Economy in Government, the 95 largest prime military contractors together had 2,072 former high military officers on their payrolls.

The cultivation of public opinion has become of great importance, especially in countries where the mass media are highly developed and the protective isolation formerly surrounding the military has been lost. Much of the "image building" is carried on by specialized organs of the various services, as well as by information offices in the defense ministries. Civilian organizations devoted to the armed forces are important channels in the public relations effort. Of even more importance is the relation between military men and political leaders, particularly in matters of the budget.

A successful war usually results in greatly increased influence and status for the military. This was notably the case in the United States after World War II. Unpopular wars, by the same token, can be disadvantageous to the military, as shown in the cases of the French in Indochina and Algeria, The Netherlands in Indonesia, and the U.S. in Vietnam.

Politicization of the military. The traditional doctrine of civil–military relations presupposes a certain balance between political and military interests. It assigns to the military the role of a specialized apparatus for violence, while exempting it from political responsibility for its actions. This is the expression of a pluralistic society, in which every institution has a great deal of autonomy in respect to politics. Totalitarian regimes take a different view: soldiers are regarded not as professional specialists at the disposal of the state but as part of the political order. In an effort to meld the military with the political superstructure, totalitarian regimes manipulate a whole sequence of control mechanisms.

Totalitarian controls

One technique is indoctrination. High officers are always members of the political party, and military training includes courses in politics. Politicization is also advanced through the selection of military personnel. By drawing upon the politically active members of youth organizations and by continuously postulating loyalty to the political regime as a criterion of selection and advancement, an officers corps is created that thinks in terms of unity with the party. This is strengthened by having a party organization within the military force at all levels, as in Communist countries where political functionaries hold positions of authority in the military hierarchy. Another way of controlling the military is to establish special units under the direct authority of the party and detached from the military command. Examples of these were the troops of the Soviet security police (NKVD) under Stalin and the German Schutzstaffel (SS) under Hitler.

Democratic indoctrination

But democratic countries also attempt to politicize and indoctrinate their troops, particularly when they are in conflict with other political systems. The United States became especially concerned with political morale during the Korean War of the early 1950s, when American prisoners appeared to be unable to withstand the ideological pressure of the Chinese Communists. The indoctrination effort was diverted from its traditional concern with military victory to that of a struggle in the name of Western democracy against a detestable political system. The wave of indoctrination also reached the European allies of the United States. In West Germany an effort was made to combat the National Socialist past by stressing the meaning of political democracy. Since it is difficult to convey the positive aspect of democratic values without conjuring up an enemy, the political indoctrination of the Bundeswehr led to a form of anti-Communist schooling that had not been originally intended. The question remains whether the modern soldier can be expected to fight effectively without being imbued with political purpose.

The militarization of politics. Sometimes the soldier intrudes himself into the political sphere, and the result is the militarization of the political order. Militaristic tendencies take several forms. They may be confined to the social elite, as in some empires and monarchies, or they may extend to the population, so that the soldier will have unusual social prestige and the military virtues be regarded as superior, as in Prussia during the 19th century and in National Socialist Germany during the 1930s.

Militarism of a different character exists when the armed forces look after their own interests at the expense of the general welfare, by aggrandizing themselves economically or socially. The military then becomes a parasite on the political body of the state. This has happened in Latin America and in various new nations with military regimes, where the internal politics of the country come to be controlled by military men.

But militarism should be distinguished from militancy. Most democratic countries are averse to militarism and are extremely sensitive to any expansion of the soldier's political and social influence. Many young democratic countries, however, are very militant—as were the United States and revolutionary France. Switzerland is a classic example of the militant spirit and Israel another. Political thinkers since Edmund Burke and Alexis de Tocqueville have pointed out that democratizing the military apparatus may increase the aggressiveness of a country and thus be a cause of wars.

MILITARY INTERVENTION IN POLITICS

Frequency of military coups

The military coup d'etat is a normal occurrence in some countries. Attempts by the military to take over the state may occur as often as ten times in one year. This is not a new phenomenon. It happened in Latin American countries throughout the 19th century, particularly in Mexico, Peru, and Chile; Spain between 1832 and 1876 was plagued by military coups, and so were the Balkans for many decades. If military intervention in politics receives more notice in the present day, it is partly because such intervention occurs frequently in the numerous new countries around the world. In many cases it produces relatively stable regimes, so that the political role of military men cannot be considered an exceptional or transitory phenomenon.

Conditions of military intervention. Most of the successful military regimes are in countries that belonged to former colonial empires. The Latin American countries are the successors of the Spanish–Portuguese empire that broke up around 1820, while those in the Middle East are the successors of the Ottoman Empire. The new nations in Africa and Southeast Asia emerged from the empires of Great Britain, France, Belgium, and The Netherlands.

Countries seem most likely to resist military intervention if they have either a strong democratic tradition or a Communist political system. In the latter case the military is kept subordinate to the political power by a strong Communist political party; the effectiveness of the party can be seen by comparing North and South Korea and North and South Vietnam.

One factor leading to military intervention in politics is evidently an inability of the civilian elements to assert their legitimacy. When the tradition reserving the exercise of governmental power to political rather than mili-

tary institutions is lacking, the political arena is opened to the military on an equal footing. Another factor is the general ineffectiveness of government in the new nations. The high hopes of the struggle for independence are often followed by disappointment and internal unrest. The military represents a relatively efficient and uncorrupted instrument of power where other social and political institutions, such as political parties and labour unions, are weak and cannot act effectively.

In countries where the struggle for independence has been settled by arms, the military is strong and often has a popular following, as in Algeria and Indonesia. If the young nation has been humiliated by foreign pressure or by military defeat, the army may step forward to defend the honour of the nation, as in Turkey and Egypt. If the country's unity is threatened by internal tensions, the army as a national institution can undertake to save it, as in Nigeria, Indonesia, and Pakistan.

Origins of military regimes

Military regimes. The rise of military regimes comes about as a reaction to a worsening political situation, as can be seen in the new African countries. Initially, after the proclamation of independence, the armies are politically neutral. They enter politics at a later period, usually in protest against the presence of European officers and the government's policy in military matters. At this stage the army's opposition begins to take on a more purposeful character, and it enters the struggle for political power. Once it has achieved power it is slow to relinquish it. Practically every military regime that succeeds to a civilian government promises in the beginning to depart as soon as order is established and general elections have been held. But elections are time after time postponed or manipulated, as in Indonesia, or declared to be invalid, as in Pakistan, Argentina, and Brazil. A situation is created in which the government can be changed only by a new military coup, and finally this method of governmental change becomes the customary one; military rule is then permanent.

The term military regime applies to governments that may be superficially quite different from one another—some are headed by monarchs; in others a single political party has power; while still others preside over a multiparty system—but the ultimate power is in the hands of the military, which may or may not have official posts in the government. The degree of military intervention in social and economic life may range from almost none to the occupation of all the key economic, social, and cultural positions, including positions in the universities and public utilities.

Military regimes are little inclined to ideology. They prefer to approach their problems technically and pragmatically; they regard the internal development of their countries as a process of economic and social modernization, not as the achievement of a political utopia. In their external relations they are moved more by nationalist sentiments than by feelings of solidarity with international political movements or coalitions.

THE FUTURE OF THE ARMED FORCES

Decreasing use of warfare among great powers

Modern warfare has become increasingly restricted to great powers and to coalitions of highly developed nations. War is now impossible among many of the countries that were once traditional enemies, such as Germany and France. Large blocs of states such as NATO and the Warsaw Pact countries have suppressed their old national differences to the extent that a violent settlement of mutual disputes among them has now become extremely unlikely. And the development of nuclear weapons has reached the point at which they can no longer be used militarily but only politically. The public attitude to war has also changed, particularly in the Western democracies, where the use of violence in service of national interests is regarded as less and less tolerable. By the 1970s, many believed the aversion to war had become so great among the younger generation as to destroy the psychological climate in which war tensions develop.

If these trends were all that counted, one could project the future of the armed forces as a technologically elaborate, extremely costly means through which great combinations of states will make their mutual spheres of interest unassailable. The military will become more and more a body of professional specialists and managers carrying out their task in close cooperation with the political leaders.

But these great coalitions of rich countries represent only half of the world. The other half, that of the new nations and developing countries, is quite different in political, social, and therefore in military terms. The tens of new nations that have originated from the process of decolonization seem to most observers fated to travel the same long road of nationalism that Europe has now left behind. Bloody civil wars such as those in Indonesia, Zaire, Nigeria, and Pakistan can easily lead to international conflicts. Revolutionary wars as in Vietnam can lead to intervention by the great powers. Ethnic and cultural differences such as those in the Indian subcontient, in The Sudan, in the Middle East, and in Indonesia may also produce violent confrontations.

One consequence is that the traditional type of military force (light infantry supported by limited motorized units) is still in many parts of the world a decisive weapon of power, though it weighs heavily on the resources of poor countries. Another consequence is that the big powers who may become involved in these conflicts will seek to maintain conventional forces that can intervene, as did Great Britain in Malaysia and East Africa, the United States in Korea, Vietnam, and Latin America, the Soviet Union in the Arab world, and France in Chad. The United Nations may not furnish a substitute for this intervention by the great powers, and the peripheral areas between the blocs may remain the scene of various disputes in which weapons will be used. As the world becomes more unified, these peripheral areas may seem to multiply.

BIBLIOGRAPHY

The military system: MORRIS JANOWITZ and ROGER W. LITTLE, *Sociology and the Military Establishment*, rev. ed. (1965), a good short introduction to the study of the armed forces as a social institution; ROGER W. LITTLE (ed.), *Handbook of Military Institutions* (1971), covering most aspects of the modern armed forces; CHARLES H. COATES and ROLAND J. PELLEGRIN, *Military Sociology: A Study of American Military Institutions and Military Life* (1965), a systematic introduction to the problems of the U.S. military forces.

The professional soldier: MORRIS JANOWITZ, *The Professional Soldier: A Social and Political Portrait* (1960), the best and most complete study of the U.S. officers corps, including comparisons with officers corps of other countries; KARL DEMETER, *Das deutsche Offizierkorps in Gesellschaft und Staat, 1650–1945*, 2nd rev. ed. (1962; Eng. trans., *The German Officer-Corps in Society and State, 1650–1945*, 1965); MICHAEL LEWIS, *England's Sea Officers: The Story of the Naval Profession* (1939).

Military bureaucracy and organization: JOHN R. BEISHLINE, *Military Management for National Defense* (1950), a manual of military planning, organization, and command; KARL LANG, "Military Organizations," in J.G. MARCH (ed.), *Handbook of Organizations*, pp. 838–878 (1965), a summary of the problems of military organizations, with extensive references to the literature; MORRIS JANOWITZ (ed.), *The New Military: Changing Patterns of Organization* (1964), a collection of studies on military selection, training, career management, and retirement.

Recruitment and integration: M.R.D. FOOT, *Men in Uniform: Military Manpower in Modern Industrial Societies* (1961), a classification of the various organizational systems, drawing upon practices in the United States, the Soviet Union, Great Britain, France, Germany, Switzerland, Belgium, Denmark, Sweden, Turkey, Israel, Canada, and Australia; HAROLD WOOL, *The Military Specialist: Skilled Manpower for the Armed Forces* (1968), a study of the demand for trained personnel in the armed forces, with much U.S. factual material; SAMUEL A. STOUFFER *et al.*, *The American Soldier*, 2 vol. (1949), studies of all aspects of the U.S. soldier, unsurpassed as source material on military behaviour; CHARLES C. MOSKOS, JR., *The American Enlisted Man: The Rank and File in Today's Military* (1970), on the U.S. soldier at home and overseas, with special attention to combat behaviour in Vietnam and to racial relations in the military forces.

Historical trends: ALFRED VAGTS, *A History of Militarism: Civilian and Military*, rev. ed. (1960), a history of the soldier in Western society, with special emphasis on the relation be-

tween soldier and state; STANISLAV ANDRESKI, *Military Organization and Society*, 2nd ed. (1968), a classification of military organizational forms on the basis of material from all historic societies; HANS DELBRUECK, *Geschichte der Kriegskunst im Rahmen der politischen Geschichte*, 2nd ed., 4 vol. (1962–66), the classic work on the history of military institutions, beginning with the Persians and the Greeks; RUSSELL F. WEIGHLEY, *History of the United States Army* (1967); WALTER MILLIS, *Arms and Men: A Study in American Military History* (1956); JOHN W. WHEELER-BENNETT, *The Nemesis of Power: The German Army in Politics, 1918–1945* (1953), a profound and extensive study of the relation between the military and the state under the Weimar Republic and National Socialism; JOHN S. AMBLER, *Soldiers Against the State: The French Army in Politics* (1968), on the alienation of the French military in the years 1945–62; B.H. LIDDELL HART (ed.), *The Soviet Army* (1956), a collection of short studies by military experts from various countries on the history, organization, and achievements of the Soviet Army.

Armed forces and society: JACQUES VAN DOORN (ed.), *Armed Forces and Society: Sociological Essays* (1968), and *Military Profession and Military Regimes: Commitments and Conflicts* (1969); MORRIS JANOWITZ and JACQUES VAN DOORN (eds.), *On Military Regimes* (1971), and *On Military Ideology* (1971), collections of short studies of the military in developed and developing countries; SAMUEL E. FINER, *The Man on Horseback: The Role of the Military in Politics* (1962), on intervention by the military in politics, with an extensive bibliography; SAMUEL P. HUNTINGTON, *The Soldier and the State: The Theory and Politics of Civil-Military Relations* (1957), a standard work that includes, along with general theory and classification, an extensive discussion of the development of civil–military relations in the U.S.; ROMAN KOLKOWICZ, *The Soviet Military and the Communist Party* (1967); WILLIAM GUTTERIDGE, *Military Institutions and Power in the New States* (1965), an introductory work on the origin, composition, and external relations of the military in the countries of the Third World; MORRIS JANOWITZ, *The Military in the Political Development of New Nations: An Essay in Comparative Analysis* (1964), a short introduction to the internal problems of military institutions and the relations between the military and society; JOHN J. JOHNSON (ed.), *The Role of the Military in Underdeveloped Countries* (1962), a collection of studies by experts; EDWIN LIEUWEN, *Arms and Politics in Latin America* (1960); CLAUDE E. WELCH, JR. (ed.), *Soldier and State in Africa: A Comparative Analysis of Military Intervention and Political Change* (1970).

United Nations forces: D.W. BOWETT, *United Nations Forces: A Legal Study of United Nations Practice* (1964); LINCOLN P. BLOOMFIELD *et al.*, *International Military Forces: The Question of Peacekeeping in an Armed and Disarming World* (1964); ALAN JAMES, *The Politics of Peace-Keeping* (1969), a summary of UN efforts at collective security.

(J.v.D.)

Armenian Language

The Armenian language, which forms a separate branch of the western group of Indo-European languages, is the mother tongue of the Turkish Armenians and of the Armenians in the Armenian Soviet Socialist Republic, where it is spoken by 2,000,000 people. In other parts of the Soviet Union, especially in the neighbouring republics of Georgia and Azerbaijan, it is used by some 1,700,000. Armenian emigrants and refugees have taken their language with them all over Asia Minor and the Middle East and from there to many European countries, especially Romania, Poland, and France, and to America, particularly the United States. In all, Armenian is probably spoken by about 5,500,000 people.

History of the language. Armenian was introduced into the mountainous Transcaucasian region (called Greater Armenia by the Greek historians) by invaders coming from the northern Balkans, probably in the latter part of the 2nd millennium BC. These invaders occupied the region on the shores of Lake Van that had previously been the site of the ancient Urartean kingdom. By the 7th century BC the Armenian language seems to have replaced the tongues of the native population. It is tempting to connect the invasion with the downfall of the Hittite empire in Anatolia.

Development of the alphabet

After the introduction of Christianity in the beginning of the 4th century AD, the language was reduced to writing; the alphabet, of 38 letters, was invented, according to tradition, by the bishop Mashtots (Mesrob) in about AD 400. Admirably suited to the phonology of Armenian, it is still used in various forms by Armenians all over the world. The oldest writings in the language date from the 5th century; they are preserved in manuscript form from the 9th century. Grabar, the written language of the 5th century, the golden age of Armenian culture, is traditionally said to be based on the dialect of Tarawn on Lake Van. To what extent the spoken language was split into dialects at that time is not known. The language of the literature from the 5th to the 8th century is remarkably homogeneous, but by the 9th century the influence of the spoken dialects was noticeable, especially in legal and historical texts. Among the Middle Armenian varieties of Grabar, the best known is the 12th- and 13th-century chancellery (court) language of the Armenian kingdom in Cilicia. More or less corrupted versions of Grabar continued as the literary language until the middle of the 19th century.

The Armenian Alphabet

letter		equivalent	letter		equivalent
capital	lowercase		capital	lowercase	
Ա	ա	a	Մ	մ	m
Բ	բ	b	Յ	յ	y
Գ	գ	g	Ն	ն	n
Դ	դ	d	Շ	շ	sh
Ե	ե	e	Ո	ո	o
Զ	զ	z	Չ	չ	ch'*
Է	է	ē	Պ	պ	p
Ը	ը	ĕ	Ջ	ջ	j
Թ	թ	t'*	Ռ	ռ	rh
Ժ	ժ	zh	Ս	ս	s
Ի	ի	i	Վ	վ	v
Լ	լ	l	Տ	տ	t
Խ	խ	kh	Ր	ր	r
Ծ	ծ	ts	Ց	ց	ts'*
Կ	կ	k	Ւ	ւ	w
Հ	հ	h	Փ	փ	p'*
Ձ	ձ	dz	Ք	ք	k'*
Ղ	ղ	gh	Օ	օ	ō
Ճ	ճ	ch	Ֆ	ֆ	f

*The *spiritus asper*, ', indicates aspiration.

Modern East Armenian and West Armenian. In the 1800s, the writers Khachatur Abovean (1805–48) and Mikael Nalbandean (1829–66) and other Armenian nationalists made efforts to reach the populace with nationalist propaganda. As a result, a national revival occurred from which a new literary language emerged that was much closer to the spoken language. This is known in two varieties. East Armenian, now the official language of the Armenian Soviet Socialist Republic, is based on the dialect of the Ararat valley and the city of Yerevan; West Armenian has its foundation in the dialect of Istanbul. East Armenian is also spoken in other parts of the Soviet Union, whereas the western variety dominates in the Armenian colonies in the Middle East, Asia Minor, Europe, and America. The differences between these two written forms of Modern Armenian are slight, constituting no barrier to mutual intelligibility.

Dialects

In addition to the two literary languages, there are a great number of dialects, some of which are so different that the speakers cannot understand each other. It is estimated that before World War I some 50 distinct dialects were spoken. Today, the spoken dialects are losing ground in the Soviet Union, under the pressure of the standard written language. Accurate statistics on the extent to which Armenian dialects are used in Turkey and in other parts of the Middle East are not available.

When the scientific study of Armenian started in the 19th century, the language was considered an Iranian dialect, a mistake easily explained by the vast number of

Iranian loanwords in the vocabulary. Subsequent studies, however, have convincingly shown Armenian to be an independent member of the Indo-European language family. According to the Greek historian Herodotus, Armenian was a variety of Phrygian, a tongue presumed to be Indo-European. What little is known of the latter is insufficient to support or confirm such a claim.

Phonological characteristics. Phonetic developments in Armenian have radically changed the sound system of the old Indo-European parent language. In particular, the pattern of the plosive consonants—the stops—has been completely reshuffled. In the more conservative central Armenian dialects three series of stops are distinguished (voiced *b, d, g*, which in some dialects are aspirated; unvoiced *p, t, k;* and unvoiced aspirated *ph, th, kh*). In the dialects of the periphery, the three series have been reduced to two (aspirated *ph, th, kh* and unaspirated *p, t, k*, or, as in Istanbul, *b, d, g*). These differences are concealed in the traditional orthography. Sibilants (fricatives) of various types and affricates have emerged through palatalization of the old palatal and labiovelar stops. Thus, Old Armenian *dz* (*z*) and *ts* may go back to the Indo-European palatal stops *ǵh* and *ǵ*, and *dž* (*ž*) to the Indo-European labiovelar $g^{w}h$ before *e* and *i;* Old Armenian *tsʿ*, *tšʿ*, and *tš* may derive from the Indo-European consonant clusters *sk, ky*, and *gy*. All Armenian dialects distinguish two types of *r*, one strongly trilled, one weakly trilled. Old Armenian also differentiated between two types of *l*, a neutral one and a velarized variety, which is made by moving the back of the tongue nearer to the soft palate at the back of the mouth. The latter type has developed into a voiced velar fricative in the modern dialects.

Sounds of Armenian

Grammatical characteristics. All of the spoken dialects and the two literary languages have maintained a fairly complicated system of noun declension, distinguishing six or seven cases. The plural stem, derived from the singular stem by the addition of the suffix *-(n)er*, is declined as a singular, according to the Turkish pattern. Characteristic of the changes in the Old Armenian verbal system is the general replacement of simple present tense forms by periphrastic expressions. These are groups of words, including auxiliaries, that take the place of a single word that is capable of being inflected to show tense or some other feature. The various types of periphrastic forms serve as the basis for the classification of the dialects. In Old and Modern Armenian, the main tense distinction is that between present, aorist (denoting occurrence without reference to completeness, duration, etc.), and periphrastic perfect tenses. The old subjunctive, still extant in classical Armenian, has been lost in the modern language. To express future time, Old Armenian used the subjunctive forms; Modern Armenian employs periphrastic expressions, as is done in the English future forms *I shall go* and *he will work*. Also characteristic of Modern Armenian is the importance of the passive forms of the verb, which are strictly parallel to the active forms, and the emergence of a special negative conjugation with differing forms for verbs in instances like "I read" and "I don't read." Whereas Old Armenian was rather close to ancient Greek in many respects, Modern Armenian is typologically much closer to Turkish. Among the features that illustrate this similarity are the agglutinative system of declension (*i.e.*, the compounding of several linguistic elements of independent meaning into a single word), the use of suffixes to indicate possession, the employing of passive and causative forms for all verbs, and the use of postpositions (grammatical elements placed after the word) instead of prepositions (as used in Old Armenian). The vocabulary of the written languages is purely Armenian, being based almost exclusively on that of Grabar, with very few loanwords from the neighbouring languages. (The Iranian loanwords mentioned above were incorporated into Armenian before the creation of the written language.)

BIBLIOGRAPHY. S.L. KOGIAN, *Armenian Grammar* (1949); H. HUBSCHMANN, "Armenische Grammatik," *Armenische Etymologie* (1897); A. MEILLET, *Esquisse d'une grammaire comparée de l'arménien classique*, new ed. (1936); GERHARD DEETERS, *Armenisch und Südkaukasisch* in *Caucasica* (1927); HEINRICH ZELLER, "Armenisch," in *Geschichte der indogermanischen Sprachwissenschaft*, vol. 4 (1927). For ancient Armenian, see A. MEILLET, *Altarmenisches Elementarbuch* (1913); and H. JENSEN, *Altarmenische Grammatik* (1959). Medieval Armenian is treated in J. KARST, *Historische Grammatik des Kilikisch-Armenischen* (1901). For modern speech, see H. ADJARIAN, *Classification des dialectes arméniens* (1909); and A. ABEGHIAN, *Neuarmenische Grammatik* (1936).

(H.K.V.)

Armenian Soviet Socialist Republic

Occupying a landlocked area just south of the great mountain range of the Caucasus, and fronting on the northwestern extremity of Asia, the Armenian Soviet Socialist Republic, or Armenia, is the smallest of the 15 republics that make up the Soviet Union. Its area—11,500 square miles (29,800 square kilometres)—is no more than 0.13 percent of the national territory. To the north and east, Armenia is bounded by the Georgian and Azerbaijan Soviet Socialist republics, while its neighbours to the west and southeast are, respectively, Turkey and Iran. Armenia lies in the southern portion of the Transcaucasian region. By the mid-1970s it was the home of some 2,790,000 people.

Modern Armenia is part of (but not identical with) ancient Armenia, one of the world's oldest centres of civilization, whose peoples have long inhabited the highlands of the area. With the loss of autonomy in the 14th century AD, Armenia was subjected to constant foreign incursions, which, together with the centuries-old rule of Ottoman and Persian conquerors, imperilled the very existence of the Armenian people. The portion of Armenia lying within the former Russian Empire achieved its current political status on November 29, 1920. Its capital is Yerevan. (See further URARTU AND ARMENIA, HISTORY OF; ARMENIAN LANGUAGE; CAUCASUS MOUNTAINS.)

THE LAND

Relief features. Armenia is a mountainous country, characterized by a variety of scenery. Its average altitude is 5,900 feet (1,800 metres) above sea level. There are no lowlands: half the territory lies at altitudes of 3,300 to 6,600 feet (2,000 metres); a mere 10 percent lies below the 3,300-foot mark.

Mountainous terrain

The northwestern part of the Armenian highland—containing Mt. Aragats, the highest peak (13,418 feet, or 4,090 metres) in the republic—is a combination of lofty mountain ranges, deep river valleys, and lava plateaus dotted with extinct volcanoes. To the north and east, the Somkhet, Bazum, Pambak, Areguni, Sevan, Vardenis, and Zangezur ranges of the Little Caucasus lie across the northern sector of Armenia. Elevated volcanic plateaus (Lory, Shirak, and others), cut by deep river valleys, lie amid these ranges.

In the eastern part of the republic, the Sevan Depression, containing Lake Sevan (525 square miles, or 1,360 square kilometres) and hemmed in by ranges soaring to a height of 11,800 feet, lies at an altitude of about 6,200 feet (1,900 metres). In the southwest, a large depression—the Ararat Plain—lies at the foot of Mt. Aragats and the Gegam Range; the Araks River cuts this important plain into halves, the northern half lying in the Armenian S.S.R. and the southern in Turkey and Iran.

Climate. Armenia's climate, because of its position in the deep interior of the northern part of the subtropical zone, enclosed by lofty ranges, is dry and continental. Regional climatic variation is nevertheless considerable. Intense sunshine occurs on many days of the year. Summer, except in high-altitude areas, is long and hot, the average June and August temperature in the plain being 77° F (25° C); sometimes it rises to 108° F (42° C). Winter is generally not cold; the average January temperature in the plain and foothills is about 23° F (−5° C), whereas in the mountains it drops to 10° F (−12° C). Invasions of Arctic air sometimes cause the temperature to drop to −22° F (−30° C): the record low is −51° F (−46° C). Winter is particularly inclement on the elevated, windswept plateaus. Autumn, long, mild, and sunny, is the pleasantest season.

The ranges of the Little Caucasus prevent humid air masses from reaching the inner regions of the republic. On the mountain slopes, yearly rainfall approaches 315 inches (8,000 millimetres), while the sheltered inland hollows and plains receive only 80 to 160 inches (2,000 to 4,000 millimetres) of rainfall a year.

The climate changes with elevation, ranging from the dry subtropical and dry continental types found in the plain and in the foothills up to a height of 3,000 to 4,300 feet, to the cold type above the 6,600-foot mark. The plains and foothills are sufficiently well warmed to permit cultivation of such cold-sensitive fruits as figs, pomegranates, peaches, and grapes. Higher up, tobacco, cereals, and some fruits are raised; and grain crops, potatoes, and fodder grasses are found up to 6,600 to 8,200 feet.

Hydrology. Of the total precipitation, some two-thirds is evaporated, and one-third filters into the rocks, notably the volcanic rocks, which are porous and fissured. The many rivers in Armenia are short and turbulent with numerous rapids and waterfalls. The water level is highest when the snow melts in the spring and during the autumn rains. As a result of considerable difference in altitude along their length, some rivers have great hydropower potential; the total for the republic is nearly 22,000,000,000 kilowatt-hours a year.

Araks River

Most of the rivers fall into the drainage area of the Araks (itself a tributary of the Kura River of the Caspian Basin), which, for 300 miles, forms a natural boundary between the Soviet Union and Turkey and Iran.

The Araks' main left-bank tributaries, the Akhuryan (130 miles), the Razdan (Hrazdan; 90 miles), the Arpa (80 miles), and the Bargushat (110 miles), serve to irrigate most of Armenia. The tributaries of the Kura—the Debet (57 miles), the Agstev (80 miles), and others—pass through Armenia's northeastern regions. Lake Sevan, holding over nine cubic miles (39 cubic kilometres) of water, is fed by dozens of rivers, but only the Razdan leaves its confines.

Armenia is rich in springs and wells, some of which possess medicinal properties.

Biogeography. The broken relief of Armenia, together with the fact that its highland lies at the junction of various biogeographical regions, has produced a great variety of landscapes; though a small country, it boasts more soil types (over 15) and plant species (over 3,000) than the vast East European Plain. There are five altitudinal vegetation zones: semidesert, steppe, forest, alpine meadows, and high-altitude tundra.

The semidesert landscape, ascending to a height of 4,300–4,600 feet, consists of a slightly rolling plain covered with scanty vegetation, mostly sagebrush. The vegetation includes drought-resisting plants such as juniper, sloe, dog rose, and honeysuckle. The boar, wildcat, jackal, adder, gurza (a venomous snake), scorpion, and, more rarely, the leopard inhabit this region.

Steppes

Steppes predominate in the republic. They start at altitudes of 4,300–4,600 feet, and in the northeast ascend to altitudes of 6,200–6,600 feet. In the central region they reach 6,600–7,200 feet and in the south are found as high as 7,900–8,200 feet. In the lower altitudes the steppes are covered with drought-resistant grasses, while the mountain slopes are overgrown with thorny bushes and juniper.

The forest zone lies in the southeast of Armenia, at altitudes of 6,200–6,600 feet, where the humidity is considerable, and also in the northeast, at altitudes of 7,200–7,900 feet. Occupying nearly 10 percent of the republic's entire territory, the northeastern forests are predominantly beech. Oak forests predominate in the southeastern regions, where the climate is drier, and in the lower part of the forest zone hackberry, pistachio, honeysuckle, and dogwood grow. The animal kingdom is represented by the Syrian bear, wildcat, lynx, and squirrel. Birds—woodcock, robin, warbler, titmouse, and woodpecker—are particularly numerous.

The alpine zone lies above 6,600 feet, with stunted grass providing good summer pastures. The fauna is rich; the abundant birds include the mountain turkey, horned lark, and bearded vulture, while the mountains also harbour the bezoar goat and the mountain ram, or mouflon.

Finally, the alpine tundra, with its scant cushion plants, covers only limited mountain areas and solitary peaks.

Regions. One of the more important of the distinctive regions of Armenia is that including the Ararat Plain and the surrounding foothills and mountains. This is a prosperous and densely populated area, the centre of Armenia's economy and culture, and traditionally the seat of its governmental institutions.

The Ararat Plain

The other regions are the Shirak Steppe, the elevated northwestern plateaus, the republic's granary; Gugark, high plateaus, ranges, and deep valleys of the northeast, covered with forests, farmlands, and alpine pastures; the Sevan Basin, the hollow containing Lake Sevan, on whose banks are farmlands, villages, and towns; and Syunik, the two parts of which—the northern (Vayk) and the southern (Zangezur)—lie in the southeast. This last region is a maze of gorges and river valleys cutting through high ranges. It is an area rich in ores, with fields and orchards scattered here and there in the valleys and on the mountain sides.

THE PEOPLE

Origins. Armenians constitute nearly 90 percent of the republic's population. They were consolidated as a nation in the second half of the 1st millennium BC.

The Russian campaigns against the Persians and the Turks in the 18th and 19th centuries resulted in large emigrations of Armenians under Muslim rule to the Transcaucasian provinces of the Russian Empire and to Russia itself (Erivan, Tiflis, Karabakh, Shamakhi, Astrakhan, Bessarabia). At the time of the massacres in Turkish Armenia in 1915, some Armenians found asylum in Russia.

The Armenians were converted to Christianity *c.* AD 300 and have an ancient and rich liturgical and Christian literary tradition. Believing Armenians today belong mainly to the Armenian Apostolic (Orthodox) Church or the Armenian Catholic Church, in communion with Rome.

Demographic trends. In the first quarter of the 19th century slightly more than 160,000 persons lived within the republic's present borders. In 1920 Soviet Armenia had a population of 780,000. By 1939 this had reached 1,282,000; by 1959, 1,763,000; and by 1970, 2,493,000. This trebling of the population in 50 years was attributable to a high natural increase, which appreciably exceeded the national average, and to an influx of more than 220,000 repatriates from foreign countries. The population of Armenia by the 1970s included Azerbaijanis (nearly 6 percent of the total), Russians (about 3 percent), and a small number of Kurds, Ukrainians, and other groups.

Settlement patterns. The average population density is rapidly rising, from 111.4 persons per square mile (43 per square kilometre) in 1939 to 153.3 by 1959 and to 216.8 by 1970. By contrast, there had been only 13 per square mile in the early 19th century. Average figures, however, are relatively meaningless, as population densities vary widely across the republic.

The density is highest in the Ararat Plain, reaching more than 1,040 per square mile (402 per square kilometre). The river valleys in the southeast and northeast are the next most densely populated area. Half the population is in fact concentrated in the zone marked by an upper altitudinal limit of 3,300 feet, which makes up only about 10 percent of the entire territory.

Many people also live in the foothills, at altitudes of 3,300–4,900 feet, and in the mountains (4,900–6,600 feet). These regions account for a further third of the entire population, with an average density exceeding 155 persons per square mile (60 per square kilometre).

The high ranges and mountains are scantily peopled: less than 5 percent of Armenia's people live above the 6,600-foot mark, and no one resides above 7,700 feet.

Fundamental changes in the distribution of Armenia's population are being caused by the urbanization resulting from economic growth, particularly the republic's industrialization. Before the Revolution, the republic's four cities—Erivan (now Yerevan), Aleksandropol (Len-

inakan), Nor-Bayazet (Kamo), and Geryusy (Goris)—accounted for 10 percent of the total population. By 1974 the urban population numbered 1,699,000, or 62 percent of the total population.

There are 23 towns and 33 other urban-type settlements in the Armenian S.S.R. Urban communities are particularly numerous in the Ararat Plain and in the valley of the Razdan River.

Yerevan

Formerly a dusty town with small mud houses, Yerevan, founded 2,500 years ago, has grown since the 1920s into an industrial city serving as the republic's administrative and cultural centre, with a population that approached 900,000 in 1975. Many of its modern buildings show original architectural design. It is followed in importance by Leninakan, an industrial centre with a population of 180,000; Kirovakan, with 123,000 people; and the small centres of Kafan, Echmiadzin (of great significance in Armenian ecclesiastical and cultural history), Alaverdi, Razdan, and Oktemberyan.

The rural population of the Armenian S.S.R. has grown much less rapidly than the urban population; comprising some 916,000 people in 1939, it fell to 881,000 in 1959, and barely exceeded the 1,000,000 mark in 1974. This pattern is a result of the movement of many villagers to towns. There are nearly 1,200 rural settlements in the Armenian S.S.R., and about two-thirds of the rural population live in communities of 1,000 to 2,000 persons. The high country to the north of Shirak, and in Syunik, has small hamlets that lie in secluded glens, on riverbanks, and near springs; in the plain, such settlements cluster around mountain streams and irrigation canals, amid orchards and vineyards.

THE ECONOMY

Economic development. Armenia has become an industrial country with an important agricultural element. The share of industry in the country's economy—14 percent in 1913 in value terms—reached 73 percent by 1970, while, over the same half century, agricultural output grew 4.5 times, and industrial output increased 180-fold.

This rapid economic advancement—unparalleled among the Soviet republics—is attributable to national financial investments that made possible the creation of a diversified industry, the reconstruction of agriculture, and the development of transport. Formerly a supplier of copper, certain farm products, and cognac, Armenia is now a major supplier of chemicals, nonferrous metals, machines, equipment, precision instruments, textiles and clothing, wines, cognacs, and canned goods. The republic also has mineral resources, especially metal ores.

Electricity

At the initial stage of industrialization, the creation of a power base utilizing the hydraulic potential of mountain streams was of decisive importance. Production of electricity was combined with the building of irrigation works and water-supply systems for industries and cities. The Sevan–Razdan series of hydraulic power stations (six in all, with an aggregate capacity of 557,000 kilowatts) became a first-priority project that used not only the waters of the Razdan but also those of Lake Sevan. This project made possible the electrification of agriculture and helped to build numerous industries. In the 1960s emphasis shifted to thermal electric power stations, which a decade later accounted for 80 percent of the power produced in the republic. An atomic electric power station was also being built at the new town of Metsamor.

Industry. Mechanical engineering, machine tools and electrical power machinery, electronics, and the chemical and mining industries hold a prominent place in the republic's heavy industry, but light and food industries are also fairly well advanced. Yerevan, Leninakan, and Kirovakan are machine-building cities. The centres of the chemical industries are Yerevan, Kirovakan, and Alaverdi.

Nonferrous metallurgy—in Gugark and Zangezur—includes the mining and dressing of copper, molybdenum, and other ores; the smelting of copper; and the extraction of precious and rare metals.

The food industry processes various farm products, meets domestic demand, and supplies other parts of the Soviet Union. The most advanced branches are involved in the primary processing of grapes and production of high-quality cognacs, wines, canned fruits, and vegetables for export.

Light industry—a modern innovation—specializes in the production of woollen, silk, and cotton fabrics; knitted goods and clothes; carpets; and footwear.

Yerevan is the foremost industrial centre, accounting for nearly three-fifths of the total industrial output in the republic; but other industrial centres and regions are developing, notably in the north, where Leninakan and Kirovakan are now major industrial centres.

Agriculture. Agriculture still plays a major role in the Armenian economy, even though industry has taken precedence.

Agriculture engages nearly half the population; many items are shipped to other republics, and farm products provide raw materials for a number of industries. By the mid-1970s there were some 376 state farms (*sovkhozy*) and 338 collective farms (*kolkhozy*) in the republic, with fleets of tractors, combine harvesters, and other farm equipment.

Agriculture in Armenia has to contend with many difficulties. Arable land is scarce; in the densely populated areas there are only 3.7 acres (1.5 hectares) of plowland per capita. Cultivated lands (plowland, orchards, and vineyards) occupy only 20 percent of the republic's total area. Pastures and meadows mowed for hay cover a larger area, approaching 28 percent of the territory. Farmlands in mountain regions form a mosaic of cornfields, orchards, vineyards, and pastures. Considerable tracts of arable land also are found in the Ararat Plain, the Shirak plateau, and the southern part of the Sevan basin.

The extensive irrigated lands in the low, sunny Ararat Plain and cultivated stretches in the northeastern and southern river valleys yield high-quality grapes and fruits. Storage lakes, dams, and pumping stations have been built and irrigation canals dug. More than half the total arable land area is irrigated. Farming, above an elevation of 3,300 feet, also combines with cattle raising; grain crops are cultivated and cattle raised in the mountains, while tobacco and potatoes are raised in the lower, warmer part of the mountain belt.

The leading branch of agriculture is viticulture. Among the many orchard crops, peaches and apricots are commonest. Apples, cherries, mazzards (sweet cherries), and pears are cultivated in the colder climate, and walnuts, hazelnuts, almonds, pomegranates, and figs are also produced in this area. Vegetables are grown in the main agricultural regions, potatoes in the cooler mountains. Quality tobaccos are widely cultivated.

Cotton and sugar beets, formerly grown in the Ararat Plain, are being succeeded by more valuable crops, such as grapes. The area under grain crops has been sharply reduced.

Extensive alpine pastures raise the productivity of animal husbandry, whose main branches are the raising of beef and milk cattle and sheep. Pig and poultry raising, as well as sericulture and apiculture, play subsidiary roles.

Transportation. Mountains are a serious impediment to the construction of land transport routes of any kind, although distances between towns and regions are not great. A railway line, leading to Tbilisi in the north and Baku in the east, runs through the northern, western, and southern regions of the Armenian S.S.R. Yerevan is linked with the Sevan basin by a line running along the Razdan River. Clustered along the rail routes are major industrial centres.

Road network

The network of motor roads is much denser, with Yerevan as the main hub. Road transport—whose role in freight and passenger traffic is growing—carries many times as much cargo as the railways, and buses remain the chief mode of travel between towns and villages.

Local air transport is being used increasingly. Air routes link Yerevan with many towns in the republic, and planes carry fresh fruits and grapes to Moscow, Lenin-

grad, and other Soviet cities. Gas pipelines link Armenia with the Azerbaijan and Georgian S.S.R.'s.

These transportation facilities enable the other union republics to supply Armenia with coal, petroleum products, metals, timber, grain, industrial raw materials and equipment, while Armenia, in turn, exports lighter but more valuable products.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitutional framework. The state structure of the Armenian S.S.R., defined by the constitution adopted in 1937, is similar to those of the other union republics.

The highest organ of state power is the Supreme Soviet of the Armenian S.S.R., elected for a four-year term from a single list of candidates. The Supreme Soviet designates its Presidium to administer the state in the periods between its own brief sessions, and it appoints the Council of Ministers and enacts laws.

The organs of state power in the districts, cities, and rural areas are the soviets of working people's deputies, elected by the population for two years.

The Supreme Court of the republic, whose members are elected by the Supreme Soviet for five years, is the highest judicial body. The procurator of the Armenian S.S.R. is appointed, for five years, by the procurator general of the Soviet Union. The Armenian S.S.R. has a coat of arms, a state flag, and a national anthem.

Communist Party

Political life is under the direction of the Communist Party, part and parcel of the Communist Party of the Soviet Union. It was founded in 1920 through the unification of Armenia's Bolshevik organizations. In 1922 it had a membership of 2,000 with a further 3,000 candidate members. By 1974 the Communist Party of Armenia had 137,000 members and candidates. The Young Communist League (Komsomol) of Armenia has a membership of more than 339,000. It was formed in 1921 in close contact with the party.

The trade unions of Armenia in 1974 had a membership of 1,200,000. By 1970, they operated 60-odd clubs and palaces of culture, nearly 100 public libraries, nearly 300 sports facilities, sanatoriums, and rest homes.

Education. By the 1970s, countrywide eight-year schooling was being introduced. In the 1973–74 school year about 1,600 general schools existed, with a combined enrollment of about 700,000 younger pupils. Forty buildings cater to the children's organization known as the Young Pioneers, and there are a number of sports establishments for children. There are also clubs for young technicians, nature lovers, and the like. Education of all types is free of tuition charges, being supported by taxes.

The republic in the mid-1970s had 50 trade schools (30,000 pupils), 64 secondary specialized educational establishments (more than 51,000 students), and 12 other institutes and colleges (52,800 students). Establishments of higher learning include Yerevan State University; polytechnical, medical, agricultural, pedagogical, and theatrical institutes; and a conservatory.

There are about 1,300 public libraries. They include the A.F. Myasnikyan State Public Library of the Armenian S.S.R. and the Matenadaran archives in Yerevan containing 10,000 Armenian manuscripts, the largest collection in the world. There are also about 1,200 clubs and 35 museums, including a comprehensive State Historical Museum of the Armenian S.S.R.

Social services. In the mid-1970s there were more than 230 hospitals with about 24,000 beds, and about 500 outpatient clinics. There was one doctor for each 315 people. The republic also has some two dozen sanatoriums. Medical treatment in hospitals and clinics is free of charge for all citizens, being supported, like education, by taxation.

More than 1,500,000 persons were rehoused between 1961 and 1970, some in the 21,000 to 26,000 apartments built annually, others in new houses built by the state, and others in new houses built by individuals on their own account.

Cultural life. Armenian written literature began in the 5th century AD, and monasteries became the principal centres of intellectual life. The earliest works were historical, Moses of Khoren's *History of Armenia* being representative. The masterpiece of classical Armenian is Eznik Koghbatzi's *Refutation of the Sects*. The first great Armenian poet (10th century) was St. Gregory Narekatzi, renowned for his mystical poems and hymns. During the 16th–18th centuries, popular bards, or troubadours, called *ashugh*, arose, outstanding among them being Nahapet Kuchak and, especially, Aruthin Sayadian, called Sayat-Nova (died 1795), whose love songs are still popular. In the 19th and early 20th centuries, Hakob Paronian and Eryand Otian were outstanding satirical novelists, and Grigor Zohrab wrote Realist short stories. Paronian was also a comic playwright, whose plays still entertain Armenian audiences. The most celebrated novelist was Hakob Meliq-Hakobian, called Raffi, and perhaps the best dramatist of recent times was Gabriel Sundukian (died 1912).

Among Armenian composers, Aram Khachaturian has achieved world-wide renown.

By the mid-1970s, publishing houses were producing more than 1,000 books (including more than 700 in Armenian), with a total of 9,800,000 copies per year. The broadcasting system has been operating since 1926 and the Yerevan television centre since 1956. Broadcasts and telecasts are conducted in Armenian, Russian, Azerbaijani, and Kurdish.

The republic boasts a State Academic Theatre of Opera and Ballet, several drama theatres, theatres for children, orchestras, a national dance company, and the Yerevan film studios, which produce feature, documentary, and popular science films. The traditional folk arts, especially singing, dancing, and artistic crafts, are thriving.

Academy of Sciences

Armenian science, like its culture, has its roots in antiquity, but research institutions are a 20th-century development. The Armenian Academy of Sciences is composed of 39 institutes engaged in research problems in natural and social sciences.

PROBLEMS AND PROSPECTS

The chief problem to be dealt with in the course of the future economic development of Armenia is believed to be the structural improvement of the production complex as a whole. There is also a growing need for a more efficient utilization of the republic's relatively limited natural resources—the mineral wealth, lands, and waters—and, consequently, for a comprehensive development of those process industries that require skilled labour for a more effective utilization of raw materials.

(A.A.M./Ed.)

Armour

Armour refers to any body covering designed for protection in combat. From very early times, armour for the fighting man and his horse was improved steadily to counteract improvements in weapons and, to some extent, tactics. For the greater part of its history only excavated fragments survive; the main evidence is found in art, where its interpretation is often difficult and subject to argument. Development was at first slow, but in Europe, between the 13th and 17th century, improvement of blade-making techniques and the introduction of more efficient missile weapons helped accelerate the development of plate armour. In the 16th and 17th centuries, improved firearms forced armourers to increase the thickness and, therefore, the weight of their product, until finally plate armour was largely abandoned in favour of increased mobility. However, armour never entirely disappeared among European cavalry, surviving in the form of a hat lining or helmet, a cuirass for the body, and high boots, until modern times. Elsewhere it survived until rendered obsolete by the introduction of firearms.

The large number of head wounds caused by overhead shell bursts during World War I led to the adoption of steel helmets by all combatant nations. Renewed interest in experiments with various types of body armour began at the time of the American Civil War, and such armour was used on a limited scale by both sides in both world wars. Continuing experiments now aim at development



Figure 1: (Left) "Coat of a Thousand Nails," reinforced with steel plates, from Rājput, India, 18th–19th century. In the Wallace Collection, London. (Centre) Mail shirt, from Sinigaglia, near Bologna, Italy, first half of the 14th century. In the Royal Scottish Museum, Edinburgh. (Right) Japanese armour, the cuirass and shoulder guards of lamellar construction, *c.* 1800. In the Armouries, Tower of London.

By courtesy of (left) the trustees of the Wallace Collection, London, (centre) the Royal Scottish Museum, Edinburgh, (right) the Ministry of Public Building and Works, British Crown Copyright; photograph, (centre) Tom Scott

of body armour that is bulletproof yet sufficiently light and flexible for the infantryman to wear in action.

TYPES OF ARMOUR

Armour falls into three main groups, depending on its construction: leather or fabric, sometimes of several thicknesses, sometimes reinforced by quilting with some material such as cotton; mail, consisting of interlocking iron or steel rings; and rigid armour of metal, whalebone, ivory, horn, bark, wood, plastic, or *cuir-bouilli* (leather hardened by boiling in wax). The last group can be further divided according to the methods of connecting the pieces: (1) small overlapping plates attached outside a leather or fabric garment (scale) or attached inside the garment (brigandine); (2) lamellar, which consists of small overlapping plates held together by laces; (3) small plates held together by mail or let into mail garments; (4) large plates, usually of metal, linked by loosely closed rivets and by internal leathers to allow the wearer maximum freedom of movement.

Cuir-bouilli

Leather armour. Armour of hide is probably the oldest of all types, and in its simplest form is indistinguishable from ordinary clothing. Coats made of five to seven layers of rhinoceros skin were worn in China in the 11th century BC, and an apparently similar armour of ox hide was used by the Mongols in the 13th century AD. North American Indians, such as the Shoshoni, wore jackets of several layers of hide glued or sewn together and also had leather horse armour. The jaguar skin jackets of Aztec chiefs seem to have been part of their regalia and not defensive. Coats made of stout buff leather were first worn under European plate armour in the 16th century. The leather sleeves and skirts, retained after plate arms were abandoned, were sufficiently sturdy to deflect a sword cut, and buff leather was used for the cuffs of cavalry gauntlets until the 19th century. Leather has also been used widely as a base for other defenses, such as the boars' tusk helmets of Mycenae and a much later jacket covered with modern coins found in Alaska.

Fabric armour. Early evidence of fabric armour is rare since fabric rarely survives on archaeological sites. A fragment consisting of 14 layers of linen was found in a Mycenaean grave of the 16th century BC, and there are occasional literary references throughout antiquity. The Assyrians of Xerxes are said to have had linen armour, and the Greek heavy infantry of the 5th–4th centuries BC wore a linen cuirass in preference to bronze, probably to increase their mobility. Throughout the Middle Ages quilted coats (aketons) were worn either alone or under mail or plate to prevent chafing. Silk quilted armour was experimentally introduced in England in the 17th century. In northern India quilted coats, sometimes with helmets and trousers to match, were worn until the 19th century and were usually velvet covered and studded with small gilt nails, with some steel plates let into their surface (see Figure 1, left). Quilted coats and helmets were also worn in China and Korea. The short jackets, with or without sleeves, illustrated in Central American art, may be the quilted armour reported by the conquistadores. Horse armour of quilting has been used in many areas, from England in the 13th century to Nigeria in the 20th. Particularly elaborate rope armours made from coir, a coarse fibre obtained from coconut husks, were worn until the 19th century in the Gilbert and Ellice Islands. Fabric armour is used today in, for example, the U.S. Navy's four-pound, synthetic-felt combined life jacket and armoured vest.

Mail. The place and date of the origin of mail is at present unknown, but pieces have been found on a site near Kiev dating from the 5th century BC. Surviving examples are of iron or steel wire, occasionally with brass rings around the edges or, in Oriental mail, forming an all-over pattern. Mail is flexible and relatively impervious to slashing strokes when worn over quilting, although a thrusting weapon can force the rings apart in spite of their rivetted closure. In the form of a simple shirt, mail was worn throughout the Roman Empire and beyond most of its frontiers (see Figure 1, centre). It formed the

The mail shirt

main armour of western Europe until the 14th century, with leg harnesses of mail appearing in the 11th century, mail hoods added during the 11th–13th century, and long sleeves ending in mail gloves added during the later period of its use. The mail shirt and aketon were worn in India and Persia until the 19th century, and in the Sudan and Nigeria until modern times. After the development of complete plate armour in Europe, mail gussets were laced to the arming clothes to close the gaps in the plates. A curtain of mail was often attached to the lower edge of the conical helmet, acting as a throat and neck protection, on the 14-century European basnet and also on many Oriental helmets, particularly Indo-Persian. Sleeves of mail were used by light troops in the 16th century, and mail shoulder straps were adopted by light cavalry in the 19th century. The Japanese used mail to a limited extent from the 14th century AD, but the Japanese rings were arranged in a variety of ways, producing a more open construction than found in Europe.

Scale armour. Although usually of metal, scale armour has been made of horn, bone, and leather and from the scales of such naturally armoured animals as the pangolin. The small plates usually have a curved or pointed lower edge overlapping a joint in the next row. The area of distribution of scale armour differs slightly from that of lamellar. Scales have been found on an Egyptian site of the 17th century BC. It is found throughout the Middle East; in the classical Greek area including southern Russia; in the Roman Empire; and in all of Europe, particularly in the 10th–12th centuries AD, but occasionally as parts of an armour until the 15th century; and in eastern Europe in the 17th century. It was worn in India and China until the early 19th century and was readopted at that time by European heavy cavalry for use as shoulder defenses.

Brigandine. The origin of brigandine is at present unknown. Although studs are shown on a garment represented in art, there is no proof of the presence of plates within the garment. Brigandine apparently was in use in the 8th century AD in China and later in Korea; in both places it was worn until the 19th century. From the 18th century some Chinese brigandines are purely ceremonial, made with rivets but without plates. A number of cuirasses made up of many plates originally rivetted to a cloth cover have been excavated on European sites of the 14th century, and apparently similar garments occur in European paintings from about 1240. These remained the most common form of body armour until about 1400, after about 1360 usually with a one-piece breastplate also under the cover. Short brigandine coats remained popular for European light troops until about 1600, particularly in a less costly version—the jack—with its plates simply tied in. Brigandine construction was also used for other parts of the armour, including gauntlets and thigh defenses. Fifteenth-century Persian and later Mughal miniatures show what are probably brigandine coats.

Bulletproof jackets

The construction of most modern bulletproof jackets is based on the principle of the brigandine or jack. Small overlapping plates of alloy steel line a waistcoat and a groin protection. The plates are not rivetted but are contained in pockets within a vest of synthetic fibre or snapped into a plastic framework. Metal is now being replaced by such synthetics as fibre glass or boron carbide, making the garment less cumbersome and more bullet resistant.

Lamellar armour. Lamellar armour consists of strips made up of numerous narrow plates with their long axes vertical, each overlapping its neighbour. The strips usually overlap upward. This armour is more flexible than scale and cheaper to manufacture than mail. Although illustrated in Egyptian paintings of the 15th century BC, lamellar probably originated in Persia and was carried to western Europe by the Avars and to the Far East by other nomads. It was usually made up into a shirt, sometimes with capelike sleeves or, in the East, with a separate bolero to which the sleeves were attached. Lamellar was worn in Persia until the 16th century, during its later use with plate forearm and shin defenses. The construction reached China and Korea by the 1st century AD, passing to Japan in the 5th century and achieving its greatest elaboration in the colourfully laced and lacquered armours of the Japanese Heian period (794–1185). Lamellar remained in use in Japan (see Figure 1, right) until armour was abolished in 1867 and was still employed in Tibet, for man and horse, early in the present century. Although it is illustrated in Mughal miniatures, no actual example of lamellar has been found south of the Himalayas. The simple cuirasses of wooden slats and rods laced together worn by North American Indians before the introduction of firearms probably derive from the bone and ivory lamellar armours of Siberia and Alaska.

Plate and mail armour. Small plates linked by mail or let into mail garments first appeared in Persia and Turkey in the 15th century AD in the form of hoops around the body with smaller overlapping plates in the arms and skirts. Sixteenth-century Russian and Turkish armours of that construction had a large circular breastplate let into the mail and narrow vertical rows of small plates at the back. Full armours of this type, including helmets and trousers, were used in northern India in the 17th–19th centuries. Horse armours, and even elephant armours, were made in the same way. Similar armour worn by the Moros of the Philippines, made with brass mail and plates of either brass or horn, probably derived from contacts with the Arabs, although their helmets were based on European models.

HELMETS

The head, the most vulnerable area, has been protected by plate since early times. The basic form of most helmets is a conical or hemispherical bowl, sometimes extended downward to protect the base of the skull, as shown in Sumerian monuments of about 3000 BC. This type, with or without a neck guard, cheek or ear pieces, with or without a mail or scale curtain to protect the throat and neck, is found in almost all periods and areas. The bowl was sometimes made up of several plates rivetted or laced together, a form carried to extremes in Japan, where a bowl sometimes consisted of as many as 62 segments. Occasionally, as in Persia, Turkey, and Russia in the 16th century, the helmet reached an acute apex. In Turkey, in the 15th century, a peak was added over the eyes with a movable nasal or nosepiece. The prestige of the Turkish armies made this form widely popular in Europe until about 1650. The movable nasal was also adopted in Persia and India. The European helm of the 13th–14th centuries developed from the uniting of the neck guard with a nasal that had gradually spread to cover all of the face but the eyes. The helm never completely ousted the basic bowl forms, which from about 1300 were often fitted with a movable visor. Close helmets, made by hinging or pivoting plates to the bowl, completely enclosing the head and turning with it, appeared in western Europe about 1410, remaining in use until about 1650. Light troops continued to wear open-faced helmets, usually peaked or brimmed, until the late 17th century. From the second half of the 15th century the bowl of European helmets was strengthened by a fore and aft comb, increasing in height until about 1600.

The visored helm

ANCIENT AND MEDIEVAL ARMOURING

A complete bronze armour, consisting of a cuirass, a deep skirt of overlapping hoops, broad shoulder defenses, and short greaves to protect the lower leg, was found in a Mycenaean grave of about 1450 BC. As the disappearance of the war chariot necessitated lighter equipment, Greek armourers of the 7th–5th centuries BC produced a combination worn by the Greek foot soldier: cuirass, long greaves sprung onto the shins, and a deep helmet—all of bronze. The large, round shield of bull's hide, sometimes faced with bronze, had both a handgrip and a forearm brace. Later bronze cuirasses, modelled closely on the muscles of the torso, extended down over the stomach, a type also worn by Roman commanders, and possibly at Byzantium, and revived for parade use in the Renaissance.

Roman steel armour

The corselet of the Roman legionary, worn between the end of the 1st century AD and the beginning of the 3rd century, consisted of a cylindrical cuirass made up of some seven horizontal hoops of steel, with openings at the front and back, where they were laced together. The cuirass was buckled or hooked to a throat piece flanked by several half-hoops protecting the shoulder. The individual plates were linked internally by articulating leathers, as on all later plate armour, allowing free movement.

Apart from helmets, armour of large plates was probably unknown in western Europe during the Dark Ages, but in the late 12th century references occur to the *cuirie*, a leather body armour reinforced with plates. During the 13th and 14th centuries the entire body and limbs were gradually enclosed in plate, although the body armour itself, until 1400, was usually of brigandine, sometimes having a steel breastplate fixed over it. The complete steel body armour, with no cover and consisting of breast, back, and a hooped skirt opening down the side, appeared about 1400. The design of the armour was perfected during the 15th and 16th centuries, with added reinforcement, and adapted for a variety of combats. By the 16th century a good-quality armour would normally have had a number of pieces of exchange to adapt it for use in the field or tournament, mounted or on foot (see Figure 2).

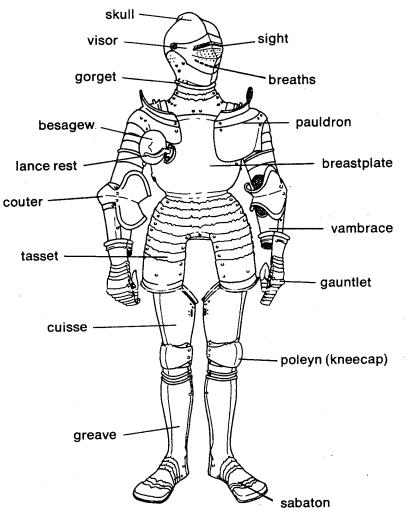


Figure 2: Complete armour, South German, about 1510.
By courtesy of the trustees of the Wallace Collection, London

Wealthy nobles and princes had great series of armours, decorated to match, for use in a variety of combinations. The major European manufacturing centres were in iron-producing areas—Milan, Brescia, Innsbruck, Augsburg, Nürnberg, and Landshut. Besides finely etched and gilded armours for the aristocracy, large numbers of uniform armours were turned out to equip the specialized regiments that developed during the 16th century. Most courts employed a small group of armourers to supply armour for their own immediate needs and to be presented as diplomatic gifts. Although full armour still appears in portraits of commanders in the 18th century, all but the cuirass and helmet had actually been discarded by about 1650.

Plate armour, consisting of forearm defenses and four large plates around the thorax, was worn in Persia and India until the 19th century, usually over a mail shirt. There was an increasing use of plate in Japan, sometimes disguised as lamellar, from the 15th century.

BIBLIOGRAPHY. A.M. SNODGRASS, *Arms and Armour of the Greeks* (1967); H.R. ROBINSON, *Oriental Armour* (1967); and C. BLAIR, *European Armour, circa 1066 to circa 1700* (1958), are three completely reliable works basic for the study of this subject. Extensive bibliographies are included in G.C. STONE, *A Glossary of the Construction, Decoration and Use of Arms and Armor in All Countries and in All Times* (1934); SIR GUY LAKING, *A Record of European Armour and Arms Through Seven Centuries*, 5 vol. (1920); and SIR JAMES MANN, *Wallace Collection Catalogue of European Arms and Armour*, part 3 (1944) and vol. 2 (1962). P. COUISSIN, *Les Armes romaines* (1926), although now somewhat dated, has not been replaced as the major study of Roman armour. Other interesting works are B. DEAN, *Helmets and Body Armor in Modern Warfare* (1920), on the modern period up to and including World War I; and W. HOUGH, "Primitive American Armor," *Annual Report of the Board of Regents of the Smithsonian Institution*, part 2, pp. 631–680 (1893–95), the only writer to deal with the armour of the American Indians. A bibliography of new works is published as they appear in the *Journal of the Arms and Armour Society* (1953–).

(A.V.B.N.)

Arms Design and Decoration

The design of arms for hunting or fighting usually has been determined by the specific purpose for which the arms are to be used. Combat weapons must be designed to cut, pierce, or crush, while armour must absorb shock or offer a glancing surface to deflect the enemy's weapon. At the same time, the embellishment of arms and armour expresses man's artistic nature. In combat armour, the decorations must be subordinated to the functional design; for example, the deflecting surface must be left intact. In parade or dress armour, however, decoration may be used for its artistic effect alone.

Decorations also have had functional purposes; for example, they acted as a means of identification. Highly ornamented battle armour and weapons often served to identify a leader. The emblazoning of shields, one of the oldest and most widespread types of arms decoration, was another means of identifying combatants.

In many cultures the right to bear arms was a privilege, and arms and armour therefore became symbols of socio-economic status. As a mark of distinction, they often were decorated so extravagantly that they could be regarded as jewelry, made in fanciful shapes from precious materials and free of all functional considerations. Hunting weapons were usually more elaborately decorated than military arms because of their association with sport and pleasure. The decoration of offensive weapons varied widely according to their function and the status of the bearer. The mass-produced pole arm of the common foot soldier could be "plain as a pike staff," while the decoration on the dagger scabbard of his captain might be highly elaborate and worked or designed by famous artists.

Psychological and symbolic aspects

Psychological effect has long been an important factor in the design and decoration of arms and armour. In the *Iliad*, the ancient Greek epic by Homer (flourished 9th century BC), "the dreadful crest nodding from the helmet's peak" served to overemphasize the height of the fighter to overawe and intimidate the enemy. Grimacing war masks, *menpo*, were worn as a part of the combat armour of the Japanese samurai war lords.

Symbolic considerations also affect arms design. A Christian knight's sword, for instance, was interpreted as a cross, and his triangular shield was symbolic of the Trinity. A Malayan kris blade if wavy symbolizes a running snake, and if straight, a sleeping snake. The hilt of a kris, though often of a highly stylized design, represents the eagle-demon Garuda as a helper in battle. A circular opening in the top of a samurai helmet was designed to allow the spirit of the Shintō war god Hachiman to enter the fighter's body, but its practical purpose was ventilation.

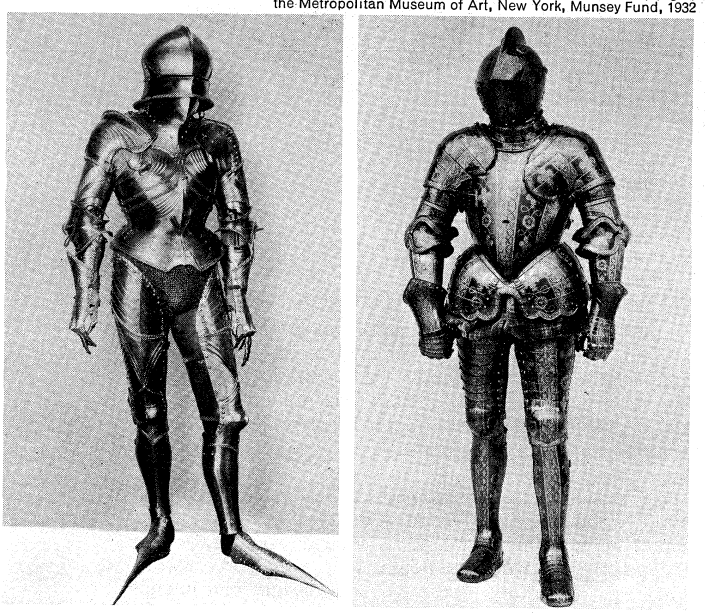
The decoration of arms and armour in most cases was subject to the overall design of the object. Usually accentuating the functional design, the ornamentation was located at the brow, or crest, of a helmet; on the upper part of a breastplate; on the face of a shield; on the hilt of a sword; on the fuller, or groove, of a blade. Swords intended for combat seldom have a decorated grip because the ornamentation would be hidden by the clenched fist. Sword decoration was concentrated on the hilt and on the parts of the blade left untouched by grinding. The obvious places for gun decoration were the lock and stock, which was often ornamented with relief carving, inlay, and mountings of precious metals. The barrel might be decorated over its entire length or only at the rear. Care was required in cutting the reliefs, which

could weaken the metal. Master craftsmen sometimes engraved delicate ornamentation on the inner parts of gun locks, which were seen only when the lock was taken apart for cleaning.

Stylistic aspects

The design of arms and armour generally reflects the prevailing artistic style of the period or region. In European design, for example, there is a striking difference in style between 15th-century armour of the late Gothic period and 16th-century armour of the Renaissance. In Gothic armour, the vertical line is accentuated, and the general outline is dominated by a spiky elegance comparable to tracery in Gothic architecture. In Renaissance armour, the horizontal is emphasized, giving a generally rounded and massive appearance, like Renaissance architectural forms. In Gothic armour, which was more often used in combat, decoration was subordinated to utilitarian considerations; in Renaissance armour, which was more frequently used for display, the decoration was more elaborate.

By courtesy of (left) the Kunsthistorisches Museum, Vienna, (right) the Metropolitan Museum of Art, New York, Munsey Fund, 1932



Comparison of Gothic and Renaissance styles of armour.
(Left) Gothic armour of the archduke Sigismund of Tirol, made in Augsburg, Germany, *c.* 1480 by Lorenz Colman (also called Lorenz Helmschmied). In the Kunsthistorisches Museum, Vienna. (Right) Blued, etched, and gilded Renaissance armour of George Clifford, 3rd earl of Cumberland, made in the royal workshops at Greenwich, England, *c.* 1590. In the Metropolitan Museum of Art, New York.

ORNAMENTAL MOTIFS AND DEVICES

The range of motifs used for the embellishment of arms and armour is endless, though, ironically, scenes of fighting and killing are rare. Animals are among the figurative motifs most widely used in all cultures and ages. They are commonly used on hunting arms but also are found on battle arms and armour. Floral motifs, either as individual decorative elements or as a background embellishment for narrative scenes, are often used, particularly in the Islāmic world, where the faith prohibits figural representation. Heraldic devices, such as coats of arms and other identifying insignia and emblems, are widely used. Calligraphic or written ornamentation was extensively employed on Islāmic arms and armour. Examples of calligraphic decoration also are found on ancient Egyptian and Mesopotamian weapons. The writing that appears on medieval and Renaissance European arms usually consists of inscriptions or protective magic formulas. The maker or place of manufacture was sometimes decoratively inscribed on European arms, as in the swords produced in Toledo, Spain.

Effects of materials and techniques

Decorative effects often were related to the materials and techniques of manufacture. Metals, for instance, in addition to their practical values, have long been appreciated for their decorative appearance—the golden shimmer of bronze, the cold glint of steel, the colourful contrast of gilding on iron. Decoration achieved through materials alone is exemplified on African shields, from ancient Egypt to the Zulu, on which the natural spotting of cowhide or leopard skin creates the decorative effect. The beauty of modern weaponry usually depends upon precision craftsmanship, and the decorative qualities of the material, such as the fine grain of a wooden gunstock.

The technique of fabricating pattern-welded (Damascus) steel produces a decorative effect of its own. When a blade, or gun barrel, is shaped from numerous steel strands of different carbon content, the surface of the steel acquires a mottled appearance, which can be arranged into regular patterns by twisting or braiding the strands. Vikings tended to see the coils of dragons or the scaly hides of snakes in these patterns called *wurmbunt*, while in Islāmic countries such patterns as "Muhammad's ladder" and the "chain of vertebrae" were highly prized. Nineteenth-century gunsmith virtuosos might

By courtesy of (left, top right) the Metropolitan Museum of Art, New York, (left) anonymous gift, 1950, (bottom right) the Kupferstichkabinett Basel



Variety of ornamental motifs.
(Left) Iron helmet with floral and calligraphic decoration of silver damascening, Turkish (Mamlūk), 15th century. In the Metropolitan Museum of Art, New York. (Upper right) German gun stock with animal motifs, *c.* 1600. In the Metropolitan Museum of Art, New York. (Bottom right) Pen drawing for a dagger scabbard decoration, showing the "Dance of Death," by Hans Holbein the Younger, *c.* 1532. In the Kupferstichkabinett, Basel.



Different techniques of decorating helmets.
(Left) Embossed steel parade helmet made by Filippo Negroli of Milan, probably for Francis I of France, Italian, 1543. (Centre) Bronze helmet with applied gold relief ornamental figures, Elamite, *c.* 1300 BC. (Right) Helmet with "schembart" visor and German style embossing, etching, and gilding, German, *c.* 1510. All in the Metropolitan Museum of Art, New York.

By courtesy of the Metropolitan Museum of Art, New York, (left) gift of J. Pierpont Morgan, 1917, (centre) Fletcher Fund, 1963, (right) Rogers Fund, 1904

spell out their names in elaborate calligraphies on Damascus steel gun barrels. The edge of the Japanese sword was hardened by coating the pattern-welded blade with clay, scraping the edge clean, and then alternately heating it and quenching it, by thrusting it in water. The hardened parts became lighter with polishing, and a skillful swordsmith, by dexterous scraping, could produce designs such as waves or scenic views.

Migration of motifs

Decorative motifs were usually disseminated through the trade or conquest of decorated objects or by the migration of the artists. Etruscan armour based on Greek prototypes suggests the first method, and Scythian gold scabbards of Greek workmanship indicate the second. Both methods might have been responsible for the spread of the spiral motif on Bronze Age weapons or of gold and garnet incrustations on the sword mountings of Central Asian steppe nomads and Germanic tribes during the migration period of the 5th to the 7th centuries AD.

After the invention of printing in the West in the 15th century, book decorations were exploited for ornamental motifs. By the early 16th century, model books of ornamental motifs were published for decorative artists. Prints by famous artists or graphic reproductions of well-known paintings were sometimes adopted either in their entirety or in part. From the 17th to the 19th centuries in Japan woodcuts served as sources for decorative devices on sword furniture. Major artists in both the East and the West have made drawings specifically for the embellishment of arms, such as those by Hans Holbein the Younger (1497?–1543) for dagger scabbards.

TECHNIQUES

Arms techniques are as diverse as the materials used. Ornamentation is either three-dimensional (*e.g.*, carving, embossing) or two-dimensional (*e.g.*, colouring, etching, inlay). In many objects both types of ornamental techniques are used. The construction of plate armour was, in fact, the forming of a sculpture in steel.

Incised decoration of arms is probably as old as carving, both techniques dating from about 30,000 years ago. At Ice Age archaeological sites in the Pyrenees mountains of Spain and France (Le Mas d'Azil, Bédeilhac, Bruniquel), spear-throwers decorated with animal carvings have been found, while engraved ornamentation was preferred by the Magdalenian inhabitants of such Pyrenean sites as Lorthet or Teyjat in the French *département* of Dordogne.

Etching

Etching of metallic surfaces was developed by medieval armourers in Europe. In the 15th century it was noticed that accidental scratches in the blacking—a metallic preservative—became imprinted by corrosion into the metal surface of the armour itself. An acid-resistant mixture of wax and resin was developed for coating polished steel surfaces. Designs were either scratched with a needle into the coating (Italian method) or painted with a brush onto the metal (German method). In both methods, acid was poured over the treated area, eating away the uncovered portions. Because the relief of the etching was so shallow, the glancing surface of the armour plate was not impaired, and etching became the most popular form of decoration for combat armour. The colour could be heightened by gilding or blackening.

Carving lends itself obviously to the embellishment of stone, wood, and bone weapons, the earliest known arms. Small parts of metal weapons, such as the gunlock or the pommel (the counterweight at the end of a sword hilt), were frequently decorated with designs and images cut in relief or in the round.

Embossing, or raised surface ornamentation, has been used since ancient times. The rough shape of the decoration was hammered out from the reverse side, using the yielding surface of a block of wood, lead, or pitch as an anvil. Fine details were hammered, filed, chiselled, or punched from the right side.

Leather was tooled and used for armour or arms accessories such as scabbards and sword belts. The dyeing of leather was widespread, particularly in the Islāmic world. Highly sophisticated stencilled designs in various colours appeared on the leather parts of Japanese armour from the 12th to the 16th centuries. *Cuir-bouilli*, or leather that had been softened by soaking and then molded, embossed, or stamped before drying, was hard enough to be used for lightweight armour.

Since ancient times arms and armour have been decorated by inlay or the application of ornament. Openwork figures and animals set in cutouts, mosaics of various materials ranging from stone to feathers, and inlaid wood and ivory have been used in various periods and regions. Even in the 20th century, inlaying of precious and semiprecious materials is used for decorative effect, as in a mother-of-pearl inlaid revolver handle. Mountings of colourful metals, such as plaques of embossed and engraved gold, brass, copper, and silver or settings of precious stones have been widely used. The surmounting of helmets with such decorative devices as crests and masks has been widespread. Feathers, shells, and beads are common decorative materials in less technically developed cultures, as are decoratively woven or plaited fibrous materials, such as the coconut fibre armour of the Gilbert Islanders in the Pacific.

Colouring techniques

Metallic arms and weapons could be decoratively coloured through the use of heat, a process usually called blueing. Skillful use of heat could produce not only blue but also shades of russet and purple. Occasionally, European medieval and Renaissance armour was blackened by burning linseed oil into the hammer-rough surface. Besides furthering the psychological effect of a "Black Knight" or a "Black Guard," this technique was a rust-proofing device. Colouring was also achieved by gilding and occasionally by coating with silver foil. Both for pro-

tection from rust and for colour effect, the Japanese lacquered the scales of their lamellar armours, usually in gold or black. Painting on metal was commonly used by European armorers of the 13th to the early 16th centuries to enliven the metallic surfaces with colour and decorative imagery.

Another surface treatment with colour effect was damascening. In this technique wires of soft metals such as gold, silver, or sometimes copper are either inlaid into grooves with a dovetail cross section cut into the steel or are more simply hammered onto a roughened cross-hatched surface. The first method is more durable, but much more time consuming. Damascening was widely used in ancient Mycenae and in other parts of Europe during the early Middle Ages, especially among the Vikings of Scandinavia and the Moors of Spain, and in Islāmic countries, where it remains one of the most important methods of arms decoration.

PERIODS AND CENTRES OF ACTIVITY

Among the earliest artifacts made by man are examples of decorated weaponry. During the stone ages elaborately carved and incised spear-throwers made of staghorn and bone were produced, especially in southwestern France and northern Spain. Finely worked flint blades with their chippings arranged in regular patterns were made throughout the prehistoric world.

Arms of the ancient Western world

Ancient Egyptian and Mesopotamian civilizations produced richly decorated arms of stone, ivory, and metal as did the Phoenecians and Mycenaeans. Other important centres of arms decoration in the pre-Graeco-Roman ancient world were in Anatolia among the Hittites and in Iran during the Luristan culture. In northern Europe the bronze weapons of the Hallstatt and La Tène cultures have spiral decorative motifs that were probably derived from Mycenaean arms.

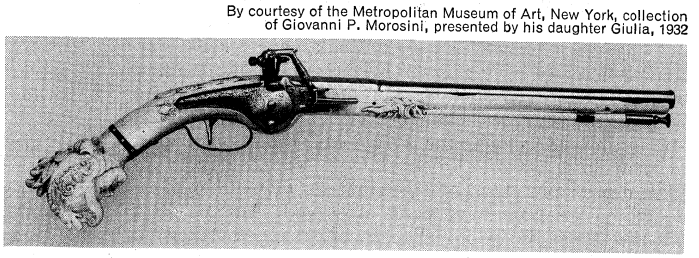
The decoration of arms and armour in ancient Greece was relatively restrained compared with the opulence of contemporary Persian work. Each Greek region, such as Attica, Corinth, or Sparta, had its own decorative style. Etruscan and Roman arms and armour were deeply indebted to Greek tradition.

Graeco-Roman arms decidedly influenced the contemporary weaponry of northern European Teutonic and Celtic tribes, while the design and workmanship of Persian arms influenced the Thracians of the Lower Danube region in the Balkans. Also at this time there flourished in the steppes of eastern Europe and Central Asia a decorative style using bizarre, contorted animals. This so-called animal style is best known through the works of the Scythians. In northwestern Europe the Celts developed a flamboyant geometric style, which was particularly prominent in the British Isles.

With the 4th-century invasion of Europe by the Huns and other steppe nomads, a new and highly distinctive style of almandine and gold incrustation spread through the Germanic tribes. The major arms-producing centre of northern Europe during this early medieval period became the German Rhineland around Cologne, where the Celtic, Germanic, and Roman cultures met. Also in the early Middle Ages, the Byzantines and Sāsānian Persians produced individually styled weapons of elaborate design, although few examples have survived.

In contrast to the Christian crusaders, whose arms and armour were simple in design and relatively austere in decoration, the warriors of Islām, particularly the Mamlūks, delighted in richly decorated arms. The important centres of manufacture were Alexandria and Damascus. Islāmic Persia, Turkey under the Ottomans, and India under the Moghuls excelled in cut steel and damascened work in scimitar and dagger hilts of Indian carved jade, which were exported to be mounted locally.

By courtesy of the Metropolitan Museum of Art, New York, collection of Giovanni P. Morosini, presented by his daughter Giulia, 1932



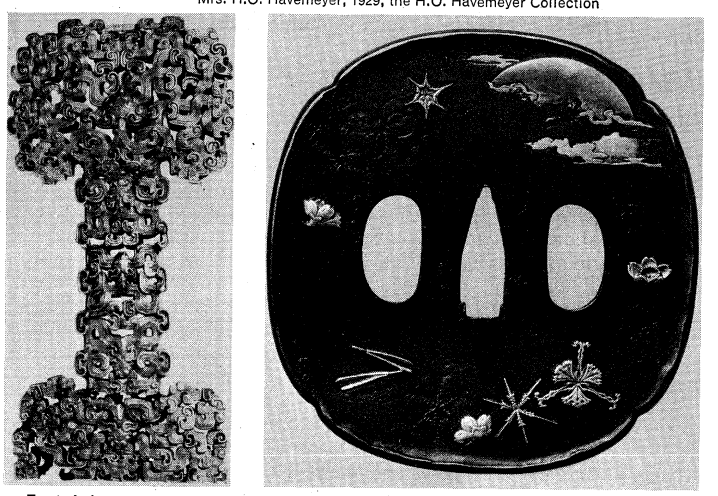
Dutch pistol with carved ivory stock, made in Maastricht, 1690. In the Metropolitan Museum of Art, New York.

15th- and 16th-century armour

The design, construction, and decoration of armour was especially outstanding from the 15th to the 17th centuries. The most renowned centre of production, Milan, Italy, produced many of the greatest armourers, notably such family dynasties as the Negrolis and the Missaglias and individual geniuses such as Lucio Piccinino (*c.* 1535–after 1595).

Initially influenced by the Milanese workmanship, independent styles of design and decoration emerged in northern Europe in the 16th century, centred in Nürnberg, Augsburg, Innsbruck, Landshut, and Cologne. The great workshops were those of the Seusenhofers in Innsbruck, Helmschmieds in Augsburg, Kunz Lochner and Valentin Siebenbürger in Nürnberg. They were assisted by famous etchers, such as Jörg Sorg and Daniel Hopfer.

By courtesy of (left) the trustees of the British Museum, (right) the Metropolitan Museum of Art, New York, bequest of Mrs. H.O. Havemeyer, 1929, the H.O. Havemeyer Collection



East Asian arms decoration.
(Left) Gold dagger hilt, Chinese, 5th–3rd century BC; Chou dynasty (1122–221 BC). Height 12 cm. In the British Museum. (Right) Metal inlaid *tsuba* made by Gotō Hokkio Ichijō, Japanese, 19th century; late Edo period (1603–1867) or Meiji period (1868–1912). 7.94 cm × 7.62 cm. In the Metropolitan Museum of Art, New York.

Sixteenth-century court workshops, such as those of the emperor Maximilian I at Innsbruck, of Henry VIII at Greenwich near London, and of Henry II of France at Paris (Louvre school), produced especially rich armour for state and presentation purposes. The Royal French workshop under Étienne Delaune (*c.* 1519–83) and that of Elisaeus Libaerts in Antwerp specialized in lavishly embossed armour. While German and English armour kept high decorative standards throughout the 16th century, Italian armour of the late 16th century tended to display rather slipshod ornamentation.

The decoration of swords to be worn ceremonially at court and firearms for duelling or sport became more important in the 17th and 18th centuries than the ornamentation of armour, which was no longer worn. The centres of production were European capitals—Paris, London, Vienna, Prague, Dresden, Munich, Venice, and Naples. Cut steelwork was especially popular. In the 16th and 17th centuries, regional styles of decoration developed. Silver hilted swords were made in London and Paris, and carved ivory pistol stocks were made by Dutch craftsmen in Maastricht. Arms from the Russian town Tula and the Caucasus Region were decorated with niello (*i.e.*, the filling of incised designs with a black alloy of sulfur and silver, copper, or lead).

During the 18th century, the French were the arbiters of fashion including that of arms design and decoration. After the French Revolution, France continued to be

the leader of fine arms production. A special workshop was installed at Versailles near Paris, under Nicolas Noël Boutet (1761–1833), that provided Napoleon with presentation pieces of superb quality. This workshop was the last great one before the decline. In the 19th century, style in arms design deteriorated, and the handcrafting of arms decoration decreased as a result of the machine manufacturing and mass production that were introduced by the Industrial Revolution.

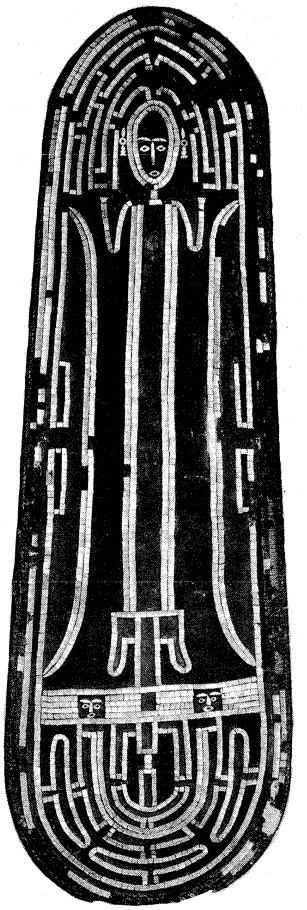
China, Japan, and India

During the Shang, or Yin, period (*c.* 1766–*c.* 1122 BC), the Chinese produced bronze blades of exquisite form and workship, sometimes richly inlaid with gold and silver wire. Most of China's later weapon types and decorative motifs were adapted from the arms and ornamentation of the nomadic peoples of the Central Asian steppes, such as the Scythians and Mongols. The traditional conservatism of the Chinese arms maker preserved many of these ancient forms and ornamental motifs until the 19th century, when more modern weaponry was introduced through Western trade.

Japanese armour was lamellar, or made of interconnected scales, of such materials as metal or lacquered leather. Leaving little surface for ornamentation, the Japanese warrior relied for aesthetic effect largely on armour accessories such as helmet crests, war masks and colourful lacings.

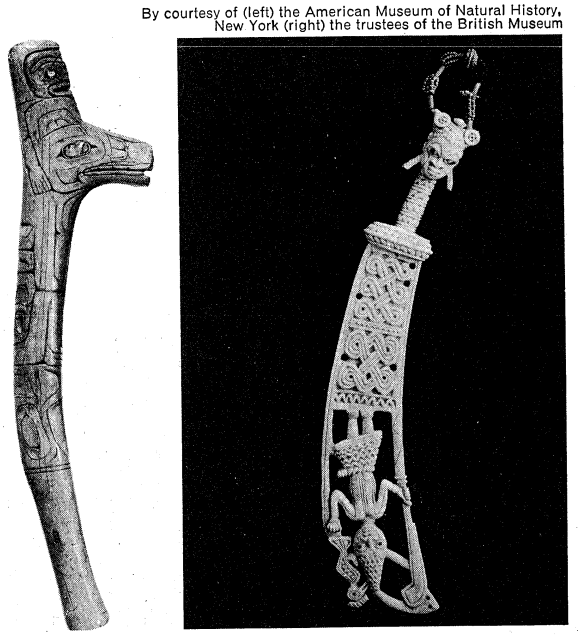
Among the Japanese, the fashioning of swords and sword furnishings, such as hilt mountings and sword guards (tsuba), was one of the main occupations of artisans working in metal. Comparable to the importance of the jewelers craft in medieval Europe were schools of sword furniture makers in Japan from the Kamakura period (1192–1333) to the 19th century. There were about 30 schools or traditional styles of design and craftsmanship. The classical style was established in the 15th century by the famed Gotō school, whose swords were considered to be the only correct ones for court use. Perhaps the most renowned individual master of tsuba was Aoki Kaneiye (flourished mid-16th century), who founded the important Kaneiye school of sword furniture manufacture.

India, apart from its Islāmic influences, has a very exuberant and colourful tradition of arms decoration. Indians have long ornamented their weapons with brilliant gems and colourful inlay or used relief carvings of floral scrolls interspersed with human figures, animals, and mythical creatures. The influence of Indian craftsmanship is reflected in many of the distinctive local styles of Indonesia.



Shield of basketry, vegetable gum, and paint, inlaid with mother-of-pearl; Solomon Islands, Melanesia. In the Museum of Primitive Art, New York.
By courtesy of the Museum of Primitive Art, New York

By courtesy of (left) the American Museum of Natural History, New York (right) the trustees of the British Museum



American Indian and African arms decoration.
(Left) War club made of caribou horn carved with a wolf's head, Chimmesyan Indian, Northwest Coast. In the American Museum of Natural History, New York. (Right) Ivory scabbard of a ceremonial sabre of the Owo-Yoruba people of Western Nigeria. In the British Museum.

Africa, Oceania, and pre-Columbian Americas

The decoration of arms in Africa varies widely according to the region and culture. The Muslim peoples of North Africa design and decorate their arms according to the general traditions followed throughout the Islāmic world. There are local variants, however, notably the renowned silver mounted Berber and Kabyle guns of Morocco. Sword hilts with gold leaf overlay are found in West Africa, but in most of the Negro cultures weapon decoration consists largely of wood and ivory carving. Chieftains' axes from the Congo area are particularly elaborate in their carved decoration.

On the islands of the South Pacific, the choice of materials for making arms and armour was severely limited. Metals, and sometimes even suitable stone, were unavailable. Nearly all of the objects for combat and hunting made by these Oceanic peoples, therefore, had to be produced from such perishable materials as wood, bone, and fibres. For this reason only recent historic examples have been preserved. Because of the geographic isolation of the islands, each group or tribe had its own tradition of design and decoration. In general, however, most of the ornamentation carved on spearheads, war clubs, and shields consists of highly stylized human and animal forms. The most distinctive styles of arms decoration are those of the Marquesas and Tonga Islands, the Solomon and Trobriand Islands, the Asmat and the Sepik region of New Guinea, and Maori work from New Zealand.

From pre-Columbian North America, few examples of decorated weapons have been preserved. Among them are the stones finely modelled in bird shapes that were probably intended as weights for spear throwers. Carved warclubs (tomahawks) and buffalo shields, which the

North American Plains Indians painted with magic protective motifs, have survived only from recent history. A few examples of ancient Meso-American decorated arms have been preserved: carved spearthrowers called *atlatl* and shields with feather mosaics. The splendour of the arms and armour of pre-Columbian Mexico and Central America is primarily known through representations in sculpture, painted murals, and the illustrated codices that have survived from Aztec and early colonial days. The arms excavated in Peru of the ancient Mochica and Inca civilizations are basically similar in form, varying only in decorative motifs.

COLLECTING AND STUDY

The first great armour collection was brought together by the Austrian archduke Ferdinand at Schloss Ambras, Tirol, in the 16th century. It was conceived as a "Gallery of Heroes," representing all great generals of the time by a suit of their armour. The Waffensammlung in Vienna —the largest and most important collection of arms and armour in the world—is largely based upon Archduke Ferdinand's collection.

Some other large collections that have been based on extant princely armories include the Real Armería in Madrid, the Historische Museum (Rüstkammer) in Dresden, the Armouries in the Tower of London, and the Livrustkammaren in Stockholm. Important collections of arms and armour assembled from artistic points of view are in the Musée de l'Armée in Paris, the Armeria Reale in Turin, the Museo Stibbert in Florence, the Wallace Collection in London, and the Metropolitan Museum of Art in New York.

During the 19th century many collections of arms and armour were assembled, and many scholarly handbooks were published, though most of them treated the subject as history rather than art. Artistic interest in the collection of armour, which dates largely from the 20th century, was influenced by the growing appreciation of Japanese arms.

BIBLIOGRAPHY. Books about arms and armour in general, but with special interest in arms decoration include: SIR GUY FRANCIS LAKING, *A Record of European Armour and Arms through Seven Centuries*, 5 vol. (1920–22); CLAUDE BLAIR, *European Armour circa 1066 to circa 1700* (1958), *European and American Arms c. 1100–1850* (1962); YIGAEL YADIN, *The Art of Warfare in Biblical Lands*, 2 vol. (Eng. trans. 1963); V.A.B. NORMAN, *Arms and Armour* (1964); R.E. OAKESHOTT, *The Sword in the Age of Chivalry* (1964); A.M. SNODGRASS, *Early Greek Armour and Weapons* (1964); H.L. BLACKMORE, *Arms and Armour* (1965); MERRILL LINDSAY, *One Hundred Great Guns* (1967); H.R. ROBINSON, *Japanese Arms and Armour* (1969), *Oriental Armour* (1967); WENDELIN BOEHEIM, *Handbuch der Waffenkunde* (1890), *Meister der Waffenschmiedekunst* (1897); ERICH HAENEL, *Kostbare Waffen aus der Dresdner Rüstkammer* (1923); WALDEMAR GINTERS, *Das Schwert der Skythen und Sarmaten in Südrussland* (1928); BRUNO THOMAS, ORTWIN GAMBER, and HANS SCHEDELMANN, *Die schönsten Waffen und Rüstungen aus europäischen und amerikanischen Sammlungen* (1963); ALEXANDER VON REITZENSTEIN, *Der Waffenschmied* (1964); GIANNI VIANELLO, *Armi in Oriente* (1966); MARTIN DE RIQUER, *L'Arnés del Cavaller* (1968); and L.G. BOCCIA and E.T. COELHO, *L'arte dell'armatura in Italia* (1969).

Books specifically treating arms decoration are: G.C. STONE, *A Glossary of the Construction, Decoration, and Use of Arms and Armor* (1934); J.F. HAYWARD, *The Art of the Gunmaker*, 2 vol. (1962–64); S.V. GRANCSAY and MERRILL LINDSAY, *Master French Gunsmith Designs of the XVII–XIX Centuries* (1970); B.W. ROBINSON, *The Arts of the Japanese Sword* (1961); H. STOECKLEIN, *Meister des Eisenschnittes* (1922); RUDOLF CEDERSTROM and K.E. STENEBERG, *Skoklosterskölden* (1945); KONRAD ULLMANN, *Schmuck alter Büchsen und Gewehre* (1964); and LUMIR JISL, *Japanische Schwertzierate* (1967).

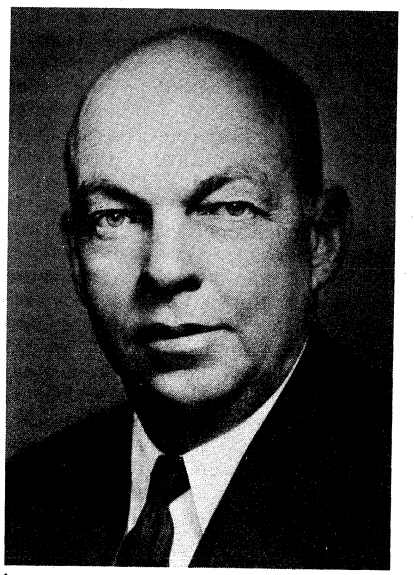
Among the vast literature about arms and armour a considerable part is dedicated to their decoration. Many articles have been published in periodicals such as: *Archaeologia* (1770–), the *Jahrbuch der Kunsthistorischen Sammlungen (Wien)* (1883–), *Zeitschrift für Historische Waffenkunde* (1897–), *Metropolitan Museum of Art Bulletin* (1905–), *Armi Antiche* (1954–), and the *Journal of the Arms and Armour Society* (1955–).

(H.Ni.)

Armstrong, Edwin H.

Edwin Howard Armstrong laid much of the foundation of modern radio and electronics in a series of brilliant and basic circuit designs. While still in college, he invented the regenerative circuit, which was at one and the same time the first amplifying receiver and the first reliable, continuous-wave transmitter. His most widely known circuit, invented in 1918, was the so-called superheterodyne circuit, a highly selective means of receiving, converting, and greatly amplifying very weak, high-frequency electromagnetic waves, which today underlies 98 percent of all radio, radar, and television reception. His crowning achievement was the invention in 1933 of wideband frequency modulation, which later became known as FM radio, a radical new system of nearly static-free broadcasting that transmits the full, natural frequency range of audible sound.

Fabian Bachrach



Armstrong, 1933.

Armstrong's career revolved around New York City. He was born on December 18, 1890, into a genteel, devoutly Presbyterian family in the old Chelsea district of Manhattan. His father, born in the same district, was a publisher, his mother a former schoolteacher, from a neighbouring family. Armstrong was a shy boy interested from childhood in engines, railway trains, and all mechanical contraptions.

At the age of 14, fired by reading of the exploits of Guglielmo Marconi in sending the first wireless message across the Atlantic, Armstrong decided to become an inventor. He built a maze of wireless apparatus in his family's attic, by then removed to the suburbs, and began the solitary, secretive work that absorbed his life. Except for a passion for tennis, acquired from his father, and later, for fast motor cars, he developed no other interest. Wireless was then in the stage of crude spark-gap transmitters and iron-filing receivers, producing faint Morse-code signals, barely audible through tight earphones. Armstrong joined in the hunt for improved instruments. On graduating from high school, he commuted to Columbia University's School of Engineering on a red motorcycle, a graduation gift from his father, to pursue his search.

Early life and schooling

In his junior year at Columbia, Armstrong made his first, most seminal invention. Among the devices investigated for better wireless reception was the then little understood, largely unused Audion, or three-element vacuum tube, invented in 1906 by Lee De Forest, a pioneer in the development of wireless telegraphy and television. Armstrong made exhaustive measurements to find out how the tube worked and devised a circuit, called the regenerative or feedback circuit, that suddenly, in the autumn of 1912, brought in signals with a thousandfold

amplification, loud enough to be heard across a room. At its highest amplification, he also discovered, the tube's circuit shifted from being a receiver to being an oscillator, or primary generator, of wireless waves. As radio-wave generator this circuit is still at the heart of all radio–television broadcasting.

Armstrong's priority was later challenged by De Forest in a monumental series of corporate patent suits, extending over 14 years, argued twice before the Supreme Court, and finally ending—in a judicial misunderstanding of the nature of the invention—in favour of de Forest. But the scientific community never accepted this verdict. The Institute of Radio Engineers refused to revoke an earlier gold-medal award to Armstrong for the discovery of the feedback circuit. Later he received the Franklin Medal, highest of U.S. scientific honours, reaffirming his invention of the regenerative circuit.

This youthful invention that opened the age of electronics had profound effects on Armstrong's life. It led him, after a stint as an instructor at Columbia, into the U.S. Army Signal Corps laboratories in World War I in Paris, where he invented the superheterodyne, a circuit going far beyond the regenerative in amplification. It brought him into early association with the man destined to lead the postwar Radio Corporation of America (RCA), David Sarnoff, whose young secretary Armstrong later married. Armstrong himself returned after the war to Columbia to become assistant to Michael Pupin, the notable physicist and inventor and his revered teacher. In this period he sold patent rights on his circuits to the major corporations, including RCA, for large sums in cash and stock. Suddenly, in the radio boom of the 1920s, he found himself a millionaire. But he continued to teach at Columbia, financing his own research, working along with Pupin, whose professorship he inherited, on the long-unsolved problem of eliminating static from radio.

Invention of FM broadcasting

In 1933 Armstrong secured four patents on advanced circuits that were to solve this last basic problem. They revealed an entirely new radio system, from transmitter to receiver. Instead of varying the amplitude or power of radio waves to carry voice or music, as in all radio before then, the new system varied or modulated the waves' frequency (number of waves per second) over a wide band of frequencies. This created a carrier wave that natural static—an amplitude phenomenon created by electrical storms—could not break into. As a result, FM's wide frequency range made possible the first clear, practical method of high-fidelity broadcasting.

Since the new system required a basic change in transmitters and receivers, it was not embraced with any alacrity by the established radio industry. Armstrong had to build the first full-scale FM station himself in 1939 at a cost of over $300,000 to prove its worth. He then had to develop and promote the system, sustain it through World War II (while he again turned to military research), and fight off postwar regulatory attempts to hobble FM's growth. When FM slowly established itself, Armstrong again found himself entrapped in another interminable patent suit to retain his invention. Ill and aging in 1954, with most of his wealth gone in the battle for FM, he took his own life (January 31 or February 1).

The years have brought increasing recognition of Armstrong's place in science and invention. FM is now the preferred system in radio, the required sound channel in all television, and the dominant medium in mobile radio, microwave relay, and space-satellite communications. Posthumously, Armstrong was elected to the pantheon of electrical greats by the Union Internationale des Télécommunications, to join such figures as the French physicist and mathematician André-Marie Ampère; Alexander Graham Bell, the inventor of the telephone; the English electrical pioneer Michael Faraday; and Guglielmo Marconi, the Italian developer of wireless telegraphy.

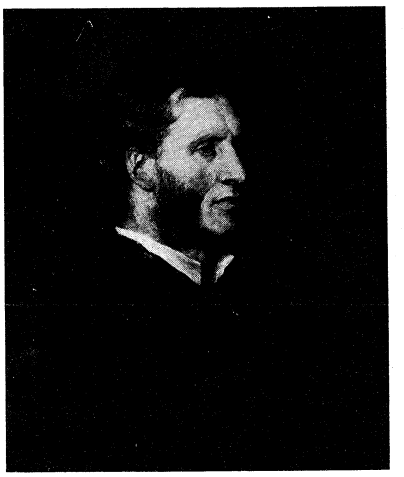
BIBLIOGRAPHY. L. LESSING, *Man of High Fidelity: Edwin Howard Armstrong* (1956; rev. paperback ed., 1969), the only definitive published biography; W.R. MACLAURIN and R.J. HARMAN, *Invention and Innovation in the Radio Industry* (1949), inventive and industrial background.

(L.P.L.)

Arnold, Matthew

An English poet who ranks among the best of his time, Matthew Arnold was also a critic whose theories have worn better than those of his Victorian contemporaries. He is above all notable for his grasp of the defects of much of the romantic poetry of his time and for opening what was then a still largely insular culture to the influence of Europe. His social, religious, and educational writings were also of great importance in their time and are of enduring interest today.

By courtesy of the National Portrait Gallery, London



Arnold, oil painting by G.F. Watts (1822–88). In the National Portrait Gallery, London.

Early life and education. Arnold was born at Laleham, Middlesex, on Dec. 24, 1822, the eldest son of the famous Dr. Thomas Arnold, who was appointed headmaster of Rugby in 1828. From his mother, Mary (*née* Penrose), he inherited Cornish blood and with it something of what he would have distinguished as the "Celtic" strain in his sensibility. After a year at Winchester, Matthew Arnold entered Rugby School (1837). His first publication was a Rugby prize poem, *Alaric at Rome* (1840). He entered Oxford as a scholar of Balliol College in 1841, won the Newdigate Prize with his poem *Cromwell* (1843), and graduated with second-class honours in 1844. At Oxford, probably in half-reaction from the earnestness of his father's Rugby, he affected a certain dandyism and ironic detachment and struck some of his contemporaries as a trifler. But he was elected to a fellowship at Oriel (1845). Arnold's levity was only a mask; though, actually, "radiant, adorn'd outside," he had even then "a hidden ground of thought and of austerity within." Far from breaking with his father, he reverenced him and spent much of his later life in advocating and furthering his educational and religious aims.

For Oxford, too, Arnold retained an impassioned affection. His Oxford was the Oxford of John Henry Newman—of Newman just about to be received into the Roman Catholic Church; and although Arnold's own religious thought, like his father's, was strongly liberal, Oxford and Newman always remained for him joint symbols of spiritual beauty and culture. Oxford, like Newmanism, might be a part of the dead or dying world; it might be "the home of lost causes and impossible loyalties," and yet: "steeped in sentiment as she lies, spreading her gardens to the moonlight, and whispering from her towers the last enchantments of the Middle Age, who will deny that Oxford, by her ineffable charm, keeps ever calling us nearer to the true goal of all of us, to the ideal, to perfection?"

Educational inspector. In 1847 Arnold became private secretary to Lord Lansdowne, who occupied a high cabinet post during Lord John Russell's Liberal ministries. And in 1851, in order to secure the income needed for his marriage (June 1851) with Frances Lucy Wightman, he accepted from Lansdowne an appointment as inspec-

Educational work

tor of schools. This was to be his routine occupation until within two years of his death, and it has often been thought an incongruous task for so rare and sensitive a spirit. The incessant travelling in all kinds of weather and the constant exposure not only to provincial hotels but to provincial rawness and philistinism certainly cost him much in strength and energy. Yet education was in his blood, and he loyally (if stoically) accepted the role of propagandist and missionary for culture and for state schools. Indeed, his fame as poet and critic should not lead us, as it has often done, to forget the extent and significance of his educational work. At a time when state action of any kind was regarded with suspicion as "interference," he saw the vital necessity in a modern democracy of organizing public education, primary and secondary, for the middle classes and for the populace.

Always critical of Victorian insularity and laissez-faire and always eager to correct British national complacency from a European standpoint, he made it his business to point out to his countrymen what had already been done in this direction on the continent. He was several times sent by the government to inquire into the state of education in France, Germany, Holland, and Switzerland, and his reports show how conscientiously he devoted some of his best energies to the work. Two of his reports on schools abroad were reprinted as books, and his annual reports on schools at home attracted wide attention, written, as they were, in Arnold's own urbane and civilized prose. Though all this routine work could have been done by lesser men, it is salutary to remember that Arnold did it, especially as a corrective to the notion of him as the "elegant Jeremiah" or the aloof and superior exponent of "culture."

Poetic achievement. The work that gives Arnold his high place in the history of literature and the history of ideas was all accomplished in the time he could spare from his official duties. Yet there is a unity in his whole output; his verse and his literary, social, and religious criticism, no less than his educational reports, proceed from a consciousness exposed to the spirit of his age and a sensibility quick to feel, and an intelligence alert to diagnose, the deeper maladies of the age. His first volume of verse was *The Strayed Reveller, and Other Poems. By A.* (1849); this was followed (in 1852) by another under the same initial: *Empedocles on Etna, and Other Poems.* In 1853 appeared the first volume of poems published under his own name; it consisted partly of poems selected from the earlier volumes and also contained the well-known preface explaining (among other things) why *Empedocles* was excluded from the selection: it was a dramatic poem "in which the suffering finds no vent in action," in which there is "everything to be endured, nothing to be done." This preface foreshadows his later criticism in its insistence upon the classic virtues of unity, impersonality, universality, and architectonic power and upon the value of the classical masterpieces as models for "an age of spiritual discomfort"—an age "wanting in moral grandeur." Other editions followed, and *Merope,* Arnold's classical tragedy, appeared in 1858, and *New Poems* in 1867. After that date, though there were further editions, Arnold wrote little more verse.

First published poems

In a letter to his mother, written in 1869, he sums up his own poetic achievement thus:

> My poems represent, on the whole, the main movement of mind of the last quarter of a century, and thus they will probably have their day as people become conscious to themselves of what that movement of mind is, and interested in the literary productions which reflect it. It might be fairly urged that I have less poetical sentiment than Tennyson, and less intellectual vigour and abundance than Browning; yet, because I have perhaps more of a fusion of the two than either of them, and have more regularly applied that fusion to the main line of modern development, I am likely enough to have my turn, as they have had theirs.

He not only "had his turn" but has continued to hold his place as, if not the most admired, yet for many readers the most congenial, of the Victorian poets and as the poet who addresses them with the voice of a human contemporary speaking directly to their condition. Not much of Arnold's verse will stand the test of his own criteria; far from being classically poised, impersonal, serene, and grand, it is often intimate, personal, full of romantic regret, sentimental pessimism, and nostalgia. As a public and social character and as a prose writer, Arnold was sunny, debonair, and sanguine; but beneath ran the current of his buried life, and of this much of his poetry is the echo:

> From the soul's subterranean depth upborne
> As from an infinitely distant land,
> Come airs, and floating echoes, and convey
> A melancholy into all our day.

"I am past thirty," he wrote a friend in 1853, "and three parts iced over." The impulse to write poetry came typically when

> A bolt is shot back somewhere in the breast,
> And a lost pulse of feeling stirs again.

Though he was "never quite benumb'd by the world's sway," these hours of insight became more and more rare, and the stirrings of buried feeling were associated with moods of regret for lost youth, regret for the freshness of the early world, moods of self-pity, moods of longing for

> The hills where his life rose
> And the sea where it goes.

At Fox How, his father's holiday home in Westmorland, Matthew's earlier life had been steeped in Wordsworthian sentiment; he knew and revered the aged poet, who lived a stone's throw away at Rydal Mount. And Arnold may be thought of as a Victorian Wordsworth—a Wordsworth exiled from the hills, dwelling on lower reaches of the river of time, further from the sources of life and inspiration, and moved to yearning rather than to rapture by Nature's grandeur, calm, and self-dependence. Yet, though much of Arnold's most characteristic verse is in this vein of soliloquy or intimate confession, he can sometimes rise, as in "Sohrab and Rustum," to epic severity and impersonality; to lofty meditation, as in "Dover Beach"; and to sustained magnificence and richness, as in "The Scholar Gipsy" and "Thyrsis"—where he wields an intricate stanza form without a stumble.

In 1857, assisted by the vote of his godfather (and predecessor) John Keble, Arnold was elected to the Oxford chair of poetry, which he held for ten years. It was characteristic of him that he revolutionized this professorship, lecturing in English instead of Latin and concerning himself not with "the trade in classic niceties" but (as the American critic Lionel Trilling put it) "with the demand which humanity makes for a poetry *adequate* to the time in which it is written." The keynote was struck in the title and contents of his inaugural lecture: "On the Modern Element in Literature," "modern" being taken to mean not merely "contemporary" (for Greece was "modern"), but the spirit that, contemplating the vast and complex spectacle of life, craves for moral and intellectual "deliverance." Several of the lectures were afterward published as critical essays, but the most substantial fruits of his professorship were the three lectures *On Translating Homer* (1861)—in which he recommends Homer's plainness and nobility as medicine for the modern world, with its "sick hurry and divided aims" and condemns Francis Newman's recent translation as ignoble and eccentric—and the lectures *On the Study of Celtic Literature* (1867), in which, without much knowledge of his subject or of anthropology, he uses the Celtic strain as a symbol of that which rejects the despotism of the commonplace and the utilitarian.

A new kind of criticism

Arnold as critic. It is said that when the poet in Arnold died, the critic was born; and it is true that from this time onward he turned almost entirely to prose. But in fact the critic had been alive in him from the start, and his poetry had itself been the outcome of his effort to apply intelligence, as well as feeling, to the problem of "the world's multitudinousness." It is a mark of Arnold's distinction and many-sidedness that, instead of writing more and worse poetry, he went on to produce those essays

that place him first among Victorian critics and high among those few European critics who have permanently influenced the course of opinion about the nature and function of poetry. In Arnold appears what is virtually a new phenomenon: the "literary" intelligence playing freely upon the great concerns of human life. He saw and proclaimed the importance, for the modern world, of those qualities of mind and spirit that literary culture can give; and literary criticism has gained immensely from his expansion of its scope.

Some of the leading ideas and phrases put into currency in *Essays in Criticism* (First Series, 1865; Second Series, 1888) and *Culture and Anarchy* (1869) will be considered briefly. The first essay in the 1865 volume, "The Function of Criticism at the Present Time," is an overture announcing briefly most of the themes he developed more fully in later work. It is at once evident that he ascribes to "criticism" a scope and importance hitherto undreamed of. The function of criticism, in his sense, is "a disinterested endeavour to learn and propagate the best that is known and thought in the world, and thus to establish a current of fresh and true ideas." It is in fact a spirit that he is trying to foster, the spirit of an awakened and informed intelligence playing upon not "literature" merely but theology, history, art, science, sociology, and politics, and in every sphere seeking "to see the object as in itself it really is."

In this critical effort, thought Arnold, England lagged woefully behind France and Germany, and the English accordingly remained in a backwater of provinciality and complacency. Even the great Romantic poets, with all their creative energy, suffered from the want of it. "Life and the world being in modern times very complex things, the creation of a modern poet, to be worth much, implies a great critical effort behind it." The English literary critic must know literatures other than his own and be in touch with European standards. This last line of thought Arnold develops in the second essay. "The Literary Influence of Academies," in which he dwells upon "the note of provinciality" in English literature, caused by remoteness from a "centre" of correct knowledge and correct taste. To realize how much Arnold widened the horizons of criticism requires only a glance at the titles (but better to read the content) of some of the other essays in *Essays in Criticism* (1865): "Maurice de Guérin," "Eugénie de Guérin," "Heinrich Heine," "Joubert," "Spinoza," "Marcus Aurelius"; in all these, as increasingly in his later books, he is "applying modern ideas to life" as well as to letters and "bringing all things under the point of view of the 19th century."

Attack on the provinciality of English literature

The first essay in the 1888 volume, "The Study of Poetry," was originally published as the general introduction to T.H. Ward's anthology, *The English Poets* (1880). It contains many of the ideas for which Arnold is best remembered. In an age of crumbling creeds, poetry will have to replace religion. More and more, we will "turn to poetry to interpret life for us, to console us, to sustain us." Therefore we must know how to distinguish the best poetry from the inferior, the genuine from the counterfeit; and to do this we must steep ourselves in the work of the acknowledged masters, using as "touchstones" passages exemplifying their "high seriousness," and their superiority of diction and movement.

The remaining essays, with the exception of the last two (on Tolstoy and Amiel), all deal with English poets: Milton, Gray, Keats, Wordsworth, Byron, and Shelley. All contain memorable things, and all attempt a serious and responsible assessment of each poet's "criticism of life" and his value as food for the modern spirit. Arnold has been taken to task for some of his judgments and omissions: for his judgment that Dryden and Pope were not "genuine" poets because they composed in their wits instead of "in the soul"; for calling Gray a "minor classic" in an age of prose and spiritual bleakness; for paying too much attention to the man behind the poetry (Gray, Keats, Shelley); for making no mention of Donne; and above all for saying that poetry is "at bottom a criticism of life." On this last point it should be remembered that he added "under the conditions fixed . . . by the laws of poetic truth and poetic beauty," and that if by "criticism" is understood (as Arnold meant) "evaluation," Arnold's dictum is seen to have wider significance than has been sometimes supposed.

Culture and Anarchy (1869) is in some ways Arnold's most central work. It is an expansion of his earlier attacks, in "The Function of Criticism" and "Heinrich Heine," upon the smugness, philistinism, and mammon worship of Victorian England. Culture, as "the study of perfection," is opposed to the prevalent "anarchy" of a new democracy without standards and without sense of direction. By "turning a stream of fresh thought upon our stock notions and habits," culture seeks to make "reason and the will of God prevail."

Arnold's concept of English society

Arnold's classification of English society into Barbarians (the aristocracy: with their high spirit, serenity, and distinguished manners and their inaccessibility to ideas), Philistines (the great middle class, the stronghold of religious nonconformity, with plenty of energy and morality but insufficient "sweetness and light"), and Populace (the masses, still raw and blind) is well known. Arnold saw in the Philistines the key to the whole position; they were now the most influential section of society; their strength was the nation's strength, their crudeness its crudeness: Educate and humanize the Philistines, therefore. Arnold saw in the idea of "the State," and not in any one class of society, the true organ and repository of the nation's collective "best self." No summary can do justice to this extraordinary book; it can still be read with pure enjoyment, for it is written with an inward poise, a serene detachment, and an infusion of mental laughter, which make it a masterpiece of ridicule as well as a searching analysis of Victorian society. The same is true of its unduly neglected sequel, *Friendship's Garland* (1871).

Religious writings. Lastly Arnold turned to religion, the constant preoccupation and true centre of his whole life, and wrote *St. Paul and Protestantism* (1870), *Literature and Dogma* (1873), *God and the Bible* (1875), and *Last Essays on Church and Religion* (1877). In these undeservedly neglected books, Arnold really founded Anglican "modernism." Like all religious liberals, he came under fire from two sides: from the orthodox, who accused him of infidelity, of turning God into a "stream of tendency" and of substituting vague emotion for definite belief; and from the infidels, for clinging to the church and retaining certain Christian beliefs of which he had undermined the foundations. To Arnold himself, it seemed that his religious writings were constructive and conservative. Those who accused him of destructiveness did not realize how far historical and scientific criticism had already riddled the old foundations; and those who accused him of timidity failed to see that he regarded religion as the highest form of culture, the one indispensable without which all secular education is in vain. His attitude is best summed up in his own words (from the preface to *God and the Bible*): "At the present moment two things about the Christian religion must surely be clear to anybody with eyes in his head. One is, that men cannot do without it; the other, that they cannot do with it as it is." Convinced that much in popular religion was "touched with the finger of death" and convinced no less of the hopelessness of man without religion, he sought to find for religion a basis of "scientific fact" that even the positive modern spirit must accept. A reading of Arnold's *Note Books* will convince any reader of the depth of Arnold's spirituality and of the degree to which, in his "buried life," he disciplined himself in constant devotion and self-forgetfulness.

Arnold died suddenly, of heart failure, on April 15, 1888, at Liverpool and was buried at Laleham, with the three sons whose early loss had shadowed his life.

MAJOR WORKS

POETICAL WORKS: *Alaric at Rome* (1840; Rugby School prize poem); *Cromwell* (1843; Newdigate Prize poem); *The Strayed Reveller, and Other Poems. By A.* (1849); *Empedocles on Etna, and Other Poems. By A.* (1852); *Poems* (1853; in-

cluding "Sohrab and Rustum," "The Forsaken Merman," and "The Scholar Gipsy"); *Poems, Second Series* (1855); *Merope* (1858; classical tragedy); *New Poems* (1867; including "Thyrsis" and "Dover Beach").

PROSE WORKS: *The Popular Education of France with Notices of that of Holland and Switzerland* (1861; revised text of the 1860 report prepared by Arnold for the Education Commission); *On Translating Homer* (1861); *On Translating Homer: Last Words* (1862); *A French Eton; or, Middle Class Education and the State* (1864); *Essays in Criticism* (1865; including "The Function of Criticism at the Present Time," "The Literary Influence of Academies," "Maurice de Guérin," "Eugénie de Guérin," "Heinrich Heine," "Joubert," "Spinoza," and "Marcus Aurelius"); *On the Study of Celtic Literature* (1867); *Schools and Universities on the Continent* (1868; reprinted from Arnold's report *On Secondary Education in Foreign Countries* of 1866); *Culture and Anarchy* (1869); *St. Paul and Protestantism* (1870); *Friendship's Garland* (1871); *Literature and Dogma* (1873); *God and the Bible* (1875); *Last Essays on Church and Religion* (1877); *Mixed Essays* (1879); *Irish Essays* (1882); *Discourses in America* (1885); *Essays in Criticism. Second Series* (1888; including "The Study of Poetry" and essays on Milton, Thomas Gray, Keats, Wordsworth, Byron, Shelley, Tolstoy, and Amiel); *Reports on Elementary Schools 1852–1882*, edited by Sir Francis Sandford (1889; new edition with added material and introduction by F.S. Marvin, 1908).

BIBLIOGRAPHY. The first collected edition of Arnold's poems was published in 1869, other editions appeared in 1877 and 1881, and a library edition in 1885. A new and complete edition was published in "Oxford Standard Authors Series" (1950), ed. by C.B. TINKER and H.F. LOWRY, who also edited *The Poetry of Matthew Arnold: A Commentary* (1940, reprinted 1970). *Arnold's Letters (1848–1888)*, collected and arranged by G.W.E. RUSSELL (1895; 2nd ed., 1901); ARNOLD WHITRIDGE (ed.), *Unpublished Letters* (1923); *Letters to Arthur Hugh Clough*, with an introductory study by H.F. LOWRY (ed.) (1932, reprinted 1968). There is a complete edition (literary contents) of *The Note-Books of Matthew Arnold*, ed. by H.F. LOWRY, K. YOUNG, and W.H. DUNN (1952). An edition de luxe of Arnold's complete *Works*, 15 vol. (1903–04), includes a bibliography by T.B. SMART, who published a separate bibliography in 1892 (reprinted 1968). For a more recent bibliography, see T.G. EHRSAM and R.H. DEILY (comps.), *Bibliographies of Twelve Victorian Authors* (1936; suppl. by J.G. FUCILLA in *Modern Philology*, 37:89–96, 1939). This gives a full bibliography of Arnold's essays, contributions to periodical literature, and editions of and introductions to literary works.

It was Arnold's expressed desire that his biography should not be written; there are, however, a number of monographs and biographico-critical works, among them LIONEL TRILLING, *Matthew Arnold* (1939), the best full-length study; J. DOVER WILSON, *Leslie Stephen and Matthew Arnold As Critics of Wordsworth* (1939); SIR E.K. CHAMBERS, *Matthew Arnold: A Study* (1947); J.D. JUMP, *Matthew Arnold*, "Men and Books Series" (1955); G. ROBERT STANGE, *Matthew Arnold: The Poet As Humanist* (1967).

Essays on Arnold: HENRY SIDGWICK, *Miscellaneous Essays and Addresses* (1904, reprinted 1968); T.S. ELIOT in *The Use of Poetry and the Use of Criticism* (1932) and "Arnold and Pater" in *Selected Essays*, new ed. (1950); BASIL WILLEY in *Nineteenth Century Studies* (1949); MERLE M. BEVINGTON in *M. Arnold's "England and the Italian Question"* (1953); JOHN HOLLOWAY in *The Victorian Sage* (1953 and 1965).

(B.W.)

Arrhenius, Svante August

A Swedish chemist, Svante Arrhenius was a founder of the modern science of physical chemistry, his best known contribution being his theory of electrolysis. Arrhenius was the first to suggest that electrolytes, certain substances that dissolve in water to yield a solution that conducts electricity, are separated, or dissociated, into electrically charged particles, or ions, even when there is no current flowing through the solution. He estimated the degree of dissociation of various electrolytes at various dilutions.

Early training

He was born at Vik, Sweden, on February 19, 1859. His father was a surveyor and estate manager. Svante, the second son, is said to have taught himself to read at the age of three and to have become interested in mathematics from watching his father add columns of figures. He attended the Cathedral School at Uppsala and went on to the university, where he studied physics, mathematics, and chemistry. In pursuit of his doctorate he migrated to Stockholm to work on electrolysis under Erik Edlund. In 1883 he published his first paper and in May 1884, at Uppsala, defended his doctoral thesis containing in embryo the dissociation theory.



Arrhenius, 1918.
By courtesy of the Kungl. Biblioteket, Stockholm

The thesis was greeted with incredulity and awarded the fourth class, a bare pass; the university in effect condemned an important and original thesis. The faculty at Uppsala were skeptical of hypotheses and devoted to accurate experimental work, while Arrhenius boasted (not quite truly) that he had never performed an exact experiment in his life; moreover, his subject fell awkwardly between chemistry and physics. Even to the sympathetic English physicist Sir Oliver Lodge, who in 1886 described the theory to the British Association for the Advancement of Science, Arrhenius seemed sometimes "to indulge in . . . manipulation of imaginary data," producing "a confusion" from which emerged so-called theoretical deductions. In reality, Arrhenius had a statistical sense and an ability to frame formulas to fit his facts, both of which were rare among chemists of his day. He had prudently sent copies of his thesis to the most prominent physical chemists of the day, who were able to understand it; and in August 1884 the German physical chemist Wilhelm Ostwald went from Riga to Uppsala to offer Arrhenius a post. He was at once given a lectureship in physical chemistry at Uppsala; and in 1886 Edlund got him a travelling fellowship from the Swedish Academy of Sciences.

Gradual acceptance of his theory

Arrhenius spent from 1886 to 1890 working with other eminent scientists—Ostwald at Riga, F.W. Kohlrausch at Würzburg, Ludwig Boltzmann at Graz, and Jacobus van't Hoff at Amsterdam. During these years he refined his theory, which gradually began to win adherents. In 1891 he was offered a chair at Giessen, Germany, where Justus von Liebig, half a century earlier, had revolutionized the teaching of chemistry; but he wished to remain in Sweden and obtained a post at the Royal Institute of Technology in Stockholm. In 1895 he became professor of physics and, in 1896, rector of the school. Abroad, his reputation stood very high; but he was not elected to the Swedish Academy of Sciences until 1901, and even then with strong opposition. In 1902 he received the Davy Medal of the Royal Society of London, which in 1911 elected him a foreign member; and in 1903 his own countrymen made amends when he became the first Swede to be awarded a Nobel Prize.

In 1905 he was offered a chair at Berlin, then the most eminent position open to an academic chemist. On patriotic grounds, he refused; and the directorship of the Nobel Institute for Physical Chemistry at Stockholm was created for him. This gave him ample opportunity for research and writing, and his later years were contented.

In 1911 he visited the United States to receive the first

Willard Gibbs Medal and to deliver the Silliman Lectures at Yale University, published as *Theories of Solutions* (1912). He died in Stockholm on October 2, 1927.

Arrhenius was a genial, energetic man who made many friends on his visits abroad. His memory was excellent, he loved nature, but he was indifferent to the fine arts and literature. His range of scientific interests was very wide: over the years, he moved away from the study of solutions into immunology, where he made pioneering studies on toxins, and then into geology and cosmology. In *Worlds in the Making* (1908), he suggested that cool stars might collide and form nebulae from which new stars and planets would arise; and so the process would go on indefinitely, life being spread about the universe by bacteria propelled by light pressure. These speculations have not found their way into modern cosmology.

BIBLIOGRAPHY. The only full-length biography of Arrhenius is E.H. RIESENFELD, *Svante Arrhenius* (1931), in German. Brief English-language biographies may be found in the *Dictionary of Scientific Biography*, vol. 1, pp. 296–302 (1970), with a full bibliography; E. FARBER (ed.), *Great Chemists* (1961); and J.R. PARTINGTON, *A History of Chemistry*, vol. 4 (1964).

(D.M.K.)

Art, Philosophy of

There are many disciplines called "philosophy of . . .": philosophy of science, philosophy of religion, philosophy of history, philosophy of education, philosophy of art. In each of these cases, what does the phrase "philosophy of" signify? Those who are concerned with "the philosophy of" a certain discipline (1) attempt to give careful analyses of the principal concepts employed in that discipline, and, by means of such analyses, (2) attempt to determine the truth or falsity of very general statements used in that discipline or about it. Thus, in the philosophy of religion, clarity is sought about the use of such terms as supernatural, God, miracle, cause, create, and in the light of these to discover whether statements such as "God exists" and "God created the world" are true or false. Similarly, in the philosophy of art, attempts are made to clarify concepts that are constantly used in discourse about the arts—for example, "form," "expression," "meaning," "symbol," "representation," "abstraction," "beautiful"—and in the light of these clarifications to examine such statements as "Art is expression" and "Every work of art is a symbol." Conclusions arrived at by philosophers of art are very general in that, while they can be applied to particular works of art, they are not primarily about particular works of art: whether a certain painting is beautiful is a matter left for critics of painting to decide; the philosopher of art is content if he has answered to his satisfaction the question, "What is it for a work of art to be beautiful?"

This article is divided into the following sections:

I. Conceptual foundations of the philosophy of art
  Distinguishing characteristics
  The interpretation of art
  The mediums of art
    Classifying arts by their mediums
    Differences in the arts related to mediums
II. Theories of the nature, functions, and effects of art
  Art as imitation (representation)
    Analysis of representation
    Subject matter
    Symbols in art
    Meaning
  Art as expression
    Expression in the creation of art
    The expressive product
  Art as form
    The formalist position
    Formal principles in art
  Pragmatic theories of art
    Hedonistic theories of art
    Art as a means to truth or knowledge
    Art as a means to moral improvement

I. Conceptual foundations of the philosophy of art

DISTINGUISHING CHARACTERISTICS

The philosophy of art is distinguished from art criticism, which is concerned with the analysis and evaluation of particular works of art. Critical activity may be primarily historical, as when a lecture is given on the conventions of the Elizabethan theatre in order to explain some of the devices used in Shakespeare's plays. It may be primarily analytical, as when a certain passage of poetry is separated into its elements and its meaning or import explained in relation to other passages and other poems in the tradition. Or it may be primarily evaluative, as when reasons are given for saying that the work of art in question is good or bad, or better or worse than another one. Sometimes it is not a single work of art but an entire class of works in a certain style or genre (such as pastoral poems or Baroque music) that is being elucidated; and sometimes it is the art of an entire period (such as Romantic). But in every case, the aim of art criticism is to achieve an increased understanding or enjoyment of the work (or classes of works) of art, and its statements are designed to achieve this end.

Difference between the philosophy of art and art criticism

The test of the success of art criticism with a given person is: has this essay or book of art criticism increased his understanding or enhanced his appreciation of the work of art in question? Art criticism is particularly helpful and often necessary for works of art that are more than usually difficult, so that the average person would be unable adequately to understand or enjoy them if left to himself.

The task of the philosopher of art is more fundamental than that of the art critic in that the critic's pronouncements presuppose answers to the questions set by the philosopher of art. The critic says that a given work of music is expressive; but the philosopher of art asks what is meant by saying that a work of art is expressive and how one determines whether it is. In speaking and writing about art, the critic presupposes that he is dealing with clear concepts, the attainment of which is the task of the philosopher of art.

The task of the philosopher of art is not to heighten understanding and appreciation of works of art but to provide conceptual foundations for the critic by (1) examining the basic concepts underlying the critic's activities to enable him to speak and write more intelligibly about the arts, and by (2) arriving at true conclusions about art, aesthetic value, expression, and the other concepts that the critic employs.

The meaning of "art"

Upon what does the philosopher of art direct his attention? "Art," is the ready answer; but what is art and what distinguishes it from all other things? The theorists who have attempted to answer this question are many, and their answers differ greatly. But there is one feature that virtually all of them have in common: a work of art is a man-made thing, an artifact, as distinguished from an object in nature. A sunset may be beautiful, but it is not a work of art. A piece of driftwood may have aesthetic qualities, but it is not a work of art since it was not made by man. On the other hand, a piece of wood that has been carved to look like driftwood is not an object of nature but of art, even though the appearance of the two may be exactly the same. This distinction is being challenged in the 20th century by artists who declare that *objets trouvés* ("found objects") are works of art, since the artist's perception of them as such makes them so, even if the objects were not man-made and were not modified in any way (except by exhibition) from their natural state.

Nevertheless, according to the simplest and widest definition, art is anything that is man-made. Within the scope of this definition, not only paintings and sculptures but also buildings, furniture, automobiles, cities, and garbage dumps are all works of art: every change that human activity has wrought upon the face of nature is art, be it good or bad, beautiful or ugly, beneficial or destructive.

The ordinary usage of the term is clearly less wide. In daily life when works of art are spoken of, the intention is to denote a much narrower range of objects, namely, those responded to aesthetically. Among the things in this narrower range, a distinction, although not a precise one, is made between fine and useful art. Fine art consists of those works designed to produce an aesthetic response or that (regardless of design) function as objects

of aesthetic appreciation (such as paintings, sculptures, poems, musical compositions)—those man-made things that are enjoyed for their own sake rather than as means to something else. Useful art has both an aesthetic and a utilitarian dimension: automobiles, glass tumblers, woven baskets, desk lamps, and a host of other handmade or manufactured objects have a primarily useful function and are made for that purpose, but they also have an aesthetic dimension: they can be enjoyed as objects of beauty, so much so that a person often buys one brand of car rather than another for aesthetic reasons even more than for mechanical reasons (of which he may know nothing). A borderline case is architecture: many buildings are useful objects the aesthetic function of which is marginal, and other buildings are primarily objects of beauty the utility of which is incidental or no longer existent (Greek temples were once places of worship, but today their value is entirely aesthetic). The test in practice is not how they were intended by their creators, but how they function in present-day experience. Many great works of painting and sculpture, for example, were created to glorify a deity and not, insofar as can be ascertained, for an aesthetic purpose (to be enjoyed simply in the contemplation of them for their own sake). It should be added, however, that many artists were undoubtedly concerned to satisfy their aesthetic capabilities in the creation of their work, since they were highly perfectionistic as artists; but in their time there was no such discipline as aesthetics in which they could articulate their goals; in any case, they chose to create "for the greater glory of God" by producing works that were also worthwhile to contemplate for their own sake.

Other senses of the term "art"

This aesthetic sense of the word "art," whether applied to fine art or useful art, is the one most employed by the majority of critics and philosophers of art today. There are two other senses of "art," however, that are still narrower, and, to avoid confusion, their use should be noted: (1) Sometimes the term "art" is restricted to the visual arts alone or to some of the visual arts. Thus, the art department in 20th-century colleges and universities is usually the department of painting and sculpture (and sometimes architecture). But as philosophers of art use the term (and as it is used here), art is not limited to visual art; music and drama and poetry are as much arts as painting, sculpture, and architecture. (2) Sometimes the term "art" is used in a persuasive sense, to include only those works considered good art. "That's not art!" exclaims the viewer at an art gallery as he examines a painting he dislikes. But if the term "art" is to be used without confusion, it must be possible for there to be bad art as well as good art; the viewer, then, is not really denying that the work in question is art (it is a man-made object presented to be contemplated for its own sake) but only that it is worthwhile.

The word "art" is also ambiguous in another way: it is sometimes used to designate the activity of creating a work of art, as in the slogan "Art is expression"; but it is more often used to designate the product of that process, the completed artwork or artifact itself, as in the remark "Art is a source of great enjoyment to me." There will be occasion later to remark on this ambiguity.

Countless proffered definitions of "art" are not definitions at all but theories about the nature of art that presuppose that the ability to identify certain things in the world as works of art already exists. Most of them are highly unsatisfactory even as theories. "Art is an exploration of reality through a sensuous presentation"—but in what way is it an exploration? Is it always concerned with reality (how is music concerned with reality, for example)? "Art is a re-creation of reality"—but is all art re-creation, even music? (It would seem likely that music is the creation of something, namely, a new set of tonal relationships, but not that it is the re-creation of anything at all.) "Art is an expression of feeling though a medium"—but is it always an expression (see below *Art as expression*) and is it always feeling that is expressed? And so on. It appears more certain that Shakespeare's *King Lear* is a work of art than that these theories are true. All that seems to be required for identifying something as a work of art in the wide sense is that it is not a natural object but something made or transformed by man; and all that is required for identifying it as art (not as good art but as art) in the narrower sense is that it functions aesthetically in man's experience, either wholly (fine art) or in part (useful art); it is not even necessary, as has been shown, that it be intended by its creator to function in this way.

THE INTERPRETATION OF ART

Works of art present problems of both interpretation and evaluation. Evaluation is not the concern of this article (see AESTHETICS); but one problem about interpretation deserves to be mentioned. Works of art are often difficult, and how to interpret them properly is far from obvious. The question then arises as to what factors should guide efforts at interpretation.

Isolationism and contextualism

At one extreme lies the view known as isolationism, according to which a knowledge of the artist's biography, historical background, and other factors is irrelevant to an appreciation of the work of art and usually is harmful in that it gets in the way, tending to substitute a recital of these facts for the more difficult attempt to come to grips with the work of art itself. If the work of art is not understood on first acquaintance, it should be read (or heard, or viewed) again and yet again. Constant re-exposure to it, so that the recipient is totally absorbed in and permeated by it, is the way to maximum appreciation.

At the other extreme, contextualism holds that the work of art should always be apprehended in its context or setting and that not merely knowledge about it but total appreciation of it is much richer if it is approached with this knowledge. According to the contextualists, not only literature (ordinarily appreciated contextually) but also the other arts, even nonrepresentational painting and music, should be apprehended in this way.

No critic or art lover need hold to either position in its undiluted form: a person could well be an isolationist about some kinds of art, such as music, and a contextualist about others, such as historical dramas and religious paintings. It is essential to be more specific, however, about the factors—other than careful and repeated perusal of the work of art itself—that the contextualist holds are either necessary or extremely helpful in the appreciation of works of art:

1. Other works of art by the same artist. If the artist has created other works, particularly in the same genre, acquaintance with them may enhance appreciation of the work at hand. Quantity of works has no particular merit in itself; but, when, say, one of the piano concertos of the 18th-century Austrian composer Wolfgang Amadeus Mozart is heard, the auditor may (often largely unconsciously) compare its mode, thematic material, and method of development and resolution with some of Mozart's 25 other piano concertos. Knowledge of the entire corpus of his work in a certain genre may heighten enjoyment of a particular work.

2. Other works of art in the same genre by other artists, particularly in the same style or tradition. Appreciation of the pastoral poem "Lycidas," by the English poet John Milton, is doubtless enhanced by a study of the pastoral tradition in poetry, with which Milton supposed his readers to be acquainted. To study "Lycidas" in isolation would needlessly deprive the reader of much of the richness of texture of the poem and would even make some of the references in it unintelligible.

3. A study of relevant facts about the artistic medium, such as the instrumental limitations or advantages of pipe organs in the time of the German composer Johann Sebastian Bach (1685–1750) or the modes of presentation of ancient Greek tragedies in the Athenian theatre. An acquaintance with the artistic conventions and idioms in which the artist operated often leads to better understanding of certain aspects of his work and helps to avert misunderstandings of it.

4. A study of the age in which the artist lived—the spirit of the time and its current ideas, the complex influ-

External influences on the artist

ences that molded the artist, even the social, economic, and political conditions of the time and place in which he worked. Sometimes such knowledge is of dubious relevance: it can be argued that no aid to the study of the 82 string quartets and 104 symphonies of the 18th-century Austrian composer Joseph Haydn is provided by reading about the political and economic conditions of his day. It is interesting to study the evolution of the string quartet or symphony from its origin through Haydn to the present; but this would appear to be an evolution traceable entirely within the art form and not dependent on factors outside it. This, however, is not always so: particularly in literature, where a study of such exterior factors seems to be of much more relevance. It would seem important to know, for example, that Milton was aware of the new Copernican astronomy but deliberately chose in *Paradise Lost* to make his cosmos Ptolemaic, the antiquated astronomical system that was already steeped in literature, mythology, and tradition.

5. A study of the artist's life. Anthologists of literature constantly assume that this is an important consideration, since they supply detailed biographies prior to their selections by each author. It is true, of course, that knowledge of the artist's life can distract attention from his work, as with those who cannot hear Beethoven's late quartets without constantly thinking, "What a pity it was that he was deaf at the time!" Yet such knowledge may also heighten experience of a work; some would say, at any rate, that it helps to know that Milton was blind when he wrote the sonnet "On His Blindness." It is the relevance of this kind of knowledge to an appreciation of the poem, as a poem, that is in dispute. In every case, however, it should be kept in mind that acquaintance with the artist's biography is a means toward an end, the enhanced appreciation and understanding of the work of art, and that otherwise it is aesthetically irrelevant. The facts about the artist's life are the means and the enhanced appreciation the end, not the other way around, as is often found, for example, in psychoanalytic essays attempting to infer facts about the artist's subconscious conflicts from his work; in these cases the work is being taken as the means and the study of his life as the end.

6. A study of the artist's intentions. It is this factor that has prompted the principal controversy in the mid-20th century. When difficulties arise as to what to make of a work of art or when several conflicting interpretations come to mind, how is the difficulty to be resolved? One obvious suggestion is to consult the artist or his records or memoirs or the testimony of people who knew him, to discover what his intentions were with regard to the work or the passage. It is tempting to believe that whichever way he intended it this is the way the work should be interpreted; for in regard to his own work, surely the author's own word should be law.

"Intentional fallacy"

This temptation is hotly decried by other critics as "the intentional fallacy"—the fallacy (if it is a fallacy) of believing that whichever way the artist intended it is by definition the way it really is. A work of art should stand on its own, without help from the artist; if he has not sufficiently realized his intentions within the work, forcing the recipient to go outside for help, this is held to be an artistic defect. Once the artist has completed his work, moreover, and presented it to the world, it belongs to the world and no longer exclusively to him: in the interpretation of it he now becomes just one critic among many, whose word should be respected but not taken as the final authority. Perhaps other critics can think of better interpretations than he did, which give a greater aesthetic reward in subsequent encounters with the work; perhaps there are even acceptable interpretations (such as the Freudian interpretations of Shakespeare's *Hamlet*) that he could not possibly have thought of himself at the time.

In the late 18th and early 19th centuries, Johann Wolfgang von Goethe set forth three criteria for critics to consider in interpreting and evaluating a work of art: (1) What was the artist trying to do? (2) Did he do it? (3) Was it worth doing? The first of the three is intentionalistic; and, says the intentionalist, surely this is plausible: an artist can hardly be blamed for failing to do what he had no intention of doing. It must first be known, then, what he was trying to do.

But the anti-intentionalist points out that the intention makes no difference, only the product does. If the ballerina excuses her fall in the middle of the dance by saying that she intended it, the dance is just as marred aesthetically as if she had fallen accidentally. And if a poet admits that he wrote rubbish and says that this is just what he intended to write, one does not rate the poem any higher because the poet's intention was fulfilled.

The persistent questioner might ask, however, if there are not at least some works of art in which the intentions of the artist have to be known? Suppose that a contemporary critic reads a dull, stodgy, moralistic Victorian novel and says at the end, "What an excellent parody of a Victorian novel!" But it was not a parody; its intentions were deadly serious—and should not this be known in order to interpret and evaluate it properly? Not at all, replies the anti-intentionalist; all the critic has to say is, "As a Victorian novel, this is deadly dull; as a parody of a Victorian novel, it is brilliant; if the author intended it in the former way, so much the worse for him—his work can still be praised for being brilliant parody, even if it wasn't intended as one. He just achieved something better than he knew at the time."

The author's statement of intention

Still, the intentionalist has a point: sometimes the clue to unlock an otherwise intransigent work may come from the author's statement of intention, and a plausible interpretation might be unobtainable without it. Such a suggestion might have come from a reader other than the author; but there is no point in disdaining helpful hints, regardless of their source. If the suggestion does come from the artist himself, that is nothing against it. Perhaps a work is less aesthetically perfect because it requires outside clues to its interpretation, but few works of art even approach perfection, and they may yet amply repay attention, all the more if some plausible suggestion comes from the outside. The 20th-century Russian composer Sergey Prokofiev intended his *Classical Symphony* to be a parody of the classical symphonies of Mozart and Haydn—and regardless of whether the suggestion that it be construed this way came from Prokofiev (as it did) or from someone else, if it is rewarding to listen to it in this way, then no one gains by refusing to accept the suggestion. A statement of intention is not the only key to unlocking the secrets of works of art, but it is one key among many, and there appears to be no good reason why it should not be used.

THE MEDIUMS OF ART

In the context of every work of art there are three items to consider:

1. The genesis of the work of art.
2. The artifact, or work of art, which is a publicly available object or thing made by the artist and viewed by the audience.
3. The effects of the work of art upon the audience.

The first item comprises all the artist's mental states, both conscious and unconscious, in the creation of the work, including his intention with regard to the work, as well as all the factors that led to these states of mind: for example, the spirit of the age, the socio-economic conditions of the times, his exchange of ideas with other artists, and so forth. Whatever factors helped to form the work of art in the artist's mind fall under this heading. The experiences undergone by the artist in the creation of the work constitute the artistic experience.

The third item includes all the effects of the work of art upon those who experience it, including both aesthetic and nonaesthetic reactions, the influence of the work of art upon the culture, on the state of knowledge, on current morality, and the like. The experience that involves the observer's attention to the work of art for its own sake and not for the sake of some ulterior end is called aesthetic, but of course art has many effects that are not aesthetic. The aesthetic experience belongs to the consumer of art, as opposed to the artistic experience, which belongs to the creator of art.

The second item is what is usually called the work of art

itself. According to some writers, such as the Italian philosopher Benedetto Croce (1866–1952), the work of art exists only in the mind of the artist, and the physical artifact then counts as an effect of the work of art. But in ordinary usage, as well as the usage of most philosophers of art, the work of art is identified with the physical artifact, as it exists in the physical medium. What goes on in the creator's mind is already contained in the first item.

Every work of art occurs in a medium; that is, there is some physical object or series of events by which the work is communicated to the recipient (listener, observer, reader) by means of his senses. In painting, the medium is paint; in sculpture, such materials as stone or wood or plastic. It might at first be thought that the medium of music consists of the musical score on which the composer writes the notes; but the written notes are not music, they are a set of visual cues for the production of the tones to be emitted by the various instruments. If every player had a perfect memory, he would have no need for the written score; indeed, music existed long before there were any written scores and was played or sung from memory from one year or generation to the next. It could be said more plausibly that the medium of music consists of the physical sound waves by means of which the sound sensations enter the consciousness of the listener. The medium of literature can truly be said to be words, yet not words as abstract entities conceived in the mind but words as spoken (in oral presentation) or written. The physical medium of literature, then, is either auditory or visual, although what is conveyed through the medium is not.

The medium of music

Classifying arts by their mediums. There are many ways of classifying the arts—by their purpose, by their intentions, by their effects. But the most usual and the most fundamental method of classifying the arts is by their mediums:

Visual art. This includes two-dimensional visual arts such as drawing and painting and also three-dimensional visual arts such as sculpture and architecture. Some of these should doubtless be called visuo-tactual art: buildings are ordinarily touched as well as seen, sculptures could be more fully appreciated if touched as well as seen, and even paintings may sometimes have enough three-dimensionality to repay touch experience. At any rate, all these arts appeal first and foremost, though not exclusively, to the sense of sight; and the artifact is an object in the visual medium.

Auditory art. This includes music in all its forms but not song, opera, and those arts that combine music with literature (see below). Just as the medium of visual art is sight, so the medium of auditory art is sound.

In auditory art there is—unlike visual art—no physical object (other than the score, which as has been seen is not the music). There is only the temporally successive series of sounds: sound waves emanating from the various instruments. While no such tones are being emitted, no sounds exist; only the musical score exists (and the memories of listeners, some of whom might enable the score to be reproduced if it were lost), from which music can be reproduced. Unlike the existence of paintings and sculptures, the existence of musical sounds is intermittent. In what sense, then, does the music exist between performances? It exists only in the sense that it is reproducible from the written score.

Verbal art. The art of literature is clearly different from both visual and auditory art. There are sound values in poetry, particularly when read aloud; but as sound alone, literature would be the most poverty-stricken of arts: what makes the sounds of poetry effective is (at least 99 percent) knowledge of the meanings of the words heard. Listening to the sounds of a poem or play uttered in an unfamiliar language gives some idea of the importance in literature of knowing the meanings of the words. Note that "murmuring," one of the most pleasant sounding words in English, has almost the same sounds as "murdering." It is almost exclusively a knowledge of word meanings that makes it possible to appreciate the art of literature.

The meanings of words

Nor is literature a visual art, although it is customary to read works of literature from a printed page. A critic who said, "I think this poem is a bad one, because it is written in unpleasant small type in double-column pages on yellowed paper," might be giving advice to typesetters and book designers (these two groups are engaged in the practice of visual arts), but he would be saying nothing about the merits of the poem. The printed or written word or for that matter the spoken word is only a vehicle for the meanings. Literature, then, must be placed in a separate class from either auditory or visual arts.

Mixed arts. Other arts variously combine the above three types of arts; this group includes all the arts of performance. Drama combines the art of literature (verbal art) with the visual arts of costuming, stage designing, and so on. Opera combines the art of music (its predominant component) with the art of literature (the libretto) and the visual arts of stage design. Dance combines the visual spectacle of moving bodies (the principal component) with musical accompaniment, sometimes with accompanying words and often with stage design. Song combines words with music. The motion picture combines the visual component (a series of pictures presented in such rapid succession that they appear to be moving) with the verbal component (the script) and usually an intermittent musical background as well.

Spatial arts and temporal arts

All the visual arts are also spatial arts, or arts of space; music and literature are both temporal arts, or arts of time. This leads to very great differences in the things each can do. In temporal arts, the parts do not appear together before the audience but appear successively in time, the second moment not beginning until the first one has finished. In spatial arts, the entire work of art is present simultaneously; attention to the parts of it is successive—it is impossible to concentrate on the whole at once, at least on first viewing—but the entire object is nevertheless there, and it is up to the viewer which part he shall examine first. In three-dimensional art, such as sculpture and architecture, the entire object is present, but it is impossible even to see (much less to look at) all of it at once: the back of a statue cannot be seen at the same moment as the front and the exterior of a cathedral cannot be viewed by someone inside it.

Temporal arts must be attended to in a certain order: it is impossible to hear the symphony played backward, or the drama, or the movie; even when technically it can be done (as in running a motion picture in reverse), the results usually are an aesthetic catastrophe. The recipient is supposed to attend to the temporal work's various parts in an order predetermined by the artist. For this reason, painting is not capable of telling a story in the way that a novel is; for a story is a series of temporally successive happenings, and a painting can at best take a series of represented persons and objects and show them as they exist at one moment only, one knife-edge of time; whereas a novel can depict the temporally successive happenings in the order of their occurrence (or in a different order, such as flashback).

The German aesthetician and dramatist Gotthold Lessing made this distinction the basis of his study *Laokoon*, contending that the function of visual art is to create beautiful objects and that the artist should select that stance or moment at which the person or object appears most beautiful, to enable the viewer to continue looking at it with pleasure; whereas literature, being temporal, is equipped to tell a story that includes many moments other than pleasing ones (moreover, the scenes in literature are not seen with the eyes but only imagined). Lessing's thesis that each art should restrict itself to what it can do best or is peculiarly equipped by its medium to do is a highly controversial one: it would virtually eliminate program music, for example, and descriptions of nature in novels. The tendency of art today is to attempt to curtail distinctions between time and space rather than to preserve them.

Differences in the arts related to mediums. Very significant differences among the arts occur because of the differences in their mediums:

Literary and nonliterary. The greatest difference among the arts is between the literary and the nonliter-

ary. Literature consists of a system of symbols with assigned meanings. A word is not simply a noise (or a mark on a printed page); a word is a noise or a printed mark with an assigned meaning. In different languages, different noises have been assigned meanings, and the language must be learned in order to understand what is being said. To appreciate the work of the 11th-century novelist Lady Murasaki Shikibu, one must learn Japanese; to fully appreciate Molière, the 17th-century playwright, one must learn French. No other art has this problem: the Englishman can appreciate German music as well as a German—or if he does not, it is not for lack of learning a language.

Assigned meanings of words

Shapes, colours, and tones do not have assigned meanings. That is not to say that these elements when present in art cannot be said to have some sense of the term "meaning." There are many meanings of the word "meaning," and a colour, for example, can have meaning in that it may symbolize something, as red symbolizes courage; or it may have strong emotional or other effects upon the viewer, evoking all manner of strong associations. But a colour or a tone has no assigned meaning: if the question were asked, "What does middle C mean?", the answer would be, "It has no assigned meaning at all; in that sense, it means nothing—it just has certain effects." But if the meaning of a certain word in a poem is not known, the reader is to that extent prevented from appreciating the poem; for the medium of poetry is not noises, not printed marks, but words, and the difference between a noise and a word lies in the fact that a word is a noise with a meaning.

This fact makes for an enormous difference between literature and the other arts. A colour in a painting may be the colour of an object represented; a colour may even "mean" something; for example, the white of a white flag that a person depicted in a battle painting is holding up as a sign of surrender. But a colour, as a colour, has no meaning at all; and the same is true of musical tones. A pattern of musical tones may occasionally acquire a meaning (the first four tones of Beethoven's *Fifth Symphony* were used to symbolize victory in World War II), but when this happens it really has very little to do with the music, and in any case most music is appreciated without any such symbolism being present. But a noise, however pleasing to the ear, is only a noise and not a word unless it has an assigned meaning; and one must know what that meaning is in order to appreciate a poem or any other work of literature.

The translation problem. Because literature consists of conventional symbols, there exists in literature the problem of translation, which does not exist in the other arts. When one seeks to make a work of literature available to a wider audience than that composed of only the native speakers of the language in which the work was written, the process of translation must be resorted to, and, in this process, a great deal of the work's original character is lost.

Difficulties in translating poetry

In a poem there are (1) the sounds, (2) the dictionary meanings of the words, and (3) the connotations of the words—the manifold associations that they evoke (sensory, intellectual, and emotional) in the minds of readers. The sounds are the least important of the three, and many a great poem as sheer sound is hardly even pleasing. The finding of like dictionary meanings is usually a simple matter, and when there is a word that has no rough equivalent in the other language, it may be simply retained in the original language (for example, the German word *Weltanschauung*, meaning something like "world outlook," is often retained in English translations of German works). As for the associations that hover about a word, they may vary from language to language, so that if a work is translated rather literally, the associative values of the words are lost. Thus, "My God!" is a much stronger expletive in English than "mon dieu" is in French, so that if the French expression is translated into the English one, it is, though literally correct, quite unfaithful to the weaker emotive force of the French expression. Words can often be found in the second language that have a roughly equivalent associative value to the original one, but these will usually not provide a literal translation; thus, the translator is faced with the dilemma of being able to provide a literal-meaning translation or a translation that renders the spirit or "feel" of the original but not both.

The question of correspondence to actuality. The arts also differ from one another, according to their mediums, in whether the items in the medium correspond to items in the world. Objects with colours and shapes are represented on canvas, and objects with colours and shapes also exist in the outside world. Even when a painting is nonrepresentational, it consists of colours and shapes, which are items in the outside world (even though certain individual colours and shapes in the painting may not exist in the outside world). But the case with music is different: though the visual arts may (to varying degrees) convey the sights of nature, music does not convey the sounds of nature. Even when a work of music attempts to represent the sound of an iron foundry or the clattering of horses' hooves, it really does not sound like these things: musical instruments emit tones, and in nature are found largely noises, and between the two there is an enormous auditory difference. Some rhythms of nature can be duplicated by musical instruments but hardly the sounds themselves.

The medium of literature, words, is indeed man created; but of course this feature is far from unique to literature. Words were devised and employed in countless situations of daily life before they were ever embodied in literature; so in literature, as in visual art, a medium is being employed that existed before the art itself.

II. Theories of the nature, functions, and effects of art

ART AS IMITATION (REPRESENTATION)

The view that "art is imitation" is at least as old as the Greek philosopher Plato (428–347 BC), and, although not widely held today, its long and distinguished history is evidence of its continuing hold on human beings as an account of the distinctive function of art. A terminological point, however, is in order here: in the interests of clarity, an artist should be spoken of as representing in his work the persons and things and scenes of the world but as imitating the work of other artists. Thus, "In this painting he represents a barn and some wheat fields, and in his style he is imitative of Vincent van Gogh." This distinction will be employed here, with the result that these traditional theories of art will be spoken of as theories of representation rather than of imitation.

At some period in the history of art, aestheticians and critics wrote as if nature should be recorded by the artist with photographic fidelity. The invention of photography (which can do this better than any painter) could plausibly be said to have relieved the artist of any such responsibility. Still, art can represent reality: the representation of a house in a painting may not look exactly like a house—it cannot, since the real house is three-dimensional and the painting is two-dimensional—but it looks enough like one to enable everyone unhesitatingly to identify it as a house.

Distinction between depiction and portrayal

A distinction should be made between depiction and portrayal. A painting may be said to depict a house if it looks more like a house than like anything else. Thus, most persons unhesitatingly classify this as a man, that as a tree, and so on; only when the painter has distorted or abstracted so much that a thing looks somewhat like a wolf and also somewhat like a bobcat, do they hesitate in saying what the object represented is. A picture may depict a rather short man in a French general's uniform of the early 19th century; but it may, in addition, portray Napoleon. It portrays Napoleon if (1) the artist intended it to represent Napoleon (for example, if the title of the painting is "Napoleon") and (2) the painting does look like Napoleon to some degree at least—at any rate it contains no important characteristics known to be incompatible with those of Napoleon. Clearly, if it is a painting that depicts a tree in someone's yard, it cannot be considered a portrait of Napoleon, no matter how much the artist said he intended it to be one. Depiction subjects

can ordinarily be recognized at once with a little knowledge of the world and the names of the things in it. Portrayal subjects require knowledge of whomever the artist intended to portray; even when that seems obvious, as in the case of Napoleon (who would be instantly recognized, unlike the portrait of a private in his army), the viewer would have to be told, by the title or otherwise, that not only does the painting depict a man in a French general's uniform but that it was intended by the artist to be a portrait of this particular man. Otherwise, how would the viewer know that it did not actually portray his double, or his stand-in? The word represent, as used in connection with art, can mean either "depict" or "portray."

Analysis of representation. Representation always involves a certain degree of abstraction; that is, the taking away of one characteristic or more of the original. Even a fairly realistic painting of a person, for example, lacks some features that characterize actual persons: the painting is two-dimensional, whereas every actual person is three-dimensional; the surface of a painting is paint, but not so the person; the actual person has very numerous pores and other marks on his face that are lacking (in whole or in part) in the painting, and so on. The depiction of a person in a painting is usually sufficient to enable human viewers to recognize the figure as a person—though it is apparently not sufficient for an animal, who sees only a coloured canvas where people see on the coloured canvas a representation. When the degree of abstraction is so great that it is no longer possible to recognize this shape as a human shape or as the shape of any identifiable object, the painting is then spoken of as nonrepresentational. (In popular parlance such paintings are called abstract; but this is misleading, for abstraction is a matter of degree, and, as has just been shown, all depictions are necessarily abstract—that is, abstracted from reality to some degree.) The actual object with all its millions of qualities is at one end of the spectrum, and the painting so abstracted that a depiction subject is unrecognizable is at the other end; between the two extremes lie all the possible degrees of abstraction.

Representation in literature

Literature can be representational but not in the same way as visual art. It is quite natural to say that in a novel or drama a number of characters and actions are represented. The representation is, of course, not a visual one; it is representation through language. The painter portrays Napoleon by making a portrait of him; the writer does so by describing him in words. The writer, unlike the painter, can also depict action. Not all literature, of course, is representational in this way: a sonnet may contain no characters at all and no action, consisting solely of an expression of feeling by some unspecified speaker.

Any of the mixed arts that include words as part of their medium, such as drama or film, can be, like literature, representational. Indeed, they have a further advantage: they can depict action not only through words but also by showing the characters and exhibiting the action before the spectator. These arts are visual as well as verbal, and since they are not limited to one moment in time, as painting and sculpture are, they are temporal arts as well as spatial. These mixed arts, then, can be doubly representational.

Is it possible for music, too, to be representational? Music cannot visually show characters or objects, nor can it describe them in words; can it "depict them in tones"? Program notes at concerts usually assume without question that it can. The audience is told about the tone poem *Don Quixote*, by the Austrian composer Richard Strauss, "The composer has given us a musical representation of the Don's adventures. The 17th-century Spanish writer Miguel de Cervantes has described them in words, and Strauss has done so in tones." But the claim to representation in music is, to say the least, quite dubious. Without the title, with the music alone, would there be any clue that the music was supposed to be "about" the adventures of Don Quixote? True, there is a passage that resembles the bleating of sheep sufficiently for that much to be guessed; but even to conjecture that this passage is a representation of sheep bleating is a far cry from being able to reconstruct the entire story. Suppose that Strauss had left every note in the score just as it was but changed the title; would the piece then have been a representation of something else? The very fact that this question can be asked shows how different music is from visual art: if a painter has drawn a house but indicated in the title that it was supposed to be a tree, the viewer could still say, on the basis of what he saw in the picture, that it was not a tree but a house. But in music the listener is never in this situation: if he says that this series of tones represents the adventures of Don Quixote, he says this because of the title Strauss used. If the composer had given it no programmatic title, one listener might think of one represented subject, another a different one, and a third none at all, and there would be no way of showing who was right or even whose opinion was to be preferred. The conclusion seems to be that music by itself—without title, without words, without depicted action (as in a combination of music and drama such as opera)—is incapable of representing anything. There is simply a series of musical tones that may suggest differing associations, programmatic or otherwise; but the musical tones by themselves cannot be said to represent anything at all.

This might be objected to as an overstatement. If a picture can represent a house by looking more like a house than anything else, cannot a work of music represent the sea by sounding more like the sea than any alternative? And is this not the case in, for example, the French composer Claude Debussy's tone poem *La Mer?* Even this, however, is highly questionable; almost no one guesses the title to Debussy's tone poem without first knowing what it is; it may seem obvious enough after the composer has channelled the listener's response by means of his title but not beforehand. And surely this is because the sounds in the tone poem do not sound more like the sea than like anything else: the tone poem consists, after all, of a series of complex musical tones, emitted by violins, cellos, clarinets, flutes, trumpets, and so on; and it would be difficult indeed for these sounds, which are musical tones, to sound very much like the sea, whose sounds consist after all of a series of complex noises. There is no great similarity between any one series of musical tones and any one series of nature's noises. Hence, the first cannot be said to constitute a representation of the second.

Musical attempts to represent the soundless

The matter is even more obvious in the case of those numerous programmatic titles in which the supposedly represented subject contains no sounds at all. Debussy's *Reflets dans l'eau* (*Reflections in the Water*) is taken by some as a musical representation of reflections in the water. But reflections in the water emit no sounds at all, not even noises. No one, then, could say that the sounds in Debussy's piano composition resemble the sounds of reflections in water. The resemblance, if there is any, is much more remote: it may be that the feeling obtained when Debussy's composition is heard is somewhat like the feeling that arises when reflections in water are seen. This is highly improbable without knowledge of the title, but at most it would provide a mood resemblance, which is far removed from a representation by music of things in the world. The conclusion seems inescapable that music is not to be classified as a representational art, at least not in the same straightforward meaning of "representation" that applies to the other arts.

So much, then, for the capacities of the various arts as far as representation is concerned. But the question remains: in those arts that are properly called representational, what should be the nature of the representation?

That art should be an outright duplication (incorrectly called "imitation") of reality is a view that was put forward by the French novelist Émile Zola in his book *Le Roman expérimental* (*The Experimental Novel*) and has been occasionally held (though not practiced) by painters reacting against Romanticism, such as the 19th-century French artist Gustave Courbet. Zola advocated a novel that resembled a scientific investigation into reality. Plot was to be of no importance, rather an aspect of reality was to be examined searchingly, and from this the story would unfold without imaginative effort. Persons or

groups of persons would be depicted, and from them the action would evolve.

Objections to Zola's theory of duplicating reality

It would be impossible, of course, to carry out such an ideal of art as "report" and undesirable even if it were possible. First, the author or painter must select a subject; and, within the subject, he must select which details to treat, for he cannot in a hundred lifetimes describe them all: since every object and event has an indefinitely large array of qualities, there is no point at which a description of it would be completed. Besides, the very language used (no matter how neutral a description is attempted) will colour the account. Even if the words were colourless, the mode of putting them together would yield a style, which would colour the account once again. Indeed, should such an ideal be achieved (as in the verbatim transcription of an actual trial) it would be the deadliest possible bore.

Art, even representational art, is not a reproduction of reality; it is a transformation of reality. How, specifically, is reality transformed in being represented in art? There is probably no general satisfactory answer to this question. Each art, each style of art, and each work of art transforms reality in its own way—the 19th-century French painters Paul Cézanne in one way, Pierre-Auguste Renoir in another; the 19th-century Russian writers Fyodor Dostoyevsky in one way, Leo Tolstoy in another. No set of rules can lead to predictions as to what transformation of reality will be conceived in the mind of the next creative artist. Reality is the common base, but each artist deals with it in his own unique way.

Subject matter. Do all works of art have a subject matter? The answer to this depends on what is meant by the term subject matter, which signifies basically what the work is about. There are several senses of being "about" that may be referred to:

1. "What is the subject matter of the *Odyssey* by the ancient Greek poet Homer?" The most natural answer would be: "The wanderings of Odysseus." This is the "representational content" of the work. A person who read it simply for the story could easily give this answer. There is contained in the work itself an account of the wanderings of a character named Odysseus, who has no counterpart in the outside world (that is, he is a fictional character) but who does resemble people in the outside world in that he is a man, he is away from home, he is beset by many vicissitudes, and so on. If the subject matter were stated in greater detail, the result would be an account of the plot.

Does painting have subject matter in this sense? If "subject matter" is taken to mean representational content, the answer is often yes; representations of people and trees and the like are easily identifiable in the painting. Sometimes the subject could not be ascertained except for the title, which is not strictly a part of the painting (the title is not a piece of visual art, consisting as it does of words). But quite clearly it cannot be said that the subject matter of the painting is whatever the title says it is: if the title says "Clouds" and the painting is obviously a still life with pears and grapes, it can hardly be said that the painting has clouds as its subject matter simply because the title says so. It was long said that the painting called "Sacred and Profane Love" by the 16th-century Italian artist Titian had sacred and profane love as its subject matter; but Titian himself gave it no such title—the title was added more than a century later by someone else. It cannot be said that the subject matter is what whatever title it came to be given says it is (it could be given numerous incompatible titles); the subject matter, in this sense, must be evident from—or at least not incompatible with—what is seen in the painting itself. The same remarks apply to music: music cannot be "about" anything that the composer happens to seize on as a title; if it were, and he changed the title without changing a note of the music, would the subject matter of the music have changed? Strong reasons already have been given to doubt that music can have a subject matter at all.

2. In addition to a subject matter in the sense of representational content, a work of art may have subject matter in the sense of theme, or underlying idea. The subject matter in the sense of representational content of *Jude the Obscure*, by English novelist and poet Thomas Hardy, is the intellectual ambitions of the main character, Jude, and the vicissitudes befalling him along the way; but the subject matter in the sense of theme or underlying idea is man's struggle to realize his ambitions; and the thesis (implicit message) of the novel is (or may be) that man's highest ambitions are doomed to frustration. Not all novels and dramas and poems have a theme or thesis; they may simply tell a story and nothing more. Works of literature do sometimes operate in both dimensions: a story and a theme or thesis underlying the story. Works of representational visual art may also do so: one might allege that the theme of "Guernica," by the 20th-century Spanish artist Pablo Picasso, is the horror of war. Again, it is doubtful whether works of music can be said to have a theme at all (indeed, the term theme applied to music has a very different meaning: it refers to a set of tones around which variations or developments are composed). If someone were to say that the theme of a work of music was the human condition or man's fear of death, how would such a view be supported? Since music is not representational, what musical passages would he take as evidence for his theory, and why? A composer might indeed title his work "The Human Condition," but this would no more show that the music was about the human condition (had the human condition as its subject matter) than giving the programmatic title "Clouds" to a musical composition shows that the composition actually is about clouds.

Thematic subjects

Symbols in art. Works of art may not only have subject matter, they may also contain symbols. Certain elements in a work of art may represent, say, a whale; but the whale thus represented may be (as it is in *Moby Dick* by the 19th-century U.S. writer Herman Melville) a symbol of evil. In Tolstoy's *Anna Karenina* is represented a gallery of characters dominated by Anna herself, and a tremendous number of actions in which these characters engage; but there is a constantly recurring item in the representational content, namely, the train. Time and again the train causes or accompanies frustration, disaster, betrayal, and other evils—so much so that before the novel is ended it becomes apparent that the train here is a symbol of the iron forces of material progress toward which Tolstoy had such great antipathy.

What is it that makes an item in a work of art a symbol? It is something represented in the work of art—an object, an action, or a pattern of objects and actions, or even (less frequently) simply a nonrepresentational item such as a colour or a line—that does the symbolizing; what is symbolized is a characteristic, such as evil or progress or courage. But by virtue of what does the first (A) become a symbol of the second (B)?

Conventional and natural symbols

The answer is not the same for all symbols, since some are conventional and some are natural. The cross is a symbol of Christianity, and it is a conventional symbol of suffering; in order for it to become a symbol, people had to adopt or accept the cross as standing for suffering. Other symbols are natural—the sun as the symbol of life and strength, a river as the symbol of eternal change and flowing, and so forth; in these cases there was no agreement (convention) as to what would stand for what, for the relation is too obvious—the symbolism is much the same in the tradition of all nations and civilizations.

Various symbols have, to varying degrees, elements of both the conventional and the natural: the eagle on the standard of the United States of America symbolizes strength—this is natural, because the eagle is strong, and conventional, because the eagle was officially adopted as the symbol of the United States. In the case of many symbols, the natural relation between symbol and thing symbolized is not strong enough by itself to achieve the symbolism, and the conventional relation was entered into in order to effect a symbolism that already had a certain natural basis. When A is a symbol of B, there must be some relation, either natural or conventional, between the symbol and what is symbolized. When the natural relation is strong, the conventional is minimal or nonexistent, and vice versa.

But there is another element in symbolism if it is to function with full-bodied effectiveness in a work of art, and that is what has been called a "vital basis" provided by history and tradition. In the U.S. flag there is very little natural basis for symbolism; it is almost wholly conventional. But the fact that it has been saluted for so long, that it has been present on battlefields, that the bodies of military dead have been wrapped in it, and so on—that is, the fact that it has acquired a life, a history—have given it an "affect" (human emotion or mood) it did not have before. Melville and Tolstoy not only represented the whale and the train: they invested these representations with such strong emotional affect that the reader is much more inclined than he would be otherwise to say that these objects are the symbols of the qualities he ascribes to them.

When some item in a work of art is construed as a symbol, it is always infused with these vital qualities. Clearly, not every case of one thing standing for something else is a symbol in this sense: a carrot in a painting does not, just by itself, symbolize growth. But barbed-wire fences do symbolize tyranny, not only because many prisoners have been enclosed in barbed-wire fences (this is the natural basis for the symbolism) but also because of concentration camps and the countless tragic events of recent history that have provided the "vital basis" for the symbolism.

Meaning. Do works of art have meaning? The answer depends once again on how the question is construed: the word "meaning" is an equivocal term that can itself mean many different things.

Clouds mean rain, a falling barometer means that a storm is coming, a twister in the sky means an approaching tornado—that is, the one is a sign of the other; these relations exist in nature and were discovered, not invented, by man. On the other hand, a bell ringing means the end of class, this note on the score means that D sharp is to be played on a certain instrument, and the word cat to someone who knows English means a certain species of domesticated quadruped; these relations are conventional, established by man. But both the natural and the conventional items are examples of meaning in its most general sense—one thing (A) standing for another (B).

Meaning in literature

Since the medium of literature is words, and words are conventional vehicles of meaning, literature has meaning in a way that the other arts do not, since every word, to be a word at all, must have a meaning. In the sense in which the word cat means something, middle C and an ellipse do not have meanings at all.

When the question is asked about meaning in art, however, it is not usually the individual ingredients in it that are being referred to: if it were, the answer could simply be, "Yes, this word has a meaning and that word has; so does the sentence as a whole; and items in paintings sometimes have meaning; for example, the halo over the Madonna's head symbolizes holiness." What is being asked is whether the work of art as a whole has a meaning. But what does the question itself mean? Several different things can be meant: (1) The inquirer may be asking, "What is it about?" in which case the question is about subject matter, already discussed. (2) He may be asking, "What is its theme?"—for example, is the motion picture *He Who Must Die* really a parable about the life of Christ? (3) He may be asking, "What is its thesis?" For example, what is the message of the Anglo-Irish author Jonathan Swift to the reader in "A Modest Proposal"? (4) The inquiry may be about the effects of a work of art on the recipient—either what these effects are or what they could or should be. In this sense, all works of art have meaning, since they all have effects (whether there is one type of effect that a given work of art should have is another question). This is, however, an extremely misleading use of the word "meaning." Indeed, the entire discussion of "meaning in art" is a most confusing one—and the fault does not lie in art, but in the human users of words. Endless unnecessary mysteries can be created by using such nebulous words as "meaning" as if they were simple, straightforward, and susceptible to one interpretation. It would contribute greatly to the clarification of discussions of philosophy of art if the word meaning were not used in them at all but some conception clearer and more specific.

ART AS EXPRESSION

The view that "art is imitation (representation)" has not only been challenged, it has been moribund in at least some of the arts for more than a century. It has largely been replaced by the theory that art is expression. Instead of reflection states of the external world, art is held to reflect the inner state of the artist. This, at least, seems to be implicit in the core meaning of "expression": the outer manifestation of an inner state. Art as a representation of outer existence (admittedly "seen through a temperament") has been replaced by art as an expression of man's inner life.

The terms "express" and "expression"

But the terms "express" and "expression" are ambiguous and do not always denote the same thing. Like so many other terms, "express" is subject to the process–product ambiguity: the same word is used for a process and for the product that results from that process. "The music expresses feeling" may mean that the composer expressed his feeling in writing the music or that the music when heard is expressive (in some way yet to be defined) of human feeling. Based on the first sense are theories about the creation of art; founded on the second are theories about the content of art and the completion of its creation.

Expression in the creation of art. The creation of a work of art is the bringing about of a new combination of elements in the medium (tones in music, words in literature, paints on canvas, and so on). The elements existed beforehand but not in the same combination; creation is the re-formation of these pre-existing materials. Pre-existence of materials holds true of creation quite apart from art: in the creation of a scientific theory or the creation of a disturbance. It applies even to creation in most theologies, except some versions of Christian theology, in which creation is *ex nihilo;* that is, without pre-existing matter.

That creation occurs in various art mediums is an obvious truth. But once this is granted, nothing has yet been said about expression; and the expressionist would say that the foregoing statement about creation is too mild to cover what he wants to say about the process of artistic creation. The creative process, he wants to say, is (or is also) an expressive process; and for expression something more is necessary than that the artist be creating something. Great care must be taken at this stage: some say that the creation of art is (or involves) self-expression; others say that it is the expression of feeling, though not necessarily of one's own feeling (or perhaps that and something more, such as the feeling of one's race, or of one's nation, or of all men); others say that it is not necessarily limited to feelings, but that ideas or thoughts can be expressed, as they clearly are in essays. But the distinctively expressionist view of artistic creation is the product of the Romantic movement, according to which the expression of feelings constitutes the creation of art, just as philosophy and other disciplines are the expression of ideas. It is, at any rate, the theory of art as the expression of feelings (which here shall be taken to include emotions and attitudes) that has been historically significant and developed: art as specially connected with the life of feeling.

When a person is said to be expressing a feeling, what specifically is he doing? In a perfectly ordinary sense, expressing is "letting go" or "letting off steam": man expresses his anger by throwing things or by cursing or by striking the person who has angered him. But, as many writers have pointed out, this kind of "expressing" has little to do with art; as the U.S. philosopher John Dewey said, it is more of a "spilling over" or a "spewing forth" than expression. In art at least, expression requires a medium, a medium that is recalcitrant and that the artist must bend to his will. In throwing things to express anger, there is no medium—or, if the man's body is called the medium, then it is something he does not have

to study to use for that purpose. It is still necessary to distinguish a "natural release" from an expression. If poetry were literally "the spontaneous overflow of powerful feelings," as William Wordsworth said, it would consist largely of things like tears and incoherent babblings. If artistic creation can plausibly be said to be a process of expression, something different from and more specific than natural release or discharge must be meant.

Emotional expression in art

One view of emotional expression in art is that it is preceded by a perturbation or excitement from a vague cause about which the artist is uncertain and therefore anxious. He then proceeds to express his feelings and ideas in words or paint or stone or the like, clarifying them and achieving a release of tension. The point of this theory seems to be that the artist, having been perturbed at the inarticulateness of his "ideas," now feels relieved because he has "expressed what he wanted to express." This phenomenon, indeed a familiar one (for everyone has felt relieved when a job is done), must still be examined for its relevance. Is it the emotion being expressed that counts or the relief at having expressed it? If the concern here is with art as therapy or doing art to provide revelations for a psychiatrist, then the latter is what counts; but the critic or consumer of the art is surely not concerned with such details of the artist's biography. This is an objection to all accounts of expression as process: how is any light at all cast upon the work of art by saying that the artist went through any expressive process or through any process whatever in the genesis of it? If the artist was relieved at the end of it, so much the better for him; but this fact is as aesthetically irrelevant as it would be if he had committed suicide at the end of it or taken to drink or composed another work immediately thereafter.

Another problem should be noted: assuming that the artist does relieve his oppressed state of mind through creating, what connection has this with the exact words or score or brushstrokes that he puts on paper or canvas? Feelings are one thing, words and visual shapes and tones are quite another; it is these latter that constitute the art medium, and in it that works of art are created. There is doubtless a causal connection between the feelings of the artist and the words he writes in his poem; but the expression theory of creation talks only about the artist's feelings, while creation occurs within the art mediums themselves; and to speak only of the former is not to tell anything about the work of art—anything, that is, that would be of interest other than to the artist's psychiatrist or biographer. Through what paroxysms of emotion the artist passed does not matter anymore, insofar as one's insight into his work is concerned, than knowing that a given engineer had had a quarrel with his wife the night before he began building his bridge. To speak of anything revelatory of works of art, it is necessary to stop talking about the artist's emotions and talk about the genesis of words, tones, and so on—items in the specific art mediums.

Distinction between art and craft

The expressionists have indeed brought out and emphasized one important distinction: between the processes involved in art and in craft. The activity of building a bridge from the architect's blueprint or constructing a brick wall or putting together a table just like a thousand others the artisan has already made is a craft and not an art. The craftsman knows at the beginning of the processes exactly what sort of end product is wanted: for example, a chair of specific dimensions made of particular materials. He knows at the beginning how much material it will take to do the job, which tools, and so forth, and, if he does not know these things, he is not a good (efficient) craftsman. But the creative artist cannot work in this manner: "The artist doesn't know what he is going to express until he has expressed it" is a watchword of the expressionist. He cannot state in advance what the completed work of art will be like: the poet cannot say what words will constitute the completed poem or how many times the word "the" will occur in it or what the order of the words will be—when he knows that, he has completed the creation of the poem, and until then he cannot say. Nor could he set about working with such a plan: "I shall compose a poem that contains the word 'the' 563 times, the word 'rose' 47 times," and so on. What distinguishes art from craft is that the artist, unlike the craftsman, "does not know the end in the beginning"; he cannot state until his work is finished what the completed product will be like—if he could, he would not have to undergo the "divine agonies" of creation in order to produce it.

The distinction seems valid enough, but whether it supports the expressionist's view is more dubious; for it can be held regardless of the attitude assumed toward the theory of expression. The open-ended process described as art rather than craft characterizes all kinds of creation: of mathematical hypotheses and of scientific theory, as well as art. What distinguishes creation from all other things is that it results in a new combination of elements, and it is not known in advance what this combination will be. Thus, one may speak of creating a work of sculpture or creating a new theory, but rarely of creating a bridge (unless the builder was also the architect who designed it, and then it is to the genesis of the idea for the bridge, not to its execution, that the word creation applies). This, then, is a feature of creation; it is not clear that it is a feature of expression (whatever is being done in expressing that is not already being done in creating). Is it necessary to talk about expression, as opposed to creation, to bring out the distinction between art and craft?

There does not seem to be any true generalization about the creative processes of all artists nor even of great artists. Some follow their "intuitions," letting their artistic work grow "as the spirit moves" and being comparatively passive in the process (that is, the conscious mind is passive, and the unconscious takes over). Others are consciously active, knowing very much what they want in advance and figuring out exactly how to do it (for example, the 19th-century U.S. writer Edgar Allan Poe in his essay "The Philosophy of Composition"). Some artists go through extended agonies of creation (the 19th-century German composer Johannes Brahms, weeping and groaning to give birth to one of his symphonies), whereas for others it seems to be comparatively easy (Mozart, who could write an entire overture in one evening for the next day's performance). Some artists create only while having physical contact with the medium (for example, composers who must compose at the piano, painters who must "play about" in the medium in order to get painterly ideas), and others prefer to create in their minds only (Mozart, it is said, visualized every note in his mind before he wrote the score). There appears to be no true generalization that can be made about the process of artistic creation—certainly not that it is always a process of expression. For the appreciation of the work of art, no such uniformity, of course, is necessary, greatly though it may be desired by theorists of artistic creation.

Problems in drawing conclusions about the creation of art

The main difficulties in the way of accepting conclusions about the creative process in art are (1) that artists differ so much from one another in their creative processes that no generalizations can be arrived at that are both true and interesting or of any significance; and (2) that in the present stage of psychology and neurology very little is known about the creative process—it is surely the most staggeringly complex of all the mental processes in man, and even his simpler mental processes are shrouded in mystery. In every arena hypotheses are rife, none of them substantiated sufficiently to compel assent over other and conflicting hypotheses. Some say—for example, Graham Wallas in his book *The Art of Thought* —that in the creation of every work of art there are four successive stages: preparation, incubation, inspiration, and elaboration; others say that these stages are not successive at all but are going on throughout the entire creative process, while still others would produce a different list of stages. Some say that the artist begins with a state of mental confusion, with a few fragments of words or melody gradually becoming clear in his mind and the rest starting from there, working gradually toward clarity and articulation; whereas others hold that the artist sets him-

self a problem, which he gradually works out during the process of creation, but his vision of the whole guides his creative process from its inception. The first view would be a surprise to the dramatist who sets himself from the beginning to write a drama in five acts about the life and assassination of Julius Caesar; and the second would be a surprise to artists like the 20th-century English artist Henry Moore, who has said he sometimes begins a drawing with no conscious aim but only the wish to use pencil on paper and make tones, lines, and shapes. Again, as to psychological theories about the unconscious motivations of artists during creation, an early Freudian view is that the artist is working out in his creation his unconscious wish fulfillments; a later Freudian view is that he is engaged in working out defenses against superego charges, "proving something to himself." Views based on the ideas of the 20th-century Swiss psychologist Carl Jung reject both these alternatives, substituting an account of the unconscious symbol-making process. Until a great deal more is known about the empirical sciences that bear on the issue, there is little point in attempting to defend one view of artistic creation against another.

The expressive product. Although talk about expression as a process is hedged with difficulties and in any case seems irrelevant to the philosophy of art (as opposed to the psychology of art), there is another way in which talk about expression may be both true and important to the philosophy of art. Mention is made about expressive properties as belonging to works of art: for example, it is said that a certain melody expresses sadness, that there is a feeling of great calm expressed in a particular painting, or that tension is expressed in the thrusts of a tower or the development of the plot of a novel or drama.

Metaphorical expression of emotion

The question arises at once of what it means to say such things. Melodies and sentences are not joyous or tense or melancholy; only persons have these qualities. The artist can have them, but how can the work of art? Clearly, to speak of a work of art as having emotions, if it is not to be utter nonsense, must be metaphorical. But what is the meaning of this metaphor? What does it mean to say that the music expresses sadness, if not in the sense of process; *i.e.*, that in writing it the composer expressed his sadness?

The music is heard, the painting is seen; each presents itself to the senses. But there is much more involved in music than simply hearing (or even listening to) the sounds and in visual art than simply seeing (or even looking at) the colours and shapes. Even very simple combinations of sounds and shapes and colours seem to express certain qualities of life: a curved line, it is said, is graceful or sprightly; the drooping willow tree is sad, as are certain passages in music. It is virtually impossible for most persons to view art as a series of sensory stimuli only. Even when a picture contains no story, no plot, no program, the viewer "reads into the script": he attributes to works of art qualities of human moods, feelings, emotions—in short, "affects." It would be safe to say that in all art, every percept is suffused with affect. The problem is: What is it that makes certain percepts expressive of certain affects?

The simplest answer, that "The melody is sad" means no more or less than "Hearing the melody makes me (or other listeners, or most listeners) feel sad," is surely inadequate. (This would be a theory of evocation, not of expression.) "The music expresses whatever feelings it arouses in me when I hear it." But often the listener does not feel emotions at all (he may imagine them), or, if he does, he feels very different emotions from the ones he believes to be expressed in the music. He may consider the rondo delirious with joy, but if he is grief stricken on a given day he hears it without feeling joy; and if he has heard the same rondo 30 times that day he feels only boredom or fatigue, while still believing that the piece is expressive of joy. Nor is it an adequate analysis to say that "The melody expresses joy" means "I am disposed to (or inclined to) feel joy when I hear it," for many people seem to recognize joy as a quality of the music without feeling it at all: or they may imagine it or just recognize the emotion without feeling it or believe that what they hear sounds the way joy feels—or any of a number of other accounts.

The true analysis of expressiveness in art must be more complex than this: it is not that the melody evokes emotion X, but that emotion X is somehow embodied in the music. But this leads back again to the question, how can an emotional quality be in a work of art? There is no single answer to this question that would be accepted by all philosophers of art, but most accounts begin by noting certain similarities, or analogies, between features of music and features of human feeling; so that when X (a passage of music, for example) is said to express Y (a state of feeling), there are certain similarities (for example, of structure) between X and Y. The physical accompaniments to a mood, say, of restlessness, such as rapid breathing and drumming fingers, have their musical equivalents: trills, quavers, increases in tempo, and the like.

Emotional qualities embodied in art

When a listener says that a certain melody is sad, he is saying that the music literally has certain qualities A, B, C, D that can be perceived in the music. Slowness is surely one such quality (the same melody played fast would not be called sad); another is the absence of large intervals between tones; another is that the sounds tend to be hushed rather than, for example, strident; another is that the tendency of the musical movement is downward rather than rising. When a listener says that the music is sad, he is saying that it has these qualities.

But why these qualities rather than others? Why is it said that the music is sad when it has A, B, C, D rather than when it was M, N, O, and P? Because A, B, C, D are the qualities that also characterize people when they are sad, such as slowness and soft and low speaking voices. If this theory or anything like it is true, it explains how emotional characteristics can be attributed to works of art—why it can be said that the melody is sad, that the horizontal lines in a painting make it calm (horizontal being the position of rest and peace, sleep and maximum relaxation, and from which one does not fall), why the lines in a painting are droopy (they are lines similar in shape to that of, say, an old woman with hunched shoulders), and so on. These qualities, it should be noted, are qualities of the work of art, not of the artist (whether he was sad when he wrote sad music is a separate question, to which the answer may sometimes be no) and not of the listener or observer (the melody is sad even if I am not sad when I hear it). They are, so to speak, embodied in the music, quite independently of the state of the artist or of the observer, or listener, though of course it requires the presence of a listener to recognize them and be moved by them.

Talk about artistic expressiveness, then, can be justified. But there is no need to resort to the language of expression in order to state it: instead of saying, "The music expresses sadness," it can simply be said, "The music is sad." But regardless of the terminology employed, it is important to have justified this conclusion.

ART AS FORM

The formalist position. Against all the foregoing accounts of the function of art stands another, which belongs distinctively to the 20th century—the theory of art as form, or formalism. The import of formalism can best be seen by noting what it was reacting against: art as representation, art as expression, art as a vehicle of truth or knowledge or moral betterment or social improvement. Formalists do not deny that art is capable of doing these things, but they believe that the true purpose of art is subverted by its being made to do these things. "Art for art's sake, not art for life's sake" is the watchword of formalism. Art is there to be enjoyed, to be savoured, for the perception of the intricate arrangements of lines and colours, of musical tones, of words, and combinations of these. By means of these mediums it is true that objects in the world can be represented, scenes from life depicted, and emotions from life expressed; but these are irrelevant to the principal purpose of art—indeed, art is much less adapted to the telling of

"Art for art's sake"

a story or the representation of the world than it is to the presentation of colours, sounds, and other items in the art medium simply for their own sake.

Most people who claim to enjoy paintings, for example, enjoy them not as presentations but as representations of things and situations in life; and thus their response is not of a kind that is unique to art, but one that takes them back to the emotions of life, from which they came. They could use art to take them into a realm of pure form unknown to anyone who is unacquainted with art; but instead they use it to direct them back to the feelings and situations of life. Thus, according to the formalists, these viewers miss the opportunity of being taken into a fresh world of purely aesthetic experience and get from a work only what they bring to it: familiar experiences and emotions they employ the work to recall.

What, then, should be brought from life to art? Knowledge of life's struggles and emotions? Knowledge at least of what people are like, and what visual objects look like? Not even these things, for even they get in the way. Representation is not bad in itself, it is merely irrelevant. Only if the representation is satisfactory as form and contributes to the general abstract design can it be said to matter aesthetically.

Most formalists have directed their attention primarily to visual art. The prerequisite for appreciating this, they believe, is a sense of form and colour and a knowledge of three-dimensional space (the last required because otherwise a cube, for example, would appear in a painting as a flat pattern and would be unable to play the architectural role intended for it). Armed with this bit of knowledge from life, they have all they need (as far as knowledge of the world outside art is concerned) for appreciating visual art. Armed with more than this, they would find their attention drawn away from the sublimities of art to the more approachable concerns of humanity (such as representation). Shorn of this extraneous knowledge and coming to painting with eyes innocent of extraneous concepts, the viewer could then be in a position to look at what painting presents directly to his vision—complex arrangements of forms and colours—which, for reasons thus far unexplained, have the capacity to move the recipient deeply with emotions utterly alien to the emotions of life.

The formalist's account of music runs along similar lines: not only representation (program music) is excluded but so is the entire realm of human feeling—not the feeling that the contemplation of pure form can give but only the feelings of life, such as love or terror. As to literature and all those arts in which words play a part (song, opera, drama, cinema), the formalist position would seem to be impossible; and indeed formalists have seldom attempted to extend their theory to literature. The medium of literature is words and sentences; and words and sentences are not merely noises but noises with meanings; and these meanings inevitably have to do with the objects, actions, qualities, and situations of life. As sound, literature is a poor thing—as a complex of meanings, it can be profound and beautiful. Take these away, and literature could cease to be. Literature does have formal properties—a drama can be as tightly knit as a fugue or a symphony—but the appreciation of it can never consist of these formal properties alone; and the reason lies in the very nature of the medium. It is the role of words to indicate images and meanings and emotions, and these are the stuff of life—therefore verbal art is inescapably humanistic. Whatever sensory beauty words have in their sounds is slight and secondary; not so with painting, where colour, line, and space have a beauty all their own and need stand for nothing outside themselves to satisfy the eye.

Inapplicability of the formalist position

Formal principles in art. What, then, are the specific qualities in works of art that the formalist is seeking? Most formalists have held that a partial account can be given of these but that, in the end, the presence of the qualities must be felt intuitively and cannot be described. Accounts of formal qualities in works of art go back as far as Aristotle's *Poetics*, written in the 4th century BC, and usually include (though sometimes in different terminology) the following as principal ingredients:

Organic unity. A work of art must have what Aristotle called "a beginning, a middle, and an end"; it must be unified, it must "hang together" as one entity. Everything, of course, has some degree of unity or other; even a collection of things, such as a woodpile, has some unity inasmuch as it can correctly be called one thing: it is a collection, but it is a single collection. But the unity desired in works of art is much greater than this: it is more like the unity of the higher organisms in which every part functions not independently of the others but interdependently with them; and it is this interdependency of the parts that constitutes an organic unity. Take away one part, and the remainder of the parts fail to function as before. This is only approximately true of organisms: without a heart or a brain a person could not continue to exist, and the activity of the other organs would cease; but without an ear or a toe they surely would. Philosophers of art have often noted that the purest examples of organic unity in the universe are not organisms but works of art: here the interdependency of parts often achieves a state of such perfection that it could often be said, of a melody or a sonnet, that if this note (or word) were not there, in just the place that it is, the effect on the entire remainder of the melody or poem would be disastrous.

Complexity, or diversity. This principle is the natural accompaniment of the first one. A blank wall has unity but no variety and is not long worth contemplating. Nor is there any triumph in achieving unity at so small a price. The work of art must hold in suspension (as it were) a great diversity of elements and unify them—the greater the complexity that is integrated into a unity, the greater the achievement. This fact is so universally recognized that the two criteria are often stated as one, unity-in-diversity, or variety-in-unity.

Unity-in-variety in works of art

Very many great works of art are much less than perfect organic unities (which is another way of saying that unity-in-variety is not the only criterion for excellence in works of art). Particularly in long poems or novels or operas, some parts are clearly more important than the others, though contributing to the whole, and some parts may be simply "padding." One could hardly allege that the entire ancient Greek epic the *Iliad* is an organic unity and that if (for example) the catalog of ships were removed the entire epic would be ruined (some even say it would be improved). In Dostoyevsky's novels there are whole chapters that are unnecessary from the point of view of relevance to the rest of the story, and, aesthetically (though perhaps not in other ways), these are a pure excrescence. In most works of art there are high spots and low spots, and there is a great deal of elasticity as to what could follow what. But in spite of this, unity-in-variety is quite universally recognized as a criterion for artistic excellence. If a drama consisted of two plots that never connected with one another, even at the end, such a play would be condemned at once for lack of unity; and a work of art is never praised for being disjointed or disunified, though it might be praised in spite of being disunified.

Theme and thematic variation. In many works of art there is a dominant theme, or motif, which stands out and upon which the other portions are centred. This theme is then varied in different ways in other portions of the work. This is a special case of unity-in-variety: if every line in a work of music or literature were entirely novel and different from the other ones, there would be enormous diversity but no unifying connecting links; whereas if there were simply a repetition of the initial theme or of entire sections of the work (as sometimes happens when a composer does not know how to develop the thematic material with which he has begun) there is unity but no variety. Both unity and variety are preserved by having central themes, with other material that is related to them (unity) but not identical with them (variety).

Development, or evolution. In works of temporal art, each part develops or evolves into the next, each part being necessary to the succeeding part, so that if an ear-

lier part were altered or deleted, all the subsequent parts would have to be altered in consequence. If a portion of Act IV could be interchanged with a portion of Act II without loss of effect, the principle of development has not been observed, for then the material occurring in between would not have made any difference.

Balance. The arrangement of the various parts should be balanced, usually in contrasting ways (the adagio movement coming between two faster movements, for example). In painting, there should be a balance between the right and left halves of the canvas. The many ways, other than simple mechanical symmetry ("for every item on the left there should be an item on the right," which soon becomes monotonous), in which a painting may have variety and yet retain balance are too complex to be discussed other than in a book of art criticism. But in its simplest essentials, the principle is acknowledged by everyone: the housewife who places all the furniture on one half of the living room while leaving the other half empty finds the arrangement aesthetically displeasing because the room lacks balance.

Generalizations based on formal principles

There are many descriptions of principles of form in art, which differ from one another in their terminology more than in their final outcome. In general, however, few if any of these principles would be denied (only the detail of their formulations might be) by most philosophers of art. Why certain principles of form are found satisfactory and others are not is a fascinating psychological question, leading back to a discussion (necessarily vague in the present state of knowledge) of the nature of the human organism and the discriminatory powers of the human mind. But however obscure the explanation, the facts of the case seem clear enough: certain formal principles in art must be observed, and, to the degree that they are ignored or violated, aesthetic catastrophe occurs: the work of art cannot evoke our interest or sustain it long once it is initiated.

Though a discussion of these formal principles is helpful particularly to those who have little native sense of form and who want "to know what to look for" in art, they are sufficiently vague so that many critics can agree on a formal principle, such as unity, and yet disagree on the degree to which a specific work possesses it.

In addition, these principles are far from complete: a work of art can possess unity and the other requirements in high degree and yet be unsuccessful even as form. The requirements listed seem only to have skimmed the surface. Yet what more is required to distinguish a formally correct but dull work of art from a brilliant one seems to defy precise analysis. Moreover, the majority of critics who have assented to these principles are not formalists: they have acknowledged and even insisted on the great importance of form in works of art, but they have not alleged, as formalists do, that these principles constitute the sole criteria of excellence in works of art. They have held that the fulfillment of formal criteria counts as a necessary condition for artistic excellence but not a sufficient condition.

PRAGMATIC THEORIES OF ART

There are theories of art that differ from one another in what they allege to be the real purpose or function of art but are at one with each other in the belief that art is a means to some end, whether that end be the titillation of the senses or the communization of the nations of the world or the conversion of mankind to belief in God or the improved moral beliefs or moral tone of the reader or viewer. In every case, the work of art is considered as a means to some end beyond itself, and hence what counts in the final analysis is not the nature of the work of art itself but its effects upon the audience—whether those effects be primarily sensory, cognitive, moral, religious, or social.

Hedonistic theories of art. According to one kind of theory, the function of art is to produce just one kind of effect upon its audience: pleasure. It may also inform or instruct, represent or express, but first and foremost it must please. The more pleasure it gives, the better the art.

If the theory is left in this simple form, it yields the result that glossy and superficial works and those containing nothing difficult or obscure are the best works of art: thus, on the hedonistic account, *King Lear* might come out far behind Henry Wadsworth Longfellow's *The Song of Hiawatha*, or Joyce Kilmer's "Trees," in view of the difficulty of comprehending Shakespeare by many people and the pleasant, easy lilting quality of Longfellow's poem; and similarly a simple ditty might come out ahead of Bach's *Mass in B Minor*. True, Shakespeare and Bach might produce more pleasure in the long run since their works have endured through more centuries. But on the other hand, the simple works can be apprehended and enjoyed by vastly more people.

The meaning of "aesthetic pleasure"

In any case, the theory has often been amended to read "aesthetic pleasure" rather than simply "pleasure"—thus placing great importance on exactly how the term "aesthetic" is to be defined. The definition of this troublesome term is beyond the scope of this article (see AESTHETICS); it will simply be said here that no quick and easy way of distinguishing aesthetic pleasures from other pleasures will suffice for the task at hand. If it is said, for example, that aesthetic pleasure consists in satisfaction taken in the contemplation of sensuous particulars (tones, colours, shapes, smells, tastes) for their own sake—that is, for no further end and without ulterior motive—then one confronts the fact that as much pleasure may be taken in single smells and tastes for their own sakes, without any reference beyond them, as may be taken in the most complex works of art. For that matter, pleasure in playing a game (one not played for money) is pleasure in doing something for its own sake, as is the pleasure of robbing a house if it is done not for money but for "kicks." If something is found pleasurable, ordinarily the pleasure is what one wants from it, not something else beyond it.

Moreover, if it is said that a work of art should be a means toward pleasure, that is treading suspiciously near to the opposed view that art should not be a means to an end but an end in itself. If someone says, "Why do you go jogging every morning for three miles? Because you feel the exercise is good for you?" and another person answers, "No, not that at all, I just enjoy doing it," this would ordinarily and quite sensibly be taken as saying that he did not do exercise as a means toward an end but as an end in itself. If something is done just because it is enjoyed, in common parlance this would be taken to be "doing it as an end in itself"; and if one objected, "No, I'm not doing it as an end in itself, I'm doing it as a means toward the enjoyment I'll get out of it," his reply would be considered sophistical, for doing it for enjoyment's sake is precisely what is ordinarily meant (or one thing that is ordinarily meant) by the statement that a thing is being done for its own sake.

In any case, the effect of great works of art upon a reader or viewer or listener can hardly be described as merely hedonistic. No one would presumably wish to deny that art can and should give us pleasure; but few would wish to assert that pleasure is all that it should give us. If one were to ask, "How did viewing Picasso's 'Guernica' affect you?" and the reply was, "I found it pleasant," we would conclude that his reaction to the painting was, to say the least, inadequate. Great art may please; it may also move, shock, challenge, or change the lives of those who experience it deeply. Pleasure is only one of many kinds of effects it produces.

Art as a means to truth or knowledge. One of the things that has been alleged to be the purpose of art is its cognitive function: art as a means to the acquisition of truth. Art has even been called the avenue to the highest knowledge available to man and to a kind of knowledge impossible of attainment by any other means.

Knowledge in the most usual sense of that word takes the form of a proposition, knowing that so-and-so is the case. Thus, it can be learned from sense observation that the sun is setting, and this is knowledge. Is knowledge acquired in this same sense from acquaintance with works of art? There is no doubt that there are some propositions (statements) that can be made after acquaintance with works of art that could not be made before: for ex-

ample, that this performance of Beethoven's *Eroica Symphony* was 47 minutes long, that this painting predominates in green, that this piece of sculpture originated around 350 BC. The question is whether there is anything that can be called truth or knowledge (presumably knowledge is of truths, or true propositions) that can be found in works of art.

The cognitive function of literature

Literature is surely the most obvious candidate; for literature consists of words, and words are combined into sentences, and sentences (at least declarative sentences) are used to convey propositions; that is, to make assertions that are either true or false. And works of literature do certainly contain many true statements: a novel about the French Revolution conveys facts about the series of events; in a verse of the English scholar and poet A.E. Housman (1859–1936), it is said that "The tears of all that be/ Help not the primal fault." Since literature contains statements, it would be surprising indeed if at least some of them were not true.

But the relevance of this fact to literature as an art is extremely dubious. If an 18th-century novel gives a true picture of English country life of that time, this makes it useful to read as history; does it also make it a better novel? Many, at any rate, would say that it does not: that a tenth-rate novel might give more facts about 18th-century life than a first-rate novel of the same century. For that matter, many of the propositions in a novel are, taken at face value, false; it is false, for example, that there was a foundling named Tom Jones who had an uncle named Squire Western. The thousands of pages of description in novels of fictional characters, ascribing to them thoughts and actions, are all false, since these characters never actually existed. Yet this fact in no way impugns their value as literature. Shakespeare, in *The Winter's Tale*, sets part of the action on the seacoast of Bohemia; but the fact that Bohemia has no seacoast does not damage *The Winter's Tale* as literature, though it would as geography. The fact that Milton used the outdated Ptolemaic astronomy does not make *Paradise Lost* less valuable, nor does the nonexistence of the lands described in *Gulliver's Travels* in any way diminish Swift's work. There is no doubt, then, that works of literature can contain true statements and false ones; but it is tempting to ask, what does their truth or falsity matter? Literature is not astronomy or geography or history or any branch of knowledge, particular or general.

Many would hold that the above statements are indeed irrelevant, as are any that encroach upon the domain of science; but, they would add, there are other assertions that matter a great deal: for example, the statements in which a world view is presented in a poem or drama or novel. The main burden of the ancient Latin poet Lucretius' *De rerum natura* ("On the Nature of Things") is a presentation of the materialism of the Greek philosopher Democritus; and an embodiment of the world view of medieval Catholicism is the very warp and woof of Dante's *Divine Comedy*—and such considerations (it would be contended) are relevant to these works as literature.

In reply, however, it might be said that while it is true that these world views must be understood and taken into consideration in the reading of these poems and that they cannot be understood or appreciated without knowing them, the truth or falsity of these views still does matter aesthetically. If Lucretius' view is true, then Dante's must be false, and vice versa, since they are incompatible; but in order to appreciate the poem it is not necessary to know which (if either) is true. Appreciating art, unlike taking a stand for or against a cause in life, does not require a yes or no to statements. It requires only that the viewer look and appreciate, that he experience as richly and fully as possible the feelings and attitudes involved in the world view that is presented. Philosophers and scientists are concerned with whether the Democritean materialism of Lucretius is true; appreciators of art are concerned only to capture the feeling appropriate to the world view in question.

Many statements in works of literature are not explicitly made at all but are implicit: Hardy never tells in his novels what his world view is, but it emerges rather clearly before the reader is halfway through any of them. Probably the most important points made in works of literature that contain a central thesis are implicit rather than explicit. How, in that case, can it be determined what thesis it is that is implied? In a court of law, if someone says, "He didn't say it exactly, he just implied it," the judge would be likely to rule that this was insufficient evidence of slander, since the person did not actually say it. Still, many statements in daily life are not stated but implied—in the sense that they are intended; the trouble lies in proving that the speaker intended them, since no one else is in a comparable position to say what his intentions were; and, in the case of deceased authors, there is no evidence of their intentions other than what they said. One is doubtless on safer ground, therefore, saying that many statements are implied in the sense that they are suggested (whether the speaker intended to do so or not) by the tone of voice and the juxtaposition of the words used. Thus, "They had children and got married" suggests, though it does not state, that they had the children before they were married; any normal user of the English language would tend to construe it thus. And it is surely no overstatement to say that Swift's *Gulliver's Travels* suggests that the author was misanthropic or that the novels of the French author Marcel Proust (1871–1922) suggest a pessimistic view of love and other human relationships close to that of the German philosopher Arthur Schopenhauer (1788–1860). A serious reader of literature will become increasingly sensitive to what is suggested in the works he is reading.

Explicit and implicit statements in literature

But, once again, the importance of the suggested statements, even when they are true, in no way shows that they must be accepted as true by the reader if he is to value them as works of art. Is the sincere Roman Catholic who finds Dante's world view congenial and Lucretius' repellent committed to saying that Dante's is the better poem? If so, he may be accused of confusing his moral and theological judgments with his aesthetic ones. Still, it should be noted that there are some critics who believe that if two works of literature are both equal in excellence on all counts, yet one presents a true view of reality and the other fails to, the one presenting a true view is better—better even as a work of art—than is the other one.

There is, however, another way of talking about truth in literature that is not or is not as obviously connected to propositions. A characterization in a novel or drama is spoken of as being true to human nature, true to the way people actually speak or behave or feel. No matter that Becky Sharp—in the English novelist William Thackeray's *Vanity Fair*—is a fictional character, it would be said, as long as she is depicted as a person of a certain type would behave, she is being depicted truly; truth in fiction does not mean truth of the statements (for the statements in Thackeray's novel describing her are false), but truth to human nature.

The notion of "truth to human nature" in literature

But what exactly does "truth to human nature" mean? The criterion is as old as Aristotle, who wrote that poetry is more true than history because it presents universal truths whereas history gives only particular truths and that poetry (dramatic fiction) shows how a person of this or that kind probably or necessarily would behave (or think, or feel). This criterion, however, is too vague as it stands: what is probable or plausible behaviour in one person is not in another, and what is probable in one set of circumstances is not so in another. The test of truth to human nature would be roughly as follows: Would a person such as has been described thus far (in the novel or drama) behave (or think or feel or be motivated) in the way that the author depicts this character as behaving in the circumstances described? It is often very difficult to decide this question, because knowledge of human beings is insufficient or because the dramatist himself has not provided enough clues. Still, once readers or critics are convinced that the character described would not have behaved as the novelist depicts him as behaving, they may criticize the characterization (at least with regard to this bit of behaviour or motivation) as implausi-

ble. If a character who has been described as spending years working toward a certain goal is represented by the novelist as abandoning it once he is within sight of it, the reader will have considerable reservations about this delineation unless the author has depicted the character as being unstable or masochistic or in some way as being the kind of person who might in these circumstances do this kind of thing. It is true that there are people in the world who abandon their goals within sight of them after years of labour, but the conviction must be implanted that the character already presented by the novelist belongs to this classification or the behaviour will seem reasonless and unmotivated.

Is truth to human nature aesthetically relevant? That is, when present does it make the work of literature better and when absent or flawed does it make the work worse as literature? Here again there would be some difference of opinion, but a very large number of critics and aestheticians, in the tradition of Aristotle, would say that it matters aesthetically a great deal. The novelist does not have to be true to geography or history or astronomy, but he must be, as the 19th-century U.S. author Nathaniel Hawthorne said of all literary artists, true to the human heart. A literary artist may tamper with all the other truths with impunity, but not this one: his characters must be convincing, and they will not be convincing if they are not depicted as having anger, love, jealousy, and other human emotions that real people have and in pretty much the contexts in which real people have them. If a novelist's characters were not motivated in much the way that human beings are motivated, the reader would not even be able to understand them—they would be alien and unintelligible to him. Even when a writer (such as the Englishman Kenneth Grahame in *The Wind in the Willows*) depicts animals as central characters in novels, however much they may differ from human beings in external appearance, they must psychologically be presented as human beings—how else and in what other terms could their behaviour and their motivation be understood? Such, then, are the reasons for saying that whatever else a literary artist does, his depictions must be truthful to human nature.

Truth to human nature in arts other than literature

Can works of art other than literature possess truth to human nature? It would seem that in a limited degree they can. Motion pictures and operas and other mixed arts clearly can, but they employ words, and literature is a principal ingredient in them. But what of arts that employ no words at all? Painting and sculpture, not being temporal arts, cannot depict action, and action is all-important in the representation of human character. These arts contain depictions of persons (real or imaginary) only in a knife-edge of time. Still, sometimes something may be inferred even from a knife-edge. The late self-portraits of the 17th-century Dutch artist Rembrandt do seem to reveal an agonized yet sometimes serene inner spirit, suggesting that there are flashes of human insight to be found in depictions of human beings in visual art. As for musical art (music without the accompaniment of words), it contains nothing that could be called depiction, not even depiction at a knife-edge of time; and, if this is so, there can be no such thing here as true depiction or false depiction. Music may be expressive of human feelings, in the sense already described, but this is a far cry from saying that it contains depictions that are true to human nature.

Even if truth to human nature in the depiction of character is aesthetically relevant (which many would question), to say this is still far from saying that it is the only criterion for excellence in works of art, or even that this is the principal thing that art gives or its main excuse for being. To go so far would be to discount colour and form and expressiveness as criteria for excellence in art; and this virtually no one is willing to do. It would seem, then, that in no case is truth (even truth to human nature) necessary in works of art, seeing that entire genres of art, such as music, exist without it; and that even when it is present and when its presence increases the merit of a work of art (which again many would deny), it is only one virtue among many. Thus, the view that the purpose or function of art is to provide truth is quite surely mistaken; perhaps the person who wants truth and is indifferent to the presence of anything else had better turn to science or philosophy rather than to the arts.

Art as a means to moral improvement. To say that a work of art is aesthetically good or has aesthetic value is one thing; to say that it is morally good or has a capacity to influence people so as to make them morally better is another. Yet, though the two kinds of judgments differ from one another, they are not entirely unrelated. Three views on the relation of art to morality can be distinguished:

Moralism. According to this view, the primary or exclusive function of art is as a handmaiden to morality —which means, usually, whatever system of morality is adhered to by the theorist in question. Art that does not promote moral influence of the desired kind is viewed by the moralist with suspicion and sometimes with grudging tolerance of its existence. For art implants in people unorthodox ideas; it breaks the molds of provincialism in which people have been brought up; it disturbs and disquiets, since it tends to emphasize individuality rather than conformity; and works of art are often created out of rebellion or disenchantment with the established order. Thus, art may undermine beliefs and attitudes on which, it is thought, the welfare of society rests and so may be viewed with suspicion by the guardians of custom. When art does not affect people morally one way or the other (for example, much nonrepresentational painting), it is considered a harmless pleasure that can be tolerated if it does not take up too much of the viewer's time; but, when it promotes questioning and defies established attitudes, it is viewed by the moralist as insidious and subversive. It is viewed with approval only if it promotes or reinforces the moral beliefs and attitudes adhered to by the moralist.

Plato's moralistic view of art

Plato is the first champion in the Western world of the moralistic view of art—at least in *The Republic* and *Laws*. Plato admired the poets and was himself something of a poet; but, when he was founding (on paper) his ideal state, he was convinced that much art, even some passages in Homer, tended to have an evil influence upon the young and impressionable, and accordingly he decided that they must be banned. Passages that spoke ill or questioningly of the gods, passages containing excessive sexual passion (and all works that would today be described as pornographic), and even passages of music that were disturbing to the soul or the senses were all condemned to the same fate. Plato's concern here was with the purity of soul of the men who would become members of the council of rulers of the state; he was not concerned with censorship for the masses, but, since one could not predict which young people would pass the series of examinations required for membership in the council of rulers and since it was (and is) practically impossible to restrict access to works of art to a certain group, the censorship, he decided, would have to be universal. The objection might be raised, to be sure, that rulers to be should not be hothouse plants separated from the influences of the outside world and that they would be better off facing all of reality, including its evils. But Plato's view was that these influences should be kept from them during their formative years—that during this critical time, when the whole tenor of their lives was being shaped, art could be an influence for evil and had to be sacrificed in the interests of morality. In other dialogues of Plato, such as the *Ion* and the *Phaedrus*, when he was not concerned with building a state, he extolled the virtues of art and even held the artist to be divine (although madly divine); but when it came to a conflict between art and morality, it was art that would have to go.

Tolstoy's moralistic view of art

The most famous champion of the moralistic view of art in modern times is Tolstoy. Long after he had finished writing his novels, he fell under the influence of primitive (prechurch) Christianity, the principal tenet of which was the brotherhood of all mankind. This one idea became such an obsession with him that everything else, includ-

ing the pursuit of art to which he had devoted his life, became subordinate to it. Almost all the literature of his own time, including all his own novels, he condemned as inimical to the brotherhood of man by emphasizing class distinction and pitting one group of mankind against another. Even art that appealed primarily (in his opinion) to "upper class" tastes, such as the symphonies of Beethoven and the operas of Richard Wagner, both 19th-century German composers, were condemned as "false art." The art that remained after these colossal excisions included such items as folk songs that peasants might sing in the fields as they worked and pictures and stories either illustrating the tenets of primitive Christianity or fostering the spirit of Christianity by promoting the brotherhood of all mankind.

The moralistic view of art is still, on the whole, the unarticulated view of art held by the masses, particularly when they are under the sway of a dominant religious or political doctrine. Historically, Christianity has been suspicious of all art except those works that depicted some aspects of biblical history or could be used to further the spread of Christian belief and practice (although this is no longer strictly true). It would probably be fair to say that the view of art held by the Soviet government is a moralistic one: works of fiction and poems must praise Communism or further its doctrines, and works of music must be melodic and singable (Soviet composers such as Dmitry Shostakovich have often been condemned by the official hierarchy as "too German" or "too materialistic"). Whenever a culture or nation is under the sway of a dominant view, whether moral or religious or political, the tendency of the rulers of that nation is to promote it at all costs—and one of the casualties in the process is art, at any rate that great body of art that is either indifferent or hostile to the reigning dogma.

Aestheticism. Diametrically opposed to the moralistic view is aestheticism, the view that instead of art (and everything else) being the handmaiden of morality, morality (and everything else) should be the handmaiden of art. The proponents of this view hold that the experience of art is the most intense and pervasive experience available in human life and that nothing should be allowed to interfere with it. If it conflicts with morality, so much the worse for morality; and if the masses fail to appreciate it or receive the experience it has to offer, so much the worse for the masses. The vital intensity of the aesthetic experience is the paramount goal in human life. If there are morally undesirable effects of art, they do not really matter in comparison to this all-important experience which art can give. When the son-in-law of the 20th-century Italian dictator Benito Mussolini waxed lyrical in his description of the beauty of a bomb exploding in the midst of a crowd of unarmed Ethiopians, he was carrying to its fullest extent the aestheticist's view of art.

Aesthetic values versus moral values

Few persons would wish to go so far. Even the most ardent lovers of art would stop short of saying that the value of art holds a monopoly over all other values. It may well be that the experience of works of art is the greatest experience available to human beings (though this, too, could be questioned), but at any rate it is not the only one available, and, this being the case, the others should be considered as well. There is a plurality of values; and aesthetic values, although far greater, admittedly, than most persons realize, are still just a few among many. It is therefore necessary to consider the relation of the values derived from art to the values derived from other things, such as the conduct of life apart from art: no one can devote every waking hour to the pursuit of art, even if for no other reason than the need for survival, and thus the values of such mundane things as food and shelter have also to be considered.

Mixed positions. The moralistic and aesthetic positions are extremes, and the truth is likely to be found somewhere between them. Indeed, art and morality are intimately related, and neither functions wholly without the other. But to trace the precise relations between art and morality is far from easy; for want of a better term, "interactionism" could be used to label the view that aesthetic and moral values each have distinctive roles to play in the world but that neither operates independently of the other.

It would be admitted, first of all, that works of literature (which will be examined first, since of all the arts the relation of literature to morality is most obvious) can teach valuable moral lessons through explicit presentation: the genre that has this as its aim is didactic literature, as exemplified by *Pilgrim's Progress* by the English Puritan John Bunyan and *Back to Methuselah* by the Irish dramatist George Bernard Shaw. But most works of literature do not exist to teach a moral lesson: possibly, Shakespeare did not write *Othello* merely to attack racial prejudice or *Macbeth* to prove that crime does not pay. Literature does teach but in a far more important way than by explicit preachment: it teaches, as John Dewey said, by being, not by express intent.

Expression of morality in literature

How does literature achieve this moral effect? It presents characters and situations (usually situations of difficult moral decision) through which the reader can deepen his own moral perspectives by reflecting on other people's problems and conflicts, which usually have a complexity that his own daily situations do not possess. He can learn from them without himself having to undergo in his personal life the same moral conflicts or make the same moral decisions. The reader can view such situations with a detachment that he can seldom achieve in daily life when he is immersed in the stream of action. By viewing these situations objectively and reflecting on them, he is enabled to make his own moral decisions more wisely when life calls on him in turn to make them. Literature can be a stimulus to moral reflection unequalled perhaps by any other, for it presents the moral choice in its total context with nothing of relevance omitted.

Perhaps the chief moral potency of literature lies in its unique power to stimulate and develop the faculty of the imagination. Through literature the reader is carried beyond the confines of the narrow world that most persons inhabit into a world of thought and feeling more profound and more varied than his own, a world in which he can share the experiences of human beings (real or fictitious) who are far removed from him in space and time and in attitude and way of life. Literature enables him to enter directly into the affective processes of other human beings, and, having done this, no perceptive reader can any longer condemn or dismiss en masse a large segment of humanity as "foreigners" or "wastrels," for a successful work of literature brings them to life as individuals, animated by the same passions as he is, facing the same conflicts, and tried in the same crucible of bitter experience. Through such an exercise of the sympathetic imagination, literature tends to draw all men together instead of setting them apart from one another in groups or types with convenient labels for each. Far more than preaching or moralizing, more even than the descriptive and scientific discourses of psychology or sociology, literature tends to unite mankind and reveal the common human nature that exists in everyone behind the facade of divisive doctrines, political ideologies, and religious beliefs.

This is not to say, of course, that those who read great works of literature are necessarily tolerant or sympathetic human beings. Reading literature alone is not a cure for human ills, and people who are neurotically grasping or selfish in their private lives will hardly cease to be so as a result of reading works of literature. Still, wide and serious reading of literature has an observable effect: people who do this kind of reading, no matter what their other characteristics may be, do tend to be more understanding of other people's conflicts, to have more sympathy with their problems, and to be able to empathize more with them as human beings than do people who have never broadened their horizons by reading literature at all. No one who has read great literature widely and for a considerable period, so as to make it an integral part of his life, can any longer share the same provincialism and be dominated by the same

The leavening influence of literature

narrow prejudices that seem to characterize most people most of the time. Literature, perhaps more than anything else, exercises a leavening influence on the temper of a man's moral life. It looses him from the bonds of his own position in space and time; it releases him from exclusive involvement with his own struggles from day to day; it enables him to see his own local problems and trials from the perspective of eternity—he can now view them as if from an enormous height.

To have moral effects, it is not necessary that a work of literature present a system of morality. Its moral potency is perhaps greatest when it presents not systems but human beings in action, so that through the exercise of the imagination the reader can see his own customs and philosophies as he sees theirs: as some among many of the countless adjustments and solutions to human problems that different circumstances and man's endlessly varied and resourceful nature have produced.

Works of literature, then, develop more than anything else the human faculty of the imagination; and the 19th-century English poet Percy Bysshe Shelley said that the imagination is the greatest single instrument of moral good. Perhaps this sounds like an absurd overstatement, but consider what morality is like without the imagination. Consider the average morality of a small community, relatively isolated from centres of culture and unacquainted with any artistic tradition. Its morality is rigid and circumscribed; the details of each member's personal life are hedged about with constant annoyances, and everyone's life is open to the prying eyes of others who are unfailingly quick to judge, with or without evidence. Outsiders are looked upon askance; people of a different religion, race, or culture are viewed with suspicion and distrust; and anyone who does not subscribe to whatever moral code is dominant in the community is condemned or ostracized. No doubt these people are sincere—they are dreadfully sincere, deadly sincere. But sincerity without enlightenment can be as harmful to the achievement of good as intelligence without wisdom when that intelligence is possessed by political leaders playing with hydrogen bombs. Generally speaking, the people of a small community have not known the leavening influence of literature. Their morality is rigid, cramped, and arid. If these same people had been exposed from early youth to great masterpieces of literature and had learned through them to appreciate the tremendous diversity of human mores and beliefs held by other groups, with the same degree of sincerity that they themselves possess, they would be less likely to be as harsh, intolerant, and rigid as they are.

People are usually inclined to separate art and morality into two hermetically sealed compartments. They talk as if morality were already complete and self-sufficient without art, and that art, if it is to be tolerated at all, can grudgingly be permitted, provided that it conforms to the moral customs of the time and place of those judging it. But this view is surely to conceive the relation between art and morality in far too one-sided a manner. If art must take cognizance of morality, equally morality must take cognizance of art. Almost everything that is alive and imaginative about morality comes from the leavening influence of art.

Influence of ancient dramatists

To consider examples from ancient Greece alone, what would morality be today without the influence of the dramatists Aeschylus and Sophocles, without Socrates as described in Plato's dialogues, even without the historians Herodotus and Thucydides with their quiet humour, gentle prodding skepticism, and tolerance for other customs and views? It is through great works of art that the most vivid conceptions of various ways of life are obtained. What is it about other times and places that people most remember? Is it their political squabbles, their wars, their economic upheavals? These events are known in general to intelligent laymen and in detail to historians, but even then such events do not usually make much of a dent in peoples' personal lives in the way that art does. What is alive today about ancient Greece is its sculpture, its poetry, its epic and drama; what is alive today about the Elizabethan period, even more than the defeat of the Spanish Armada and the reign of Queen Elizabeth I, is its poetic drama, with its vivid characterizations and boundless energy. Other civilizations and cultures may be sources of facts and theories that enlighten modern understanding, but what enables contemporary man to share directly their feelings and attitudes toward life is not their politics nor even their religion but their art. Art alone is never out of date. Science is cumulative; even the science textbooks of ten years ago are now discarded as obsolete; the science of the ancient Greeks and the Elizabethans is studied today primarily for its historical value. But great art is never obsolete; it can still present to modern man its full impact, undiminished by time. Shakespeare will not be out of date as long as human beings continue to feel love, jealousy, and conflict in a troubled world. A biblical statement might be paraphrased and applied to past cultures: "By their arts shall ye know them." The artists whose works are now revered may have died unsung; most of them, even those who were appreciated during their lifetimes, were considered far less important than the latest naval victories or the accession of the current king; yet today these things have all passed into history, but art survives with undiminished vigour. The art of the past molds in countless ways the attitudes, responses, and dispositions of modern man's daily life. Most of what is perceptive and imaginative in morality owes its origin to art, and, when morality loses contact with the tradition of art, it becomes dead and sterile. Yet, in spite of this, some people tell us that art is merely the salve of morality, to soothe its stringency.

Already, in the preceding paragraph, mention has begun of arts other than literature. How, it could be asked, can they have any moral effects on those who view them or listen to them? Yet there are effects of these arts on the observer that, in a broad sense, are moral (as opposed to nonmoral) and that account for the attempts of many people to censor them.

Aristotle's theory of catharsis

Historically, the most famous supposition about the moral effect of art on its audience is Aristotle's theory of catharsis; Aristotle applied the theory to tragedy only, but many since his day have applied it to art in general. According to this view, art acts as an emotional cathartic and achieves a "purgation of the emotions." Certain emotions man would be better off without (Aristotle limited them to pity and fear, but they could easily be extended) are generated during the course of daily life. Art is the principal agency that should help to dispel these emotions. By observing works of art (witnessing a drama, listening to a powerful symphony, looking at certain works of sculpture or painting) the recipient can work off these emotions rather than let them fester inside him or take them out in unpleasant ways on his fellow men. Art siphons off these disturbing inner states rather than letting them grow rancid within man.

As it stands, this view is undoubtedly somewhat crude, especially in the light of modern psychology; and fault could be found in many respects with the Aristotelian doctrine of catharsis. Yet the experience of reading, viewing, or listening to a work of art does give a peculiar release, a feeling of freedom from inner turbulence. The mere act of plunging, for a few hours, into an entirely different world when attending a play or a concert is often enough to transform, however temporarily, the tone of people's daily lives. It is not merely that for a few hours they can forget their troubles—any form of entertainment, however worthless, might do this. It is not merely that art provides a break or interruption in the course of people's lives at the end of which they are exactly what they were before. It is that through the aesthetic process itself, in the very act of concentrating energies on an art object of great unity and complexity and depth, a kind of inner clarification is achieved that was not present before.

It is not true, therefore, that reading novels of crime and detection leads people to indulge in a life of crime; on the whole, those who read such novels are law-abiding people, and, if anything, the reading of such novels is a

substitute for aggressive activity (it is aggression vicariously experienced) rather than an incitement to it. Nor do works of art of a licentious nature usually incite people to rape or adultery; far from acting as incitements to action, they are safety valves against action by providing a kind of substitute gratification. It has been said, for example, that Shakespeare's *Antony and Cleopatra* is an immoral work because it celebrates the passionate surrender to an illicit love and the victory of this love over practical, political, and moral concerns. But is there any evidence that people who read this play will behave like the lovers in question because they read the play? On the contrary: it could be argued that reading the play has an instrumental value in that it presents another example of a complex moral situation, the perusal of which provides many avenues for moral reflection, and that the play also possesses the intrinsic value of acute characterization, dramatic power, and poetry whose imagery and intensity are among the most splendid in the English language. Again, it is said that American youth has been demoralized by such 20th-century U.S. writers as Ernest Hemingway and William Faulkner in that these writers set an example of bad behaviour. But to say that they are capable of demoralizing an entire generation is certainly to attribute to them too much moral power, especially over people who have never even heard of them. Even among those who do read serious literature, the effects are probably more beneficial than harmful: through books the horizons of such readers have been expanded to include other ways of life than they would have previously known.

The essential morality of the aesthetic experience

Quite apart from the ultimate effect of a work of art on a man's emotions, it would appear that the very act of experiencing the work may itself have a moral effect. If he is really concentrating on the details of a work of art and not just passively letting it play upon his senses, this effect—the heightening of his sensibilities and the refining of his capacities for perceptual discrimination—will make him more receptive to the world around him, thus raising the tone of his daily life and making his experience of the world richer than before.

Most of what passes for aesthetic appreciation does not begin to have this effect; but its failure is only because it is not aesthetic appreciation at all—it is a kind of tired reverie rather than an intense absorption in the aesthetic object. Most people, when they hear music, simply allow themselves to be inundated by the sheer flow of sound. Such people do not actively listen to the music and are not even aware of the most elementary kind of ebb and flow occurring within it; they only receive it passively, perhaps using it as a springboard for a private reverie or an emotional debauch of their own. Music has for them not an aesthetic effect but an anesthetic effect. It is not just hearing music that will have the required effect. The aesthetic experience, which involves nothing less than a total concentration on the perceptual details of the aesthetic object, is an experience that heightens consciousness, exercises man's capacity for perceptual awareness and discrimination, and helps him come alive to the sight and texture of the world around him. After a viewer has seen an exhibition of landscapes by Cézanne, the entire world may seem to him to have changed its structure and complexion: it may, indeed, take on the look of Cézanne's landscapes. And is not anything that increases awareness and subtlety of discernment and discrimination a potentially moral agent? Art provides the most intense, concentrated, and sharply focussed of the experiences available to man. Because of this, art can have an enormous influence on the tenor of a person's life, more influential no doubt than any particular system of morality. In its ability to do this, it has an effect on man's life that, in an extended sense at least, can surely be called moral. Morality transcends particular systems of morality; and art, by being for many persons the dominant influence in their lives, thus transcends them also.

BIBLIOGRAPHY

The definition of art: PAUL ZIFF, "The Task of Defining a Work of Art," *Phil. Rev.*, 62:58–78 (1953).

The creation of art: Anthologies of readings containing various accounts of creation in different art mediums are BREWSTER GHISELIN (ed.), *The Creative Process* (1952); and VINCENT TOMAS (ed.), *Creativity in the Arts* (1964). See also ROBERT GOLDWATER and MARCO TREVES (eds.), *Artists on Art, from the XIV to the XX Century* (1945); STUART GOLANN, "Psychological Study of Creativity," *Psychol. Bull.*, 60:548–565 (1963).

The mediums of the various arts: DEWITT PARKER, *The Analysis of Art* (1926), esp. ch. 3; SUSANNE K. LANGER, *Feeling and Form* (1953); and JOHN DEWEY, *Art As Experience* (1934), for imaginative discussions of the various art mediums.

The place of intention in art interpretation: See first the classic attack on intention by WILLIAM K. WIMSATT and MONROE C. BEARDSLEY, "The Intentional Fallacy," in *The Verbal Icon* (1954); for an opposed view see LESLIE A. FIEDLER, "Archetype and Signature: A Study of the Relationship Between Biography and Poetry," *Sewanee Rev.*, 60: 253–273 (1952); and HENRY D. AIKEN, "The Aesthetic Relevance of the Artist's Intentions," *J. Phil.*, 52:742–753 (1955).

Art as imitation (representation): Classic sources are ARISTOTLE, *Poetics*, and LONGINUS, *On the Sublime.* Recent works include WALTER ABELL, *Representation and Form* (1936), particularly in visual art; THEODORE M. GREENE, *The Arts and the Art of Criticism* (1940); and STEPHEN C. PEPPER, *The Work of Art* (1955). Perceptive articles include A.O. LOVEJOY, "Nature As Aesthetic Norm," *Mod. Language Notes*, 42:444–450 (1927); ARNOLD ISENBERG, "Perception, Meaning, and the Subject-Matter of Art," *J. Phil.*, 41:661–75 (1944); DAVID F. BOWERS, "The Role of Subject-Matter in Art," *J. Phil.*, 36:617–630 (1939).

Art as expression: A classic source is LEO TOLSTOY'S *What Is Art?* (1898; Aylmer Maude translation in "Oxford World's Classic Library," 1930, reprinted 1960); defense of forms of expression theory are ROBIN G. COLLINGWOOD, *The Principles of Art* (1938); and CURT J. DUCASSE, *The Philosophy of Art*, rev. ed. (1966). An anthology of readings on artistic expression is JOHN HOSPERS (ed.), *Artistic Expression* (1971). For critical accounts, see LOUIS A. REID, *Meaning in the Arts* (1969); D.W. GOTSHALK, "Aesthetic Expression," *J. Aesthetics Art Criticism*, 13:80–85 (1954).

Art as form: For a defense of art as form, see CLIVE BELL, *Art* (1914, reprinted 1958); ROGER FRY, *Transformations* (1927); EDUARD HANSLICK, *Vom Musikalisch-Schönen*, 7th ed. rev. (1885; Eng. trans., *The Beautiful in Music*, 1891, reprinted 1957). For mixed accounts, see DEWITT PARKER, "The Problem of Aesthetic Form," in *The Analysis of Art* (1929).

Art and truth: JOHN HOSPERS, "Implied Truths in Literature," *J. Aesthetics Art Criticism*, 19:37–46 (1960); and "Literature and Human Nature," *ibid.*, 17:45–57 (1958); ARNOLD ISENBERG, "The Problem of Belief," *ibid.*, 13:395–407 (1955).

Art, morality, and society: JOHN DEWEY, "Art and Civilization," in *Art As Experience* (1934); GEORGE SANTAYANA, *Reason in Art* (1910); SIDNEY ZINK, "The Moral Effect of Art," *Ethics*, 60:261–274 (1950).

Histories of the subject: ALBERT HOFSTADTER and RICHARD KUHNS (eds.), *Philosophies of Art and Beauty* (1964); ALEXANDER SESONSKE (ed.), *What Is Art?* (1965); MONROE C. BEARDSLEY, *Aesthetics from Classical Greece to the Present* (1966).

(Jo.Ho.)

Art Conservation and Restoration

Conservation, including maintenance, preservation (protection from damage or deterioration), and restoration, has become an increasingly important aspect of the work not only of museums but also of civic authorities and all those concerned with works of art, whether artists, collectors, or gallery goers. Technical advances of the 20th century have made possible new methods of safer cleaning and repairing of objects. Art restoration has become an important tool of research, and it enables the viewer to appreciate the original intention of the artist.

ARCHITECTURE

Development of architectural restoration. Conservation and restoration of architecture is a relatively modern phenomenon. The oldest buildings that have survived tend to be those that received religious veneration. When they were no longer venerated, they disappeared like other buildings. Even the famed ancient Egyptian Sphinx at Giza lay for centuries under the sand, and it was not until the early 19th century that the Forum of ancient Rome was uncovered and explored.

Medieval builders treated the work of their forebears with a healthy lack of awe. Every new Gothic chapel or chantry and, sometimes, every stage in the development of a single Gothic building followed the style of its own day. With the Renaissance in Europe grew a new respect for classical antiquity and a new interest in its architectural forms. By the end of the 18th century a knowledge of archaeology had become an accepted accomplishment of the educated man. Architectural design itself became a matter of "correctness." Old buildings everywhere began to be "restored" to the style of periods especially favoured. The French architect and writer E.-E. Viollet-le-Duc brilliantly restored the Sainte-Chapelle (1840–67) and the cathedral of Notre-Dame de Paris (1845–64). The ancient walls of Carcassonne in France and of Windsor Castle in England were not only repaired but also largely rebuilt in the 19th century.

With the spread of the Industrial Revolution, the labour of hands became more costly, and the value of craftsmanship gained a new significance. Old buildings began to command a new respect, and the English art critic John Ruskin (1819–1900) was even able to assert that "the greatest glory of a building is its age." In 1877 the pioneers of the conservation movement, led by the English artist and writer William Morris (1834–96), founded the Society for the Protection of Ancient Buildings (SPAB). Nicknamed Anti-Scrape, the society vehemently opposed the indiscriminate refacing of old stonework and the "conjectural restorations" still so fashionable, such as the new west front of St. Albans Cathedral in England (1880–83). The movement gathered force, and in the 20th century groups throughout the world now devote their efforts to architectural conservation.

An added local impetus has been given by national pride; postwar reconstruction became, for countries like Poland, the symbol of national resurgence. Almost every civilized country is increasingly conscious of its heritage of ancient buildings, while cultural bodies such as the United Nations Educational, Scientific, and Cultural Organization have lent to the conservation movement a powerful international impetus.

Effects of economic and social change. Architecture is a sensitive index of social change. The economic climate and social preoccupations of each age have combined to generate its own architecture and its own towns. Almost every decade of new building displays its own peculiar characteristics and modifies by constant adaptation the buildings and towns of yesterday. But the urban environment is society's investment in its future, and the cycle of renewal is continuous, if often slow. Thus, the problems of building maintenance and renewal are complicated by long-term economic and social change.

Perhaps the most marked trends are still those that brought about the conservation movement itself. First is the accelerated pace of physical growth. Old buildings have become not only relatively rarer but often virtually irreplaceable in terms of labour and craftsmanship, as well as in their use of building materials that are rare in more recent structures. In many cases, old buildings give to a locality much of its special character and identity, as, for example, in any English country village where thatched roofs still predominate. Another and rarer asset is a sheer and intrinsic merit of architectural form. And alongside all these is the tangible evidence that any old building provides for its community of a kind of social and environmental continuity—a reassuring reference point in a constantly changing world.

The factor of land value

Under the increasing pressure of population, the value of land has climbed steeply, with some curious effects on old buildings. Increased demand brings increased values and, at first, better prospects of repair and maintenance. But as values rise higher, the older building must also justify itself in terms of economic efficiency. All over the world, the town houses of the 19th century and earlier serve with varied grace in the 20th century as centres of modern industry and commerce. Their fabric is subjected to new strains, and their room shapes and capacity may become incompatible with new and changed demands. There comes a point at which the old building on a valuable town site can compete no longer with redevelopment. Then it is quickly overtaken, and financial subsidy is powerless. The old building in a deteriorating neighbourhood is at the same time likely to be in no better a situation. Its maintenance may become no longer worthwhile, condemning it to early death by neglect. The destructive effects of both over- and undervalue are clearly displayed side by side in a fine Georgian city like Dublin or in once-distinguished neighbourhoods like Bloomsbury in London or the Marais in Paris.

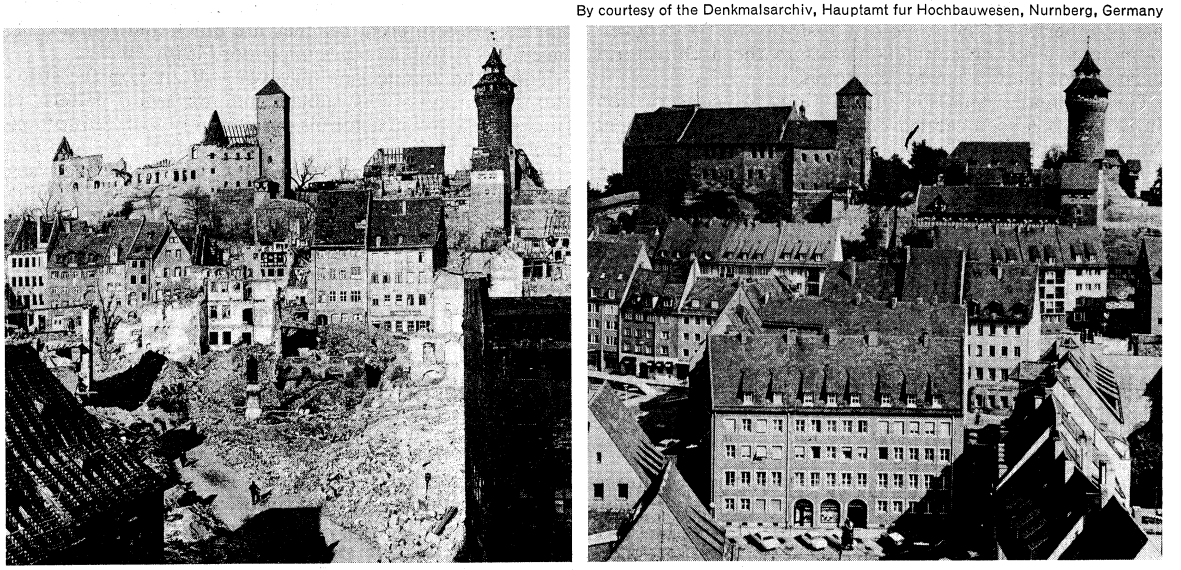
The most successful neighbourhood conservation occurs where values have been held in pace with the architectural capacity of a community, as at Bath in England, or in the Georgetown section of Washington, D.C.

The factor of increased mobility

Another social change is the rapid increase in mobility. The automobile brought better roads and an incentive to use them. Old city centres, after centuries of essentially domestic life, begin to be abandoned in favour of ring upon ring of suburbs. This peripheral accretion of cities is allied with their central decay as communities. As a universal result, the twice-daily thrombosis of the highways urges on a constant process of road widening, in which many an intervening historic area has been completely eroded away.

New possibilities of travel are increasingly open to

By courtesy of the Denkmalsarchiv, Hauptamt fur Hochbauwesen, Nurnberg, Germany



Albrecht Dürerplatz, Nürnberg, Germany, after bombing in 1945 (left) and after reconstruction (right).

many. Cultural tourism can bring an antidote to the destructive forces, so that in countries such as Italy and Spain architectural conservation attracts quite a new order of investment and national energy.

Role of the law. In all conservation, the first effective step is to decide and define what is worthy of protection. For most countries this has involved a systematic process of inventory and survey. In Britain, for example, the Royal Commission on Historic Monuments (RCHM) was set up in 1908, and the Civic Amenities Act of 1967 was promoted by an independent body, the Civic Trust. The act required local planning authorities to define special areas for "conservation and enhancement." In France, the Commission des Secteurs Sauvegardés was set up in 1962 under André Malraux, minister for cultural affairs, to pursue an active program for public protection of historic areas. In the United States, the Historic American Buildings Survey was designed to assemble a national archive of historic American architecture.

Criteria for conservation

Criteria for conservation are rarely well defined. Architectural merit clearly must rank highly—especially in the case of any building which authentically exemplifies its period. Historical associations, such as the birthplace of a famous person, are less easily rated. One pernicious effect of all selection is the way in which it is the most outstanding example of any period, rather than the truly typical, that in the end remains to represent it. Another is that defects as well as merits may be kept warm under the same blanket. This is particularly so in the larger groups of buildings that are coming to be recognized as worthy of conservation.

Once defined, a building's next defense is in specific legal powers for its protection. These may be of varied degree and effectiveness. The most obvious form of legislation is the restriction against demolition. A higher degree of legal sophistication occurs in powers for the annexation of property and its maintenance by the state. Covenanted rights and restrictions are a variant of this principle. Next in the scale of effectiveness comes positive encouragement to owners by means of grants, bringing a public share and interest in the actual work of repair. In this way, actual legal rights over private property may be confined to a minimum while finance is encouraged from private pockets. Probably the most effective ultimate defense has been a selective protection, exercised as a regular part of everyday town- and country-planning control.

Negative legislation itself varies in degree. In Italy it is possible to insist upon the return even of certain pictures or chattels illegally dispersed from a building where these are adjudged to be of sufficient national importance. But negative powers are inherently weak. They convey no control over the philistine or intransigent owner and, at best, can only slow down neglect and demolition, deliberate or otherwise. The value of a historic site for redevelopment is a strong counter-magnet to the architectural heritage for many a private conscience.

The national acquisition of buildings for conservation in Britain has been carried out chiefly under the Ancient Monuments Consolidation and Amendment Act of 1913, by which suitable unoccupied properties can be "taken into guardianship." A much more rigorous application of the principle is sometimes possible in the United States, whereby the owners of whole groups of buildings held to be of sufficient distinction can in fact be legally dispossessed. These erstwhile owners may then be allowed to remain in residence on condition of the repair and rehabilitation of their buildings to a specified standard. In this way, whole areas of buildings, such as Society Hill in Philadelphia, have been taken over, concentrated redevelopment by high-rise apartments being permitted in selected inner locations but the buildings with frontage being restored in period styles.

Williamsburg, Virginia

The most exhaustive of all restoration projects is in the United States, at Williamsburg, Virginia. This small town, the colonial capital of Virginia from 1699 to 1780, has attracted the most expensive restoration program ever. Commenced in 1926, the project is dedicated to the purpose "that the future may learn from the past." By 1970 over $70,000,000 had been spent on the careful and scholarly restoration of buildings. Environmental management is of a high order. Tourist automobile traffic is excluded from the restored area in season, when a free bus service is provided. The emphasis is frankly educational. The enterprise not only owns its buildings but also staffs them, its employees wearing correct period costume.

Perhaps the most dramatic rescue operation has been in Egypt. There the ancient temples (*c.* 1250 BC) of Abu Simbel were threatened with destruction by the rising waters of the new Aswan High Dam. They were sawed into giant blocks and successfully reassembled 200 feet (60 metres) above the original site. This act of preservation was the result of intense and international negotiation and expertise.

Another variant on public ownership may be found in acquisition by a private body, such as the National Trust in Britain. Founded in 1895, this property-owning body opens to the public over 200 of its properties. The trust receives no direct government subsidy and relies upon careful economic management, although certain legal preferences operate in its favour. In the United States the American National Trust operates in the same manner.

Among bodies devoted to grant aid, the Historic Buildings Council (HBC) for England was set up in 1953 and disburses grants within a small annual budget, largely to help building owners penalized by heavy estate duties. These grants are administered to encourage owners to take a pride in their own buildings.

A pioneer architectural-conservation program has been established by the Faculty of Architecture of Rome University. Of six months' duration, the course is aimed at providing specialist training in conservation for architects of all nationalities. At national level in various countries, several comparable courses are becoming available, meeting a demonstrable need for suitably qualified and experienced architects.

Techniques of building conservation. The first requisite in any conservation project is an adequate historical survey. The conservator first researches the life history of the building and its development through the ages. Every building has its own biography. The Parthenon in Athens was originally built from 447 to 432 BC as a temple and subsequently served as a Christian church, a mosque, and a powder magazine before it became one of the world's greatest attractions for the tourist and art lover. A knowledge of the whole life of a building brings an essential understanding of its features and its problems.

Next, the conservator will need a thorough, measured survey. Generally, this is prepared by hand, with tape and rod and level. Modern measuring techniques, including photogrammetry and, sometimes, stereophotogrammetry, are also used, for they are quick and remarkably accurate. A full photographic record is invaluable.

Assessing the structure's soundness

Third, the architect or surveyor will analyze the structural stability of his subject and its living pattern of movement. No structure is permanently still. Subsoil expands and shrinks, thrust moves against thrust, and materials move with heat and wind. Complicated exercises, like bell ringing, have an even greater effect on a building's stability. Clay soil is the worst: the building protects the ground underneath but not around; and, with every downpour, a wall of accumulated clay may vary the lean of the building. Many ancient buildings had piled foundations—Winchester Cathedral's east end was built on oak piles that rotted after centuries, once the subsoil drainage was improved; and the only answer was to send down a diver to underpin. Framed structures can move a great deal. The skeleton of a timber-framed medieval house can be incredibly crooked without losing strength, if it is well triangulated and its joints are made sound. A wall is theoretically safe until it leans far enough to develop tension on one side, yet even then it may be stiffened by structural cross-walls. Generally, the old, evenly spread load will be stable, and any new point load or thrust will be suspect. Where it is necessary, the surveyor checks his observations over a period; *e.g.*, by measure-

ment with plumb lines or by marking devices placed across a crack.

The survey will not only analyze but also clearly convey its observations, defining remedies and their relative urgency and relating all to the function of the building and the available budget.

The next job in architectural conservation is to stabilize and consolidate the structure. Ideally, this is best done by restraining, or tying, the point of active thrust and, perhaps, by replacing, splinting, or in some way giving fresh heart to a failing or defective member. Adding heavy weights such as buttresses often does more harm than good. A load can frequently be spread more widely or more evenly. A structure can, in effect, be corseted by inserting (for example, around a tower) a continuous beam or ring of concrete. This can be done even in delicate masonry and, as in underpinning, by removing alternate sections of a wall, threading in reinforcement, and casting one and then another set of concrete stitches, until the whole becomes one strengthening beam. Or all that is necessary may be a single tie bar, a metal rod inserted along a direct line of thrust or weakness to join the structural elements in need of support.

Damp in buildings

After structural movement, the next serious adversary in building conservation is damp. Not only of itself but also allied with almost every other trouble, damp accelerates decay. Weather may be penetrating through whole surfaces, such as porous brickwork, or finding its way through cracks or defects in the roofing. Especially vulnerable are gutters or any part of the rainwater-collecting system. Wet weakens walling, rots timbers, and spoils finishes. The eradication of damp may mean patching or renewing a roof. Frequently it can be cured by inserting a continuous moisture barrier, perhaps even in a modern material like stout polyethylene. In this case, special care is needed to avoid future damage by concentrating more trouble at any possible defect. Modern techniques of waterproofing wet walls include the insertion of high-capillary tubes, designed to draw the moisture to themselves and to expel it, and also the injection of latex and similar water-repellent solutions, deposited in the heart of the walling. A new technique is that of electro-osmosis, designed to reverse the moisture path from ground to walling. The traditional ditch, or dry area, drained if necessary, disposes of the water before it reaches the wall. Double or cavity walls, with air between them, are another defense against damp.

Again, dampness compounds decay, and the first attention should be to such structural devices as sloping wall coverings, such as copings, which are designed to keep walling dry. Both in stonework and in brickwork, much harm can be caused by damp, especially when allied with an overly hard mortar jointing. This traps moisture along the lines of the joints, bringing any harmful salts to the surface, where they crystallize and damage the facing. Mortar jointing should always be softer than the brick or stone of a wall.

Unsound walling

Much decay is the result of poor construction. Defects are almost always accelerated by the simple contravention of good building practice. In walling, a typical cause of structural instability is a double-skin construction with rough rubble between in which, by uneven loading, one skin has been caused to bulge and to release loose material in the core of the wall. Once on the move, this rapidly gains momentum as a live wedge, forcing apart its two faces. The conservator generally inserts temporary support, then distributes the load more evenly and rebuilds the affected area. In some cases, through washing out loose material and then, either by gravity or at high pressure, grouting up the unseen cavities, it may be possible to restore a wall without disturbing the facing stonework.

The roof is a building's first defense. It must be impervious and collect water clear of a building. Roof finishes are commonly either of unit materials such as tiles, slates, or stone, or of boarding covered in sheet metal, preferably lead. The failure of unit materials is usually caused by decay of the materials that hold them in place. Iron nails are especially destructive and are usually replaced by nonferrous materials, such as copper. The battens that carry the tiles or slates have a longer lifespan but also need periodic renewal. Leadwork failure is usually the result of sheer age. This material has a very long life, but, if it has been used in sheets of excessive size, there is a tendency to buckle and creep as a result of expansion—especially in sunshine. Leadwork can readily be recast. It needs firm support and elements to fix it in place. Temporary repairs are executed by lead burning, in which a patch may be actually welded with the original lead, rather than by soldering, which tends to crack away.

The chief enemies of timber are the natural predators of the forest—fungi and wood-boring insects. In ancient buildings, the former are especially damaging.

The most voracious fungus that attacks building timbers is dry rot (*Merulius lacrymans*). This can spread along infected wood to sound timber, carrying its own moisture supply. It extracts cellulose, which forms the chief part of plant cells, and leaves behind a tindery and useless shell. Stagnant air and warmth accelerate its spread. Eradication must be thorough, or the trouble will rapidly re-establish itself. Modern fungicides are highly effective.

Wood-boring insects include the furniture and deathwatch beetles. From eggs laid in cracks, the larvae tunnel into timber and damage it before emerging as beetles to lay more eggs.

The deathwatch beetle inhabits mostly the outer sapwood of oak, preferably when wet or softened by rot. The furniture beetle lives mostly in deal, especially when sappy or damp. Both can be eradicated with modern pesticides.

The conservator lastly tests all services, especially electrical wiring, with its risk of fire; gas lines, with their danger of seepage and explosion; and plumbing, with its danger of leaks. These services are frequently redesigned and are often simplified as well as improved. Lightning conductors and fire-fighting equipment are an important part of the protection of any ancient building and need regular servicing. (D.W.I.)

PAINTINGS

The conservator of paintings aims above all at "true conservation," the preservation of the objects in conditions that, as far as possible, will arrest material decay and delay as long as possible the moment when restoration is needed. The correct choice of conditions of display and storage is, therefore, of the first importance. Ideally, each type of painting requires its own special conditions for maximum safety, depending on the original technique and materials used to compose it. Broadly speaking, most paintings can be divided into (1) easel paintings, on either canvas or a solid support, usually wood; (2) wall, or mural, paintings, executed in a variety of techniques; and (3) painting on paper and ivory.

Easel paintings. More or less portable paintings on canvas or panel are called easel paintings. Basically, they consist of the support (the canvas or panel); the ground, ordinarily a white or tinted pigment or inert substance mixed with either glue or oil; the paint layer itself, which may be complex in structure; and, finally, the surface coating, usually a varnish, to protect the paint and modify its appearance aesthetically. These four layers have many variants but must be constantly borne in mind when considering the problems of conservation.

Paintings on wood. Wood-panel supports were used almost universally in European art before about 1450, when canvas began to gain ground. Wood has the disadvantage of swelling and shrinking across the grain with variations in the relative humidity of the atmosphere. In northern temperate climates, variations in humidity can be considerable. In England, for example, the seasonal variation in a museum that is centrally heated in the winter can be from 25 percent in midwinter to 90 percent in summer. Although paint has a certain elasticity, it cannot usually take up much movement and generally cracks in a network referred to as craquelure. In continental land masses, such as the United States, the average relative humidity in dry zones may be consistently low, so that European paintings with wooden supports

Effects of humidity

air-seasoned to a higher humidity may suffer considerably. In both Europe and America, the effect of an unsuitable background of low or changing relative humidity and the restraining effect of the paint layer often produces a permanent bowing of the panel, which is convex at the front surface. To counteract both the shrinkage and the bowing (especially the latter), restorers in the past placed wooden strips called battens or more complex structures across the back of the panel as constraints. This often led to severe distortion of the front surface and cracking of the whole panel in lines along the wood grain. Extensive damage to the paint sometimes occurs, and drastic restoration is needed. In terms of preservation, the ideal solution is a form of air conditioning in which the relative humidity is maintained as nearly constant as possible at what is generally agreed to be the most reasonable level; *i.e.*, about 55 percent. The apparatus for achieving this is expensive and can be considered only by wealthy museums. In smaller museums and private houses, humidifiers either incorporated into central heating installations, if possible, or in the form of small portable machines are often used during the winter months.

Use of secondary supports

When warping and cracking have already occurred or when the latter seems likely as a result of the mistaken application of secondary supports, such as cross-battens, expert restoration treatment is required. In principle, this consists of removing the cross-battens and applying a reinforcement to the back that imposes a uniform but gentler constraint over the whole surface. It is normal in the 20th century to accept as inevitable some permanent convex curvature. The adhesives used and the composition of the new secondary support take many forms. One consists in backing the panel with strips of a very light, open-textured wood (balsa), using as a cement a mixture of beeswax, a natural resin, such as dammar or colophony, and an inert filler. This thermoplastic cement, applied as a hot, creamy liquid, solidifies without contraction. The epoxy group of resins, which harden without contraction, have also been used and have the advantage of not requiring heat. The material and cement used are chosen according to the nature of the original panel. Some restorers reduce the strength of the original panel, before applying the secondary support, by reducing its thickness. This practice is not universally approved. Occasionally, when the panel is badly worm eaten or severely cracked, it has to be removed from the paint and ground altogether in the process known as transfer. This is accomplished by pasting a substantial support of paper and, possibly, canvas to the front surface of the paint and then gently gouging away the wood on the back. An entirely new, inert support of balsa wood or compressed board is then cemented to the back surface and the front facing removed.

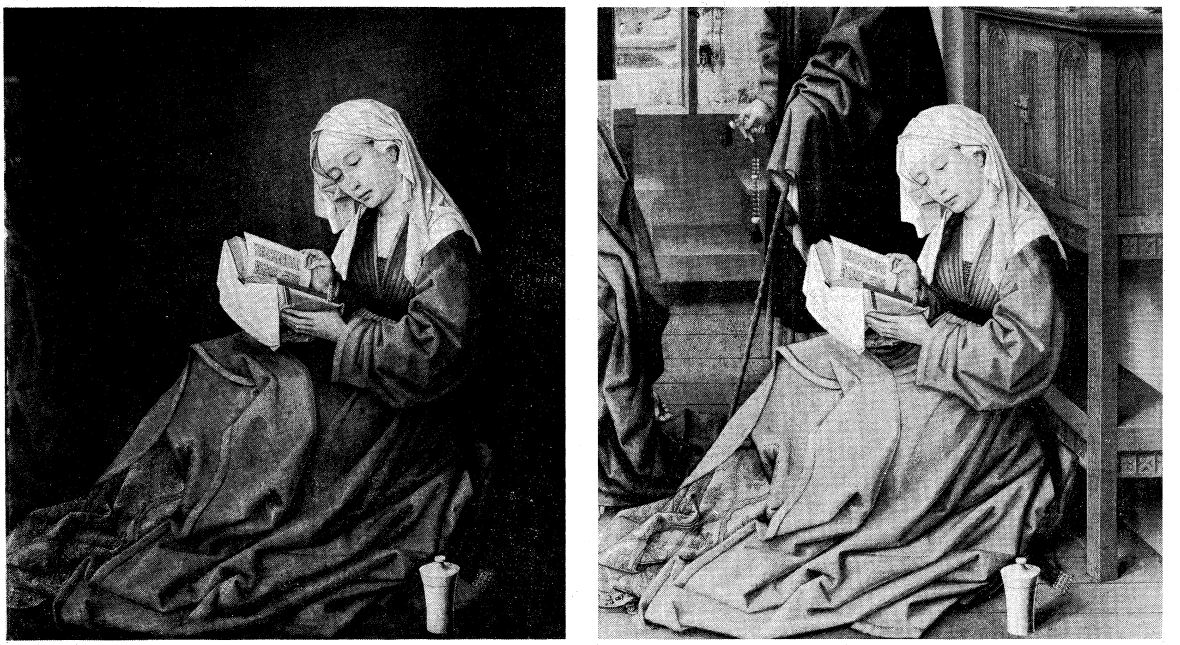
Paintings on canvas. A canvas support expands and contracts with variations in relative humidity, but the effect is not as drastic as with wood. Canvas, however, will deteriorate with age and acid conditions. In many cases, parts of the paint and ground will lift from the surface, a condition described as flaking, blistering, or scaling. In the case of paintings on canvas, the process of transfer is almost never performed. Instead, the canvas is reinforced at the back by attaching a new canvas to the old. This lining process (almost always referred to as relining) can be done in two ways. The traditional method consists of ironing the new canvas to the old, using as adhesive a warm, fluid mixture of animal glue and a farinaceous paste, sometimes emulsified with a resinous plasticizer, such as Venice turpentine. The second, formerly known as the Dutch method, uses as adhesive the wax–resin mixture mentioned above, at a temperature of about 70°–80° C. It is ironed as it cools and provides a thermoplastic bond. The advantage of this method is that it penetrates the front surface and fixes loose paint to the canvas, whereas, with the glue–paste method, fixing the paint would require a second operation. Also, the permanence and moisture impermeability of the wax mixture helps to protect the painting and the canvas from humidity changes and atmospheric impurities. A notable advance in the wax–resin method since 1950 is becoming universally adopted. This consists of employing the pressure of the atmosphere instead of a heavy iron during solidification. This is achieved by carrying out the process on an electrically heated platen (the "hot table"). The painting and its new canvas, treated with layers of the adhesive, are placed on the smooth metal surface of the table and covered with a latex sheet, after which the space between is evacuated with a pump through holes in the table. The cooling process is accelerated with fans. A notable advantage of this method is that the impasto (*i.e.*, the raised brushstroking of the paint) is not flattened, as used often to occur when heavy irons were used.

Advantage of the Dutch method

The ground (*i.e.*, the inert paint layer covering the support below the painting itself) can ordinarily be regarded for conservation purposes as part of the painting layers. Occasionally, when the ground is composed of glue and an inert substance such as whiting or gypsum, the glue may deteriorate and the ground lose its adhesion to either the support or the paint layers. In extreme cases, with wood-panel supports, a complete transfer is required, in which not only the support but also the ground must be removed. The restorer has a brief opportunity of seeing the painting or at least its lowest layers in reverse before applying a new ground and support.

The paint layers themselves are subject to a number of maladies as a result of natural decay, faulty original technique, unsuitable conditions, ill treatment, and improper earlier restorations. It must be remembered that, whereas housepaint usually has to be renewed every few years, the paint of easel paintings is required to survive indefinitely and may be already 600 years old. The most prevalent defect, as mentioned above, is flaking, or scaling, where, in local areas of paint, small particles become partly or wholly detached from the support. If the loss is not total, the paint can be secured, according to circumstances, with either a dilute gelatin adhesive or the wax–resin adhesive. The paint is usually pressed firmly into place with an electrically heated spatula. For easel paintings the binder for the coloured pigments is usually either egg yolk, oil, or, occasionally, glue. The first type of method, called egg-tempera painting, was universal before the mid-15th century, when oil painting began to be used increasingly. The condition of egg-tempera paintings where damage has not been caused by deterioration of support or ground is usually good. The condition of oil paintings is often less satisfactory. Sometimes the original technique of the artist is at fault, and this becomes increasingly so from the 18th century onward. Too much oil may have been used, leading to ineradicable wrinkling, or superimposed layers may have dried at different rates, producing a wide craquelure as a result of unequal shrinkage. An enhanced version of the latter occurred increasingly, as the 19th century progressed, by the use of a brown pigment called bitumen. Bitumenous paints never dried completely, producing a surface effect similar to crocodile skin. These defects cannot be cured and can be visually ameliorated only by judicious retouching. The most notable defect arising from poor conservation is the fading or changing of the pigments by excessive light. Although this is more evident with thin-layer paintings, such as watercolours, it is also visible in oil paintings. The palette of the earlier painters was, in general, stable to light; however, some pigments, notably the lakes, which consisted of vegetable dyestuffs mordanted onto translucent inert materials, often faded easily. A transparent green, copper resinate, much used from the 15th to the 18th century, became a deep chocolate brown after prolonged exposure to light. After the discovery of synthetic dyestuffs in 1856, a further series of pigments was created, some of which were later discovered to fade rapidly. Unfortunately, it is impossible to restore the original colour, and in this case conservation, in its true sense of arresting decay, is important; *i.e.*, to limit the light to the lowest level consistent with adequate viewing—in practice about 15 lumens per square foot (15 footcandles; 150 lux).

Problems in the original paint



"The Magdalen Reading," oil painting by Rogier van der Weyden (1399/1400–64). In the National Gallery, London. Cleaning and restoration in 1956 (right) revealed that the painting was a fragment from an altarpiece, not a complete painting as it had appeared (left).
By courtesy of the trustees of the National Gallery, London

Almost every painting of any degree of antiquity will have losses and damages, and a painting of earlier than the 19th century in perfect condition will usually be an object of special interest. Before a more conscientious approach to restoration became general in the mid-20th century, areas that had a number of small losses were often—indeed, generally—entirely repainted. It was considered normal in any case to repaint not only losses or gravely damaged areas but also a wide area of surrounding original paint, often with materials that have visibly darkened or faded with time. Large areas with significant detail missing were repainted inventively in what was supposed to be the style of the original artist. It is customary nowadays to repaint only the actual missing areas, matching carefully the artist's technique and paint texture. In some cases, as with skies in which areas have been worn away by injudicious cleaning, very little repainting is done at all. Some restorers adopt various methods of inpainting in which the surrounding original paint is not imitated. The inpainting is done in a colour or with a texture that is intended to eliminate the shock of seeing a completely lost area without actually deceiving the observer. The aim in inpainting is always to use pigments and mediums that do not change with time. Egg tempera has been a preferred medium, though it is difficult to use, and various stable, modern resins are employed in place of or in addition to tempera. Exact imitation of the original entails close study of the painter's technique, especially with the earlier multilayer methods, since the successive layers, being partly translucent, contribute to the final visual effect. Minute details of texture, brushstrokes, and craquelure must also be simulated.

Inpainting

This work of inpainting ordinarily has to be done after the top layer, or the varnish, has been removed. Because all varnishes before about 1930 and many since have undergone changes in colour and transparency, they partially obscure the appearance of the original paint and, therefore, must be taken off.

While the use of varnish was partly to protect the paint from accidental damage and abrasion, its main purpose was to improve the appearance. Oil paint, over the course of time, changes chemically by oxidation and polymerization and becomes harder and more brittle. A remarkable range of injurious cleaning agents, including sand and caustic alkali, is known to have been used for cleaning the surface of the paint in the past. As a consequence, some of the hardened oil medium is lost, and the painting becomes matte, or without lustre, and lifeless. A varnish, visually at least, revives it, and if a varnish itself becomes matte a further coat of varnish revives the former varnish coat, and so on. Unfortunately, the varnishes used consisted of hard resins, such as copal, or, more often, soft resins, such as mastic and dammar. These become yellow, brittle, and slightly opaque and also less soluble in harmless solvents, such as turpentine. Occasionally, as in London's National Gallery in the mid-19th century, lead driers were added to the varnish to quench the bloom, a blue haze that covered the varnish as a consequence of the prevalent coal smoke combined with a high and variable humidity. This varnish, known as the "Gallery Varnish," yellowed even more rapidly than the resins alone. The removal of these disfiguring natural-resin varnishes from paintings is an operation that must be carried out with great skill to avoid damaging the original paint. The work of restorers in removing varnish (usually described as cleaning) has been the subject of occasional vehement public criticism from 1850 onward. The criticisms centre on two points: first, that the solvents used are likely to damage the paint, which is said to contain proportions of soluble material, even when soluble resinous additives are not present (as they sometimes were in the 18th and 19th centuries); secondly, that some earlier artists themselves used a toned varnish to unify their colours and that, if this still remains, varnish removal is likely to produce an overcleaned effect that was not intended.

Varnishing

The removal of the varnish is usually done in the 20th century by solution in a solvent mixture of which the active ingredient is often one of the lower alcohols (ethanol or isopropyl alcohol) or acetone. Careful, minute tests are always made on representative areas, and the solvent is sparingly applied on a cotton-wool swab. For revarnishing, the natural resins have been almost universally abandoned in the mid-20th century in favour of a limited group of synthetics whose common characteristics are stability to light, oxygen, and moisture and, in general, a chemical stability that will permit eventual removal by safe solvents. These synthetic varnishes are also optically effective, adhere satisfactorily, and are adequately protective. The first synthetic resin varnish with these characteristics was polyvinyl acetate, which was introduced in the United States in 1930 and has been used continuously since. To this may be added a group of resins of the polycyclohexanone type (almost but not completely

stable to light) and an acrylic resin that is finding increasing use.

Wall paintings. From the point of view of conservation, the different types of wall painting have a number of features in common, though the techniques of restoration inevitably differ in detail.

Among the wall painting techniques is buon fresco, or true fresco, in which pigments mixed with water are painted onto a freshly prepared layer of damp lime plaster. Fresco secco is a method, often used in conjunction with buon fresco, in which a mixture of pigment and egg tempera is painted onto the dry plaster or is used as a retouching or enhancement of a dried buon fresco painting. Wall paintings are also executed with pigments mixed in oil applied either to a prepared dry plaster wall or on canvas, which is then fixed to the wall.

As far as pure conservation is concerned, there are two outstanding factors. The first, which applies to all methods of wall painting and especially to aqueous, or water-based mediums, is the exclusion of damp. This can attack the painting from several sources. One source is damp rising through the walls of a building; this first affects the bottom of the wall painting and then spreads upward. This is prevented by inserting a metallic or resinous damp course. New damp courses in old buildings are often prohibitively expensive, in which case a possible amelioration is to dig out exterior soil to a depth of at least six inches below the interior floor. The second source of damp is from the outside wall. It is important at least to avoid treating the painting with a water-impermeable material, such as wax or silicates, so that the damp can penetrate freely without meeting a barrier at the inner surface. The third source is condensation on the inner surface, which is particularly prevalent in churches that are heated only on weekends. More continuous and uniform heat is the solution, provided that the air is not dried out so rapidly that efflorescence, the formation of a powdery surface, occurs. The fourth and most easily remedied source, though often neglected, is from leaking roofs and clogged drainpipes.

Deterioration of lime

The second important hazard is more insidious. It affects solely those murals painted on lime mortar, which inevitably, by the action of air, becomes calcium carbonate. Since 1900, with the increasing use of motor vehicles and of fuel-burning industries, the percentage of sulfur dioxide in the atmosphere has greatly increased. In the presence of moisture the calcium carbonate is changed to calcium sulfate, whose volume is almost twice that of the original carbonate of the mural. As a result, disintegration in some areas of a mural can be rapid. In Italy this sort of disintegration has greatly increased and has made necessary the development of drastic though highly expert methods of transfer of frescoes from the original walls. These range from the method of *strappo* to that of *stacco*. While in practice they are not always clearly distinguishable, *strappo*, the more usual method, consists in gluing canvas firmly to the surface of the fresco, followed by pulling and easing away with long spatulas a thin layer of the plaster that contains the pigment particles of the fresco. The bond between the facing and the fresco must be stronger than the internal cohesion of the plaster. Excess plaster is removed, revealing the fresco in reverse. This is then fixed to a rigid support with synthetic resins, using inert substances mixed with resins as an intermediate layer to stimulate optically the original underlying plaster. In the *stacco* method, a thicker layer of plaster is removed with the fresco and is smoothed flat on its back surface before sticking the rigid composite layer to a board. Where possible, consolidation without detachment is performed. The removal of previous repaintings and overlying whitewash is often the most tedious part of the work.

In humid, temperate climates, such as England's, limewater is usually used as a consolidant. Earlier consolidations, often of wax or natural resins, are not only difficult to remove but also have frequently accelerated deterioration. In dry parts of the world, synthetic resins such as polyvinyl acetate have been used with success as consolidants.

Paintings on paper and ivory. Environmental conservation for these objects, which are ordinarily painted in an aqueous medium, consists in maintaining a stable relative humidity in the region of 50–60 percent. At lower humidities, both paper and ivory (the latter often used as a thin layer for portrait miniatures) tend to shrink, and the former becomes more brittle. The thin ivory of miniatures, which often tends to crack, may crack along the grain if constrained under conditions of varying humidity. At higher humidities, there is a possibility, especially when ventilation is poor, of mold growth, which can occur above about 68 percent relative humidity. Watercolour paintings are particularly vulnerable to light, which ideally should not exceed 5–10 lumens per square foot (5–10 footcandles, 50–100 lux). Some pigments fade rapidly, whereas others do not alter, and there is inevitably not only a loss of colour but also a distortion of the artist's intention. It should be noted that a warm light, as from an ordinary incandescent lamp, is less damaging in general than an equal amount of daylight or light from a fluorescent lamp. Daylight should be particularly avoided.

Restoration of paintings on paper has many detailed variations. After removal of the material on which the paintings are mounted, local brown stains, usually known as foxing, are sometimes reduced. The painting is freshened by washing gently in water with or without a little neutral detergent (which should not be of the household variety). Often a watercolour painting will resist washing without loss of colour, but it is generally advisable merely to damp the back before proceeding to reduce the stains locally with an oxidizing bleach. A mild form of bleaching agent known as chloramine-T is sometimes used. Other, stronger oxidizing bleaches can be used subsequently, but there are various disadvantages. Stains other than the characteristic foxing must be identified and the specific solvents used. Japanese prints may be treated similarly, though with even more care, since some colours (notably, a range of mauves) must never be damped.

Portrait miniatures on ivory require expert treatment, and it is even possible to damage them irrevocably in removing them from their lockets.

The restoration and conservation of prints are dealt with in the article PRINTMAKING. (N.S.B.)

SCULPTURE

Efforts have been made for some time to develop a means of preserving valuable marble and limestone sculpture. It was only in the mid-20th century, however, that the seriousness of the problem became universally recognized. Much sculpture exhibited out-of-doors suffers from the effect of air pollution. With Italian sculpture from the 14th century onward, the problem is acute. Such work often suffers from its age or what could well be called stone fatigue. The two main elements detrimental to stone sculpture are agreed to be the high concentration of sulfur dioxide usual in modern industrial environments and frost, which obviously has been a constant historic factor. When both detrimental effects are concurrent, a relatively rapid destructive cycle is activated. The degeneration of the stone surface by the repeated expansion of freezing water and the fast thawing in the morning sun opens pores of the stone to the industrial gases, which quickly change the chemical nature of calcium carbonate stones to calcium sulfate. This brings about an increase of the stone's volume, resulting in a shedding of the degenerate layer.

Cleaning sculpture

Attempts to clean sculpture in the past have often proved ineffective or positively harmful. Before cleaning can be undertaken, the nature of the dirt must be determined. A marble statue may be affected by incorrect cleaning, deliberate toning (oiling or waxing; the latter includes the result of handling), and sulfation, the reaction of sulfur oxides in the air with the calcium in the stone. Later, dirt must be dealt with before removal of the film caused by the sulfation is possible. Most superficial dirt can be removed by applying a solvent such as methylene chloride or toluene on a patch at a time and

removing the dissolved methylene dirt with a series of clean swabs. This treatment leaves the marble clean except for the film of sulfation, which can vary in colour from cream to black. Attempts to remove this deposit with soap and water are unlikely to be successful, as tapwater does not dissolve the calcium sulfate, and the soap reacts with the calcium carbonate to form still another waterproof film. The problem is easily solved by using deionized water (ultrapure water), which dissolves the sulfation film. The deionized water is suspended in position in a mudlike pack for some 12 hours, the suspender being magnesium silicate. At the end of this period, the now dry mud is brushed off, and the marble is rinsed once more with deionized water. The remaining problem is to prevent a repetition of the decay.

One method is to use a solution of two parts of cosmoloid wax and one part of Ketone N resin made up in white spirit to form a thin cream. This is massaged into the surface of the marble, leaving no excess. This coating inhibits the action of sulfur dioxide and sulfation. The whole work is then brushed over with talc, which eliminates the dust-attracting property of the wax–resin coating. Subsequent cleaning need only take the form of wiping with damp cotton when necessary. This treatment, however, is not suitable for painted or gilded sculpture. Mist spraying with tap water is more practical for outdoor sculpture.

Repair of breaks

Modern adhesives are so effective in repairing broken sculpture that care must be taken in using them, as the chances of altering the repair after they have set is remote. Both the epoxy and polyester resins employed as adhesives are used only on clean, dry stone. The adhesive must be applied so that it does not appear on the surface, as it tends to yellow with age and become visible.

Dowelling is a commonsense operation. If a life-size arm is broken in two, for example, a stainless-steel dowel of about three-eighths of an inch diameter will suffice, and, if the stone is sound, it need only be six inches in length. A spot of coloured paint is applied to the centre of one half, and the two pieces are fitted together. Upon removal of one half, the paint will have reported on the other half. A hole is then bored with a mason's drill, a little over three inches in depth, in each piece, and slightly broader than three-eighths of an inch in diameter, great care being taken to ensure that both holes are aligned. The dowel is then inserted and the halves fitted together dry. If the union is undisturbed by the introduction of the dowel, the polyester adhesive should be applied to both holes and a little to cover the break. The pieces are then firmly fitted together and held in position until set.

Chips and missing areas in a piece exhibited indoors are generally replaced by using solid polyvinyl acetate, which can be coloured to match the surrounding material and is easily molded. An item such as a hand can be produced satisfactorily by using a water-clear epoxy as a casting medium.

The traditional, although usually unsatisfactory, method for doing this work out-of-doors is to use a cement with crushed marble as an aggregate. An epoxy resin with an aggregate of ground, white glass is usually found to be much more aesthetically pleasing and durable. Grafting a new piece of marble onto a statue not only means that the break must be trimmed back but also calls for a degree of sculpting comparable to the original.

While there is a reluctance on the part of painters to repaint pictures, painted sculpture has not experienced this repression. Wooden sculpture from the 14th century may have five to seven layers of repainting. The age and insolubility of these old repaintings inevitably means that to uncover the original layer the subsequent layers have to be removed manually, a long and delicate job. The techniques are similar to those for restoring panel paintings. Final varnishing, however, is of a lower gloss than that usual with panels.

DECORATIVE ARTS

Furniture. The need for restoration of furniture stems from several causes—physical damage, destruction by woodworm, decomposition of glue caused by dampness. A major problem is damage by central heating. Old furniture was constructed with timber seasoned to be compatible with a more natural humidity level. Today much of the world's fine furniture has found its way into centrally heated museums, galleries, and private homes. The drier conditions affect the wood and glue, causing shrinking, which results in splitting of wood, peeling veneer, and loosening of inlay. The correct use of humidifiers usually averts this problem if introduced in time.

Woodworm is best treated with methyl bromide gas. Liquid pesticides are often impractical because they may destroy or mar painted or gilded surfaces.

The restoration of a damaged piece of furniture requires many techniques and skills. It is, first, very difficult to obtain properly seasoned rare woods with which to execute repairs. The staining and polishing of the repair element present problems of subsequent fading and must not interfere with the original patina of the surrounding area. Polyvinyl acetate and epoxy resins have largely superseded traditional animal glue. The restoration of veneers and inlays of diverse materials such as tortoiseshell, ivory, pewter, and mother-of-pearl all require individual processes of restoration, as do the restoring of leather coverings, fabric upholstery, or caning.

Stained glass. Generally damaged by war, vandalism, or degeneration of the lead support and, less commonly, by devitrification of very old glass, stained-glass windows can be restored. Missing pieces are usually recut from new glass especially made for the purpose. If there was painting on the original glass, it must be skillfully redrawn on the new glass. Care is taken to match the delicate original shading. After painting or staining, the glass is fired and then annealled. To restore flashed glass (clear glass coated with coloured glass), the coloured coating is masked with wax so that the unmasked area can be removed or refined with hydrofluoric acid. Occasionally, the coating is ground off to produce a design.

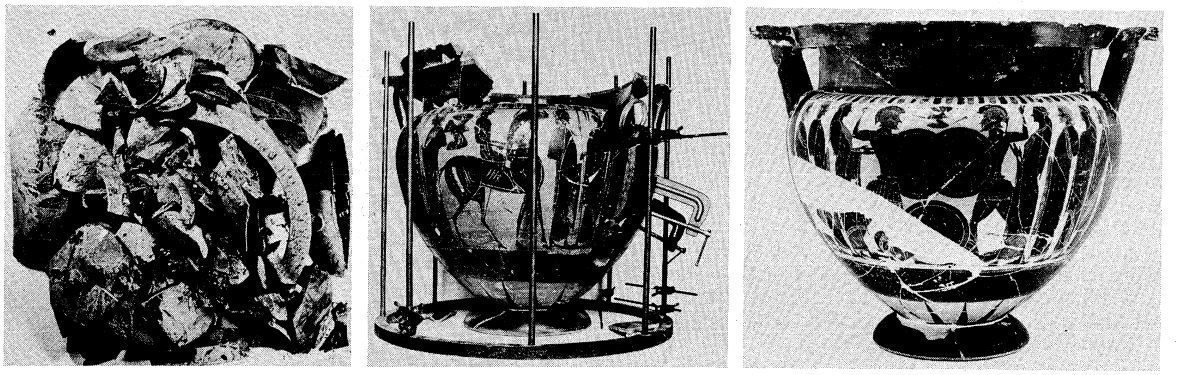
Broken fragments are often found leaded together, but this tends to confuse the design. Often no attempt has been made to produce a new missing piece, but an odd fragment, perhaps from another window, has been cut to fit. Stained-glass windows have to be completely releaded if the original leading has decomposed. Very old glass sometimes becomes completely opaque, and sometimes the external surface becomes chalky. With great care and patience, this film can be ground or etched off with acid, restoring the glass to something of its original beauty.

Textiles. In terms of preservation, textiles are one of the most fragile mediums of the decorative arts. Made of plant fibres and animal hair, textiles are very perishable and generally do not survive the test of time except in dry desert climates such as parts of Central Asia, Egypt, and Peru. Sunlight fades the dyes used to colour textiles. Sometimes only certain colours will be so affected. Usually the colours that are darker in hue are more faded or decayed because they absorb the most light. When exhibited, therefore, the rooms or cases where these materials are installed are air-conditioned and lighted with ultraviolet filtered light.

Hanging textiles, such as tapestries, often tear from their own weight or cause the weaving of the fabric to separate. Such tears can be averted by sewing the hanging to a backing of another textile. If there are torn or moth-eaten areas, new warp threads can be inserted and the damaged design rewoven into the work.

The storage and display of delicate fabrics is even more crucial than for tapestries. Damaged silks that are beyond needle restoration are sometimes reinforced by being welded onto a supporting material with polyvinyl acetate emulsion. Consolidation of decaying fabric is also effected by soluble nylon or polyvinyl acetate. A whole range of techniques is called for in the cleaning of the many types of textile materials. The success of cleaning still depends on the skill of the conservator more than on the use of modern chemicals.

Ceramics. The principles that govern the restoration of pottery are exactness, durability, and noninterference



Greek black-figure vase from Orvieto destroyed during the flood of 1966 in Florence, Italy, shown (left to right) before, during, and after restoration.
By courtesy of the Soprintendenza alle Antichita d'Etruria, Florence

with visual appreciation. Breaks are sometimes disguised by complete overspraying, which often alters the character of the glaze. Much skill is required in the reconstruction of a fragmented object, even with the use of modern adhesives such as epoxy resins and polyvinyl acetate emulsions. It is crucial to induce the maximum intimacy between the breaks. This means little adhesive, no dirt or loose particles, and precise positioning. Missing parts are usually made with a modern cellulose filler or a composition formed by mixing kaolin with epoxy resin. This exercise must be perfect, for, however well the join is painted, it will never be satisfactory if the filling is imperfect. When the restorer is satisfied with the form and surface of his repair, he views it in an oblique light to show up imperfections.

The missing heads, arms, or hands of figurines are often modelled in low-temperature firing clay and fired onto the original. This is rarely done by museum restorers, for the process may alter the appearance of the original or endanger the glaze. A cold curing resin can safely be used as a medium for retouching. (K.F.B.H.)

BIBLIOGRAPHY

Conservation of buildings: JOHN H. HARVEY (comp.), *Conservation of Old Buildings* (1958), a general bibliography; DONALD W. INSALL, *The Care of Old Buildings: A Practical Guide* (1972); *Set of Four Pilot Reports on Historic Towns* (HMSO, 1968), on Bath, Chester, Chichester, and York; PRESERVATION POLICY GROUP, *Report to the Minister of Housing and Local Government* (1970), a concerted attack upon the problems of historic city conservation in Britain; ORIN M. BULLOCK, *The Restoration Manual: An Illustrated Guide to the Preservation and Restoration of Old Buildings* (1966); JANE JACOBS, *The Death and Life of Great American Cities* (1961).

Conservation of paintings: HELMUT RUHEMANN, *The Cleaning of Paintings*, with a comprehensive bibliography by JOYCE PLESTERS (1968); HAROLD J. PLENDERLEITH and ANTHONY E. WERNER, *The Conservation of Antiquities and Works of Art*, 2nd ed. (1972); UNESCO, *The Conservation of Cultural Property* (1968); ARTS COUNCIL OF GREAT BRITAIN, *Frescoes from Florence*, with an introduction by U. PROCACCI (1969).

Conservation of sculpture: Proceedings of the 1970 IIC Conference on stone and wood; *La conservazione delle sculture all'aperto* (1969).

Conservation of other works of art: By far the most valuable accounts are contained in the quarterly publications of the International Institute for Conservation of Historic and Artistic Work, London (IIC). See also C.S.M. PARSONS and F.H. CURL, *China Mending and Restoration* (1963); JOHN RODD, *The Repair and Restoration of Furniture* (1954); and *Studies in Conservation*, vol. 16, no. 2 (1971), on ceramics.

(D.W.I./N.S.B./K.F.B.H.)

Artevelde, Jacob van

Jacob van Artevelde was a 14th-century Flemish leader who played a leading role in the preliminary phase of the Hundred Years' War. A citizen of Ghent, van Artevelde kept the Flemish cities neutral at first; then he sided with England, in spite of the nominal allegiance of Flanders to France.

Life. Van Artevelde was born in Ghent *c.* 1295; his profession is unknown, but he belonged to the wealthy bourgeoisie and owned land both in Ghent and in the surrounding area. His children's marriages indicate a connection with the nobility. His second wife, Kateline de Coster, took an active and capable part in public life. She travelled several times to England, in van Artevelde's name, to obtain payment of sums promised by the English king to the Flemish towns. One of their sons, Philip (born 1340), led a revolt against Count Louis II of Flanders in 1382. Van Artevelde had already reached middle age when he began to take part in public affairs. The only mention of him before 1338 is as a supporter of Louis I, count of Flanders, during a revolt against Louis in Ghent in 1325. But as relations between England and France worsened in the 1330s, tension arose between the Count and the Flemish towns. Louis, a vassal of the French king Philip VI, sided with France. The towns, although Philip offered them inducements, needed English wool for their textile industry and could not afford to alienate Edward III of England.

Artevelde's rise to power

At that point, van Artevelde emerged as a leader. Early in 1338, the people of Ghent, under his leadership, declared their neutrality. Bruges and Ypres, the other major towns, followed suit. France was forced to acquiesce, and the vital trade with England was safeguarded.

Van Artevelde governed Ghent with a group of four other "captains." At least three of his colleagues were wealthy merchants. A dean of the weavers was also elected, and the extent of the power conferred on a member of the less prestigious artisan class was characteristic for the social aspect of van Artevelde's reforms.

Flanders remained neutral for only two years. At the beginning of 1340, it joined the English, thereby obtaining further commercial advantages and a promise by Edward III to help reconquer some Flemish areas under French control. To give the new situation some show of legality, the English king, probably on van Artevelde's initiative, let himself be proclaimed king of France in Ghent (Jan. 26, 1340). Edward overcame the French fleet near the Flemish port of Sluis and then, with van Artevelde, besieged the French city of Tournai.

Unlike his subjects, the Count of Flanders fought on the French side, and the Flemings, under van Artevelde, repudiated his authority. When a truce was declared after the siege of Tournai, the Count returned to Flanders, but he was subjected to the control of a council dominated by van Artevelde's supporters. When Louis extricated himself by flight, his relative Simon van Mirabello, a wealthy Lombard who had become a citizen of Ghent, was made regent by van Artevelde to replace him. Van Artevelde never made formal changes in institutions. His name never appears on an official document, but Edward III's correspondence shows that he exercised power. He owed his influence to the support of the great towns, whose interests he promoted. When necessary, he sacrificed the interests of the small towns and the rural population to his end.

Van Artevelde maintained his position unchallenged until the beginning of 1343, when an unsuccessful attempt to overthrow him was made by a onetime alder-

man, Jan van Steenbeke. A second attempt, in May 1345, was more successful. A conflict arose between the weavers and fullers, putting an end to the policy of balance among all classes, on which van Artevelde's authority rested. He lost his position as chief captain but retained the confidence of Edward III. That was fatal to him. In July 1345 Edward came to Sluis to negotiate the continuation of the alliance with Ghent's representatives. Van Artevelde quarrelled with his colleagues, who thought him too compliant. On his return to Ghent, he was murdered during a riot (*c.* July 22). Jacob van Artevelde long remained a controversial figure; more or less forgotten in the 17th and 18th centuries, his memory was resurrected by Belgian-nationalist historians in the 19th century.

His death

Personality. It is difficult to assess van Artevelde's personality. Most of the surviving information about him comes from his enemies. His only known writings are three letters, written in French during the winter of 1342–43 and addressed to the king of England, the queen, and the prince of Wales. They are businesslike in tone, without exaggerated compliments. It is possible to deduce from his actions some of the characteristics that made him stand out: a strong conviction about the needs of his people and how to secure them, a willingness to break with accepted values, and the ability to act without hesitation. His violent disposition, which led him to kill an opponent in a quarrel, was a quality he shared with many of his contemporaries.

BIBLIOGRAPHY. N. DE PAUW, *Cartulaire historique et généalogique des Artevelde* (1920), a nearly complete, but rather uncritical, collection of sources on Jacob van Artevelde and his family; H.S. LUCAS, *The Low Countries and the Hundred Years' War, 1326–1347* (1929), an analytic description of van Artevelde's part in the Hundred Years' War in its first stage; HANS VAN WERVEKE, *Jacques van Artevelde* (1942), a synthetic view on his life and political career.

(H. van W.)

Arthropoda

The phylum Arthropoda is the largest and probably the most diverse phylum in the animal kingdom. Its members, which form about 75 percent of the known animal species, differ widely in anatomy, physiology, behaviour, and habitat. One characteristic that distinguishes arthropods from other animal phyla is a resistant outer covering called a cuticle, which is composed in part of a protein (chitin) and is shed at periodic intervals (molt, or ecdysis). The cuticle serves as a protective device and functions as an exoskeleton, providing surfaces for muscle attachment. Other distinguishing characteristics of arthropods include a segmented body, a six-segmented head in Mandibulata, chitinous jointed appendages (hence the name Arthropoda, meaning jointed legs), many specialized sense organs, a circulatory system containing vascular spaces collectively called a hemocoel, and bilateral symmetry.

More than 925,000 arthropod species have been described, of which about 90 percent are insects. The total number of arthropod species has been estimated at 6,000,000 or more. The phylum Arthropoda may be divided into three subphyla: Trilobitomorpha (containing only the extinct trilobites), Mandibulata, and Chelicerata. The Mandibulata include the water-inhabiting crustaceans (class Crustacea) and the terrestrial insects (class Insecta), springtails (class Collembola), and myriapods, a collective name commonly used to refer to members of four closely related classes (Pauropoda; Diplopoda, or millipedes; Chilopoda, or centipedes; and Symphyla) with similar leg-bearing body regions unlike those of insects. The Chelicerata include the marine class Merostomata (xiphosurids and eurypterids), the class Arachnida (scorpions, spiders, mites, opilionids), and the marine class Pycnogonida (sea spiders).

Of uncertain relationships are the arthropod-like tardigrades, pentastomids, and onychophores; they are described fully in the article ONCOPOD.

Arthropods are common on land and in marine waters and fresh waters. In the sea, minute crustaceans are a major component of the zooplankton, which serves as food for other invertebrates, fishes, and whales. The land is dominated by the insects, which are of economic importance both as pests and as pollinators of crop plants. Spiders, mites, scorpions, and other arachnids also live on land. Centipedes, millipedes, symphylids, and pauropods, which live in damp habitats, are of little economic importance. The trilobites, which become extinct by the Permian (about 280,000,000 years ago), were dominant arthropods in the Early Paleozoic seas (about 550,000,000 years ago); their earliest fossil remains, found near the bottom of the Lower Cambrian rocks, are about 570,000,000 years old.

GENERAL FEATURES

Size range and diversity of structure. The hard external skeletons of arthropods influence their ultimate size. Only aquatic forms are able to attain substantial sizes because their bodies are supported in part by the surrounding water. The extinct Eurypterida, for example, reached a length of 1.8 metres, and some present-day crustaceans grow to more than twice that size; giant spider crabs may weigh up to 6.4 kilograms and span 3.8 metres. The strongly calcified body coverings of crustaceans provide weight that is advantageous to bottom-dwelling aquatic forms.

Influence of external skeleton

The weight of the exoskeleton, however, is not the only factor that limits arthropod size; molting also plays a role. The cuticles of terrestrial arthropods, strengthened by sclerotization (a hardening process analogous to the tanning of leather), are not only tough and rigid but also light in weight. Animals encased in rigid exoskeletons, be they light or heavy, can increase in size only by molting; and, after attaining certain sizes, their linear dimensions increase only slightly during intervals between molts (called instars). The necessity for molting, or ecdysis, therefore, sets an upper limit to the size of arthropods.

Terrestrial arthropods do not attain large sizes. The largest insects and spiders do not weigh more than 100 grams. The beetle *Goliathus regius* measures 15 centimetres in length and ten centimetres in width, while the butterfly *Ornithoptera victoriae* of the Solomon Islands has a wing span exceeding 30 centimetres. The phasmid *Pharnacia serratipes* is one of the longest insects at 33 centimetres. It is probable that the size of terrestrial arthropods may be limited not only by the weight of the exoskeleton but also by their vulnerability; significant, perhaps, is the fact that scorpions and Solifugae are nocturnal, a habit that may be related to preventing attacks by vertebrate enemies.

The smallest arthropods include some parasitic wasps, beetles of the family Ptiliidae, and mites that are less than 0.25 millimetre in length, despite their complex structures, and may weigh less than the nucleus of a large cell. Ants 1.5 centimetres long weigh less than one gram.

Distribution and abundance. The only region of the earth in which insects are not dominant invertebrates is the sea; the probable reason is that every suitable habitat had been exploited by other arthropods, the crustaceans, before the insects evolved. The crab *Ethusina abyssicola* is known from a depth of more than 4,000 metres, while Collembola and jumping spiders (Salticidae) are found on Mt. Everest at heights exceeding 6,700 metres. Collembola and orabatid mites are among the permanent inhabitants of Antarctica. Many arthropods are parasites in or on other animals.

Importance. Arthropods are of great economic and medical importance. Despite increases in knowledge and the development of powerful insecticides, insects and ticks and other mites (Acarina) remain serious disease threats to man and other animals throughout the world; arthropod-carried human diseases include malaria, yellow fever, rickettsial diseases, plague, filariasis, and other worm infections. Arthropods also cause harm by their poisonous stings and bites and by their destruction of crops, timber, and stored products.

Soil arthropods, which include some crustaceans, myriapods, insects, and collembolans, play an important role in the formation of humus from decomposed leaf litter.

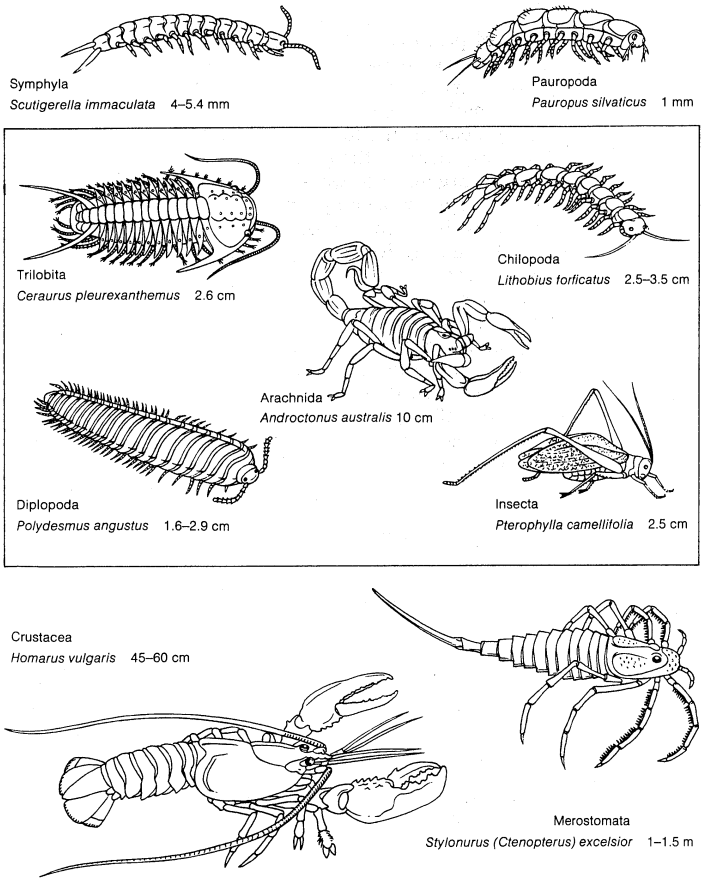


Figure 1: Diversity of arthropod body plans.

Crustaceans are important food supplies of man and other animals.

NATURAL HISTORY

Reproduction and life cycle. With few exceptions, the sexes are separate in arthropods; *i.e.*, there are both males and females. The paired sex organs, or gonads, of each sex are connected directly to ducts, called gonoducts, which in turn are connected to regions that lead outside the animal (gonopores); glands frequently are associated with the ducts. The gonoducts open through gonopores to the exterior at different regions of the body of the various arthropod groups (*e.g.*, gonopores are near the rear of the body in centipedes and insects; at the hind end of the thorax in Crustacea; close to the head in millipedes, pauropods, and Symphyla; and near the middle of the body in Arachnida). The abdomen of the pycnogonids contains little more than the terminal part of the alimentary canal, so that the gonads are displaced into the leg segments, and eggs pass through openings in the second segment.

Spermatophores

Throughout the Arthropoda, spermatozoa often are transferred to the female in sealed packets known as spermatophores. During this primitive method of transfer, the spermatozoa are not diluted by the surrounding medium, in the case of aquatic forms; nor, on land, do they suffer from desiccation. Spermatophores, characteristic of centipedes (Chilopoda) and Arachnida in which mating is delayed by elaborate courtship behaviour, also occur typically among certain insect groups—*e.g.*, Orthoptera (grasshoppers, crickets), Dictyoptera (cockroaches, mantids), Lepidoptera (butterflies, moths), many Coleoptera (beetles, weevils), and Trichoptera (caddisflies). The transfer of free spermatozoa occurs in the Diptera (true flies), some Heteroptera (true bugs), Diplopoda (millipedes), Opiliones, and some mites.

In its simplest form in arachnids such as Solifugae, Ricinulei, and some Acarina, the spermatophore, a mucilaginous mass of spermatozoa, is transferred to the female by the male with the aid of his appendages (*e.g.*, chelicerae, pedipalps, or legs). In other arachnids (*e.g.*, Scorpionida, false scorpions, Amblypygi, Thelyphonida, Schizomida, and some Acarina) in which the complex, rodlike spermatophore is deposited on the ground, mating involves a nuptial dance, during which the male manoeuvres the female into a position such that she can gather the spermatozoa into her genital opening. In spiders (Araneida), spermatozoa are transferred to the female by processes, called palpal organs, attached to the male prosoma. The tarsal and metatarsal segments of the third pair of legs of Ricinulei are modified for a similar purpose.

Unusual methods of reproduction found among arthropods include the development of unfertilized eggs (parthenogenesis), the birth of living young (viviparity), and the appearance of several embryos from a single fertilized egg (polyembryony). Arthropod eggs are usually rich in yolk; in all groups, especially among Crustacea, however, are species whose eggs have little yolk.

Most crustacean and many myriapod embryos have fewer body segments than the adult forms. A free-swimming microscopic larval stage (nauplius) of some crustaceans, for example, has only three segments in addition to an anterior nonsegmented portion (acron); additional segments appear at regular intervals when shedding of the exoskeleton (molting) occurs. In contrast, in many chelicerates and insects, most segments form during embryological development within the egg.

Locomotion. *Walking and running.* The locomotion of living arthropods is similar to that of the more primitive wormlike onychophorans, the bodies of which are extensible. The number of locomotory appendages, which varies among the arthropod classes, is greatest in burrowing forms—*e.g.*, millipedes that push their way through crevices and certain centipedes that move through soil. The ability to move rapidly, associated with a reduction in the number of limbs and a shortening of the body, is best developed in arachnids, which are equipped with four pairs of legs, and in insects, with three pairs. It is noteworthy that Solifugae, the fastest of the Arachnida, run with the last three pairs of legs only, holding the first pair above the ground.

In order to maintain stability, an insect must keep at least three of its legs on the ground. No foreleg or middle leg can be lifted until the limb behind it has taken up a supporting position, and each leg alternates with the corresponding leg of the same segment. The legs usually move in a set sequence—foreleg, opposite middle leg, and hindleg—but many variations occur. Claws on the fifth leg segment (tarsus) are used to walk over rough surfaces; padlike adhesive organs on the tarsus (*e.g.*, the arolium) are employed on smooth surfaces. Most insects that leap (*e.g.*, grasshoppers, fleas) do so by suddenly extending the fourth segments of the hindlegs (tibiae). In contrast, all of the legs of jumping spiders (Salticidae) are extended together by an increase in blood pressure.

Flight. Insects are unique among animals in that they have acquired wings directly and not through the modification of limbs, as is the case with birds, bats, pterodactyls, flying frogs, and flying fishes.

The origin of flight is still a matter of dispute. Some authorities believe that wings first evolved in large insects that were forced to escape predation by ancestral mygalomorph spiders. According to this hypothesis, the spiders evolved the ability to construct aerial webs after their prey developed ways to fly. An alternative hypothesis postulates that wings first arose in small ancestral insects that comprised part of the aerial fauna.

In many flying insects except Odonata (dragonflies), the power by which the wings are moved is provided by indirect wing muscles; in Dictyoptera, however, the muscles that drive the wings also are concerned with moving the legs. Analysis of the possible sequence of muscular movements in walking and in flight suggests that, in the primitive condition, a common central nervous pattern of control could have served the needs of both running and flying.

Modifications of insect wings

In many insects, the two pairs of wings are able to act as a single pair. In some groups, bristles and lobes couple the fore- and hindwings on each side. In others, one pair of wings may be used for functions other than flight; *e.g.*, in the Diptera (flies), the hindwings are reduced to structures called halteres, which contain many sense organs that act to control flight.

Ecology. With the exception of wood lice and land crabs, the Crustacea inhabit salt water or fresh water. Not only do they form an important element of zooplankton but they also are the primary scavengers of shore regions. The other mandibulate arthropods (*i.e.*, the myriapods and the insects) are primarily terrestrial in habit; some centipedes and insects, however, inhabit the shoreline between tide marks (littoral), and some insects (*e.g.*, among the Coleoptera, Diptera, Hymenoptera, Lepidoptera, Heteroptera) have developed an aquatic habit secondarily, mostly in fresh water. Least modified for an aquatic habit are insects with an open respiratory system in which air enters tubules called tracheae through holes known as spiracles. Other insects have developed a closed (apneustic) tracheal system, into which oxygen passes from the surrounding medium; elaborations of this system include tracheal gills and spiracular gills, the latter being adapted for respiration both above and below water in streams whose water level may fluctuate rapidly.

Extant chelicerates are terrestrial, except for horseshoe crabs (Xiphosura), the Pycnogonida (sea spiders), and a few spiders and mites that, like some of the insects, have returned to the water. Terrestrial arthropods differ very little from their marine relatives, with the exception of respiratory modifications and the evolution of wings (in insects). Arthropods function efficiently either in water or on land. Most of the differences between aquatic and terrestrial forms are concerned with functional variations.

Many primitive arthropod and arthropod-like animals, which live in soil, humus, and leaf litter, represent an intermediate stage between aquatic and terrestrial habits; tardigrades (water bears), for example, frequently inhabit mosses. Some crustaceans, including copepods and ostracods, are soil dwellers, as are, among others, myriapods, collembolans, termites, ants, flies, and mites.

The soil fauna lives in comparative security and has sufficient oxygen and little risk of desiccation. The large numbers of primitive arthropods that live beneath objects (known as cryptozoic arthropods) include Pauropoda, Symphyla, and a few Arachnida (Palpigradi, Ricinulei). Usually small and lacking pigmentation and efficient respiratory mechanisms for controlling water loss by evaporation, cryptozoic arthropods have poorly developed visual sense organs but well-developed tactile and taste sensillae. Cryptozoa may molt throughout life even though growth no longer occurs; these soil arthropods, which have a long fossil history, play a vital part in soil formation (see also SOIL ORGANISM).

Parasitic arthropods

Arthropods also exhibit many highly specialized parasitic forms. Pentastomida are entirely parasitic, and parasitic adaptations are common among certain crustaceans, cirripedes, copepods, and isopods; *Sacculina*, for example, is a cirripede that parasitizes crabs. Among insects, Phthiraptera (fleas), Mallophaga (biting lice), and Anoplura (sucking lice) are entirely parasitic in habit; many Heteroptera (bugs), Diptera (flies), and Hymenoptera (wasps) also are parasitic during some stage of their life cycle. Dipterous and hymenopterous parasites play an important ecological role in regulating the numbers of their insect hosts. Although most arachnids are carnivorous, many mites are parasitic—*e.g.*, harvest mites (Trombiculidae), the itch mite (*Sarcoptes scabei*), follicle mites (*Demodex*), and ticks (Ixodidae).

FORM AND FUNCTION

Immature forms. The newly hatched young of arthropods often are very different from adults in appearance and habits. One biological advantage of larval stages in development is that the young do not compete with the adults; a disadvantage is that larval young may be vulnerable to predators. The developmental stage during which an animal hatches depends upon the amount of yolk present in the egg; among arthropods in which embryonic development is direct and the young hatch as miniatures of adults, fewer eggs are produced per female.

Trilobites had larval forms. Crustaceans typically hatch as minute, nonsegmented nauplius larvae with three pairs of appendages and a single median eye. This stage may be passed in the egg, however, and the animal then hatches as a metanauplius larva (or even as a zoea) that has developed beyond the nauplius stage but is not yet an adult. Special larval stages occur in several crustacean groups—*e.g.*, cypris larvae in Cirripedia, cyclops larvae in copepods, zoeae in Malacostraca.

The pauropod egg hatches into an immobile stage with two pairs of legs; after molting, a six-segmented larva appears with three pairs of legs. Most millipedes hatch with seven trunk segments: the second, third, and fourth each with one pair of legs; the fifth and sixth each with two pairs of limb buds. The young of certain centipedes (Scolopendromorpha and Geophilomorpha), because they hatch with the adult number of segments, are said to undergo epimorphic development; other centipedes (Lithobiomorpha and Scutigeromorpha), however, like Pauropoda, Diplopoda, Symphyla, and insects of the order Protura, hatch with fewer than the adult number of segments and legs (anamorphic development).

Some insects hatch as nymphs that superficially resemble the adults (exopterygote insects); others (endopterygotes) characteristically pass through several immature stages (*e.g.*, larva, pupa) before becoming adults. Endopterygote insect larvae (*e.g.*, caterpillars, maggots, grubs) differ markedly from adults, inhabit different environments, and eat different foods. A pupal or chrysalis stage bridges the gap in form and habit between larva and adult. The differences between adult and immature stages of exopterygotes, however, are so slight that the necessary changes can be accomplished without metamorphosis.

The young of most arachnid classes develop directly. In the case of scorpions, Amblypygi, and wolf spiders, the young climb onto their mother's back immediately after birth. The eggs of some scorpions, rich in yolk, develop in the mother; in the family Scorpionidae, however, the fertilized egg lacks yolk and is nourished by nutrient fluids from the wall of the mother's intestine. The wall of the egg-producing organ (ovary) in pseudoscorpions be-

comes glandular after the eggs are formed and secretes a nutrient fluid on which the embryos feed. The young of false scorpions usually emerge as nymphs called protonymphs that molt to form tritonymphs and then adults. The young of Opiliones, Solifugae, and spiders resemble their parents; in mites (Acarina), however, four postembryonic development stages occur, separated by molts. The six-legged larvae may pass through several stages (protonymph, deutonymph, tritonymph), which usually resemble the adults.

Body regions and appendages

Adult structure. The success of the arthropods derives from the presence of an exoskeleton and division of the adult body into regions that may, or may not, be marked off sharply. The specialized body regions of arthropods are sometimes called tagmata. In Onychophora and mandibulate arthropods, for example, the head is followed by a trunk region, which, in crustaceans and insects, is divided into thoracic and abdominal tagmata. The foremost region (tagma) of arachnids, called the prosoma or cephalothorax, bears the legs and the appendages used in feeding; the divisions of the hinder part of the body, called the opisthosoma or abdomen, may be divided into a mesosoma and metasoma, or tail. A nonsegmented terminal part of an arthropod often is called a telson. Any arthropod tagmata may be nonsegmented externally as a result of the fusion of segments during embryonic development. Head and thorax (or cephalothorax) are often covered by a protective cover (carapace).

The number of segments incorporated into the head varies among the different arthropod classes—*e.g.*, two in trilobites, and probably six in crustaceans, myriapods, and insects. The arachnid cephalothorax, or prosoma, consists of seven segments. The second and third segments bear the feeding appendages (chelicerae and pedipalps); the remaining four carry the legs.

The paired appendages of arthropods consist of several parts arranged as a linear set (uniramous) or as a bifurcated set (biramous). Two basic limb types are distinguishable in Crustacea, a flattened unsegmented one (phyllopodium) and a rounded one (stenopodium). The primitive limb type may have consisted of nine segments bearing a biting process, or gnathobase, on the first segment and an outer branch, or exopodite, on a more distal segment.

Changes in shape and relative size of parts of arthropod limbs have resulted in organs variable in appearance and function. Limbs may be modified for swimming, walking, respiration, reproduction, or as sense organs or mouthparts; the spinnerets of spiders, for example, are modified abdominal limbs. The mouthparts of primitive insects (*e.g.*, cockroaches) resemble crustacean limbs; parts of the maxilla (lower jaw), for example, are similar in origin to those of a crab. Primitive mouthparts adapted for chewing, as found in cockroaches, earwigs, locusts, and beetles, have evolved various modifications for piercing and sucking; these occur in the mouthparts of numerous insects—*e.g.*, bugs, bees, fleas, flies, butterflies.

Exoskeleton

The exoskeleton surrounding each body segment consists basically of four plates—a dorsal tergum, a ventral sternum, and two lateral pleurites. The plates around each segment and those of adjacent segments are attached by internal segmental muscles.

Functional features. Arthropods have exploited every conceivable source of food; in many cases, digestion is assisted by bacteria or protozoans that live in symbiotic associations in the arthropod alimentary canal. The alimentary canal consists of a chitin-lined foregut and hindgut and a midgut, or mesenteron. Salivary glands open at the mouth or the anterior part of the foregut; digestive glands open into the mesenteron. The hindgut terminates at the anus, which is usually situated on the under side of the telson. The excretory system of aquatic groups consists of glands, called coxal glands or green glands, which are found at the bases of certain of the segmental appendages. Terrestrial arthropods and Amphipoda (Crustacea), but not Collembola, possess Malpighian tubules for excretion of nitrogen-containing wastes; the principal excretory product is uric acid (guanine in arachnids), and the urine is semisolid.

The blood, or hemolymph, that bathes the internal organs, is circulated by a tubular, segmented heart. In arthropods with localized respiratory organs, such as gills and lung books (paired respiratory structures internally arranged as pages in a book), the blood contains a respiratory pigment, hemocyanin. No respiratory carrier is found in arthropods with tracheate respiratory systems because oxygen is conveyed directly by small tubes (tracheoles) to a pigment (cytochrome) in the tissues.

The nervous system consists of segmented ganglia that may unite during embryonic development into larger units, such as cerebral and subesophageal ganglia; these may have both nervous and neurosecretory functions. The gonads of most arthropods have no segmentation; in a number of groups, the left and right gonads are fused. The cavities in which the gonads lie represent the remnants of an ancestral body cavity (coelom).

Cuticle

The exoskeletal integument, or cuticle, of arthropods consists of a thin, impermeable, nonchitinous outer layer (epicuticle) and a thick, elastic, permeable, lamellar inner layer (endocuticle), composed largely of the protein chitin. The outer layers of the endocuticle usually are hardened by sclerotization or in Xiphosura, Crustacea, and Diplopoda by deposition of a salt (calcium carbonate). Molting is controlled by hormones; before the hard outer layers of the cuticle are shed, the inner layers are digested by an enzyme (organic catalyst). The arthropod swallows water or air to create a pressure that is distributed by the vascular spaces (hemocoel) to all parts of the body, which remain swollen until the new cuticle has hardened.

From H. Weber, *Grundriss der Insektenkunde*, 4. A. (1966); G. Fischer, Stuttgart

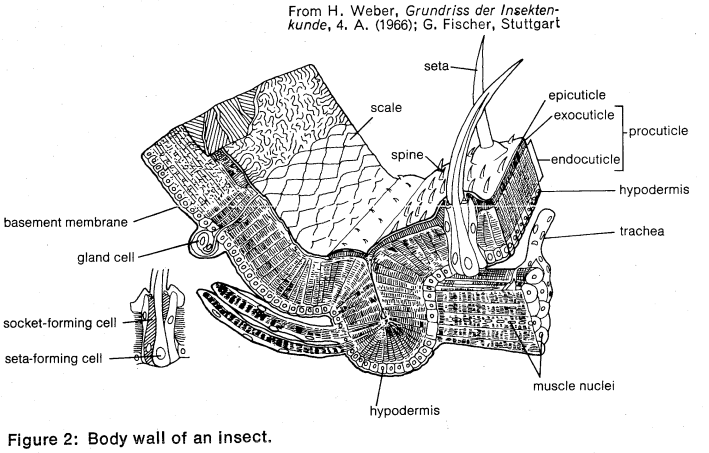


Figure 2: Body wall of an insect.

The integuments of terrestrial arthropods function not only in support and protection but also in preventing water loss, of vital importance in view of their small size and proportionately enormous surface area. Many cryptozoic forms (*e.g.*, wood lice, myriapods, Collembola) live in damp or humid environments, such as those found beneath stones, leaves, bark, and in soil and crevices. Most insects and arachnids, however, avoid desiccation by an epicuticle, which is waterproofed with a wax layer that is protected from the environment by a cement layer secreted by epidermal gland cells. The cuticle thus is impervious not only to water vapour but also to oxygen and carbon dioxide. A specialized respiratory mechanism sensitive to carbon dioxide permits gaseous exchange to take place while minimizing water loss; *e.g.*, the spiracles of insects and the lung books of arachnids open to facilitate respiration only when carbon dioxide begins to accumulate in the body. Before molting takes place, a new wax layer is secreted beneath the old cuticle, so that molting occurs with minimal water loss.

Excretion and nutrition also are closely concerned with water conservation. Insoluble excretory compounds, such as uric acid and guanine, were probably first acquired in association with the evolution of an enclosed egg. Uric acid and guanine are equally important to the periods between molts (instars) of larval and adult forms of terrestrial arthropods, however, since both compounds can be concentrated in the Malpighian tubules and excreted

with feces from which water has been withdrawn. It is probable that the fate of water ingested with food may be controlled by factors produced by a cerebral neurosecretory system.

Arthropod hormones control colour change, retinal pigment migration, growth, molting, changes in body form (metamorphosis), and sexual development.

Arthropods may develop poison glands that are associated with a variety of functions. In centipedes, scorpions, false scorpions, and spiders, poison glands are used to aid in the capture and digestion of prey. In certain hymenopteran insects (*e.g.*, wasps), poison glands are associated with the egg placement structure of the female (ovipositor) or a sting; the larvae of several lepidopterans (butterflies) have poison glands that may be associated with projections such as setae or spines. When broken, these projections allow the discharge of a fluid that causes an itching sensation in man. Certain glands that produce substances with repugnant odours function in defense and are found in wood lice, millipedes, harvest spiders, Schizomida, and Thelyphonida.

Sense organs

Apart from the eyes, which may be simple or compound, the sense organs, or sensilla, of arthropods are either tactile hairs (trichoid sensilla) or modifications sensitive to smell (olfactory), taste (gustatory), sound (auditory), or internal (proprioceptive) stimuli. Chemotactile pegs and cones often are located on the antennae, as are humidity and temperature receptors; chordotonal sensilla are attached to tympanal organs of sound reception. Receptors called proprioceptive campaniform sensilla, which measure tensions in the cuticle, are found near insect wings. Lyriform organs on the legs of spiders respond to vibrations of the web, so that the struggles of potential prey can be distinguished from the web plucking of a courting male. Arthropod senses usually are acute.

EVOLUTION AND PALEONTOLOGY

The arthropods have much in common with annelid worms of the class Polychaeta, and it is generally accepted that both are derived from the same ancestral stock. The peculiarities of arthropods, like their successors, may be related to the development of the cuticle into a jointed exoskeleton, which allows the segmented appendages to become adapted to a variety of functions—*e.g.*, feeding, locomotion, respiration, sensory reception, and reproduction. Associated with the development of muscular limbs is the replacement of the body cavity (coelom) by vascular spaces (hemocoel), in which the organs of the body are bathed in blood. The nervous system is like that of the Annelida.

The earliest known arthropods, the trilobites, were marine, as were the scorpion-like Eurypterida. Of the extant arthropods, only the crustaceans are predominantly aquatic; the Arthropoda, however, contains the largest proportion of terrestrial members of any invertebrate phylum.

Although fossil arthropods have been reported from Precambrian rocks (more than 570,000,000 years old), the validity of such records has not yet been established. Since fossils of certain trilobites, branchiopod Crustacea, and merostomes date from Early Cambrian times (about 550,000,000 years ago), the Arthropoda had a long Precambrian existence, during which they became modified from their annelid ancestors and branched along various evolutionary lines. The earliest known arachnid fossils appear in Upper Silurian strata (about 400,000,000 years ago); Collembola, insects, millipedes, and sea spiders (Pycnogonida) appear in the upper parts of the Devonian system (about 350,000,000 years ago). Fossils of tardigrades and pentastomids have not been found.

CLASSIFICATION

Distinguishing taxonomic features. The phylum Arthropoda is divided into classes on the basis of comparative anatomy, embryology, and function. Differentiation of habits in similar environments appears to have been of great importance in the evolutionary differentiation of terrestrial groups. Modification, specialization, number, and appearance of body segments and appendages (especially anterior ones such as antennae and mouthparts) are important criteria in distinguishing arthropod classes. Other structural features of taxonomic significance include location of the gonopore, structure of the head, and adaptations of the respiratory and excretory systems. In the classification below, the group marked with a dagger (†) is wholly extinct and known only from fossils.

Annotated classification.

PHYLUM ARTHROPODA

Bilaterally symmetrical invertebrates with joined exoskeleton covering body and legs; appendages specialized for specific functions; body wall containing cuticle secreted by epidermal cells; outer endocuticle thickened to form skeletal plates (sclerites); waxy compounds of cuticle reduce water loss in terrestrial forms; cuticle shed periodically (molt, or ecdysis); body musculature complex, sometimes specialized (*e.g.*, for flight, sound production); coelom reduced, replaced by hemocoel; nervous system consists of dorsal brain and a double ventral nerve cord; cilia absent; development highly modified with eggs typically rich in yolk.

†**Subphylum Trilobitomorpha** (trilobites)

Extinct arthropods; head (or cephalon) composed of 5 segments bearing a pair of segmented antennae, compound eyes; body with 3 regions, cephalon, thorax, and pygidium.

Class Trilobita

Numerous in Cambrian (about 500,000,000–570,000,000 years ago) and Silurian (395,000,000–430,000,000 years ago) but extinct by the Mesozoic Era; marine organisms with oval, flattened body molded longitudinally into 3 lobes; trunk and pygidium (tergites usually fused) with biramous appendages; eggs hatched into protaspis larvae; length 3.5–75 cm; over 4,000 fossil species known.

Subphylum Mandibulata

Head with antennae, mandibles, and food-handling appendages (maxillae); thorax with appendages (uniramous or biramous) sometimes associated with mouthparts; abdomen undivided from thorax, with appendages, or distinct from thorax, with or without appendages.

Class Crustacea

Mostly aquatic in habit; respiration by gills; strong exoskeleton, 2nd and 3rd segments with antennae, 4th with a pair of mandibles; eyes, when present, stalked or unstalked; evolutionary tendencies include specialization of limbs, shortening of body, reduction in number of segments, and development of carapace; many parasitic adaptations; size range 0.5 mm–2 m; over 25,000 living species known (see CRUSTACEA).

Class Pauropoda

Very small myriapods with 2 pairs of appendages transformed into mouthparts and 8–11 (usually 9) pairs of walking legs; antennae with 4 (rarely 6) segments with short branches and long, multiarticulate flagella; length up to 1.9 mm; about 180 living species known (see MYRIAPOD).

Class Diplopoda (millipedes)

Elongated myriapods; abdomen an indefinite number of double segments, each with 2 pairs of legs and spiracles; head with mandibles, maxillae fused (form a gnathochilarium), sometimes simple eyes (ocelli), and short, club-shaped antennae; thorax of 4 single segments with gonopores in the 3rd (progoneate); length 3 mm–28 cm; about 8,000 living species known (see MYRIAPOD).

Class Chilopoda (centipedes)

Elongated myriapods with many distinct abdominal segments; each with 1 pair of legs; head with ocelli, flagellate antennae, mandibles, 1st and 2nd maxillae; limbs of 1st abdominal segment modified as poison claws; gonopore on last segment (opisthogoneate); length about 5 mm–26.5 cm; about 2,800 living species known (see MYRIAPOD).

Class Symphyla

Small myriapods with 3 pairs of mouthparts, 12 pairs of walking legs, and a posterior pair of spinnerets; gonopore usually on 4th trunk segment; length up to 8 mm; about 120 living species known (see MYRIAPOD).

Class Collembola (springtails)

Small, widely distributed, insect-like opisthogoneate arthropods; mouthparts ectognathous, antennae usually 4-segmented; eyes simple; 3 thoracic segments with legs; 6-segmented abdomen with forked springing organ; tracheae usually absent; no Malpighian tubules; length up to 5 mm; about 1,500 living species known (see APTERYGOTE).

Class Insecta (*Hexapoda*)

Mandibulate arthropods with 3 pairs of appendages modified to form mouthparts; head of 6 segments with 1 pair of flagel-

late antennae, usually both median and lateral eyes; thorax of 3 segments, each with a pair of legs and wings (usually on 2nd and 3rd), abdomen of 11 segments, without appendages in adult; gonopores posterior; length 0.25 mm–33 cm; about 750,000 living species known (see INSECTA).

Subphylum Chelicerata

Prosoma without antennae but with pincerlike chelicerae, and sometimes pedipalps; uniramous walking legs on thoracic segments; if present, appendages on abdomen highly modified.

Class Merostomata

Large marine chelicerates with gill books; prosoma completely covered by carapace; opisthosoma bears a long spine; of 2 orders, only Xiphosura has living representatives (4 species); Eurypterida (= Gigantostraca) includes 200 fossil species, the largest 1.8 metres long.

Class Arachnida

Chelicerate arthropods with body divided into prosoma and opisthosoma, sometimes connected by a narrow pedicel; prosoma bears chelicerae, pedipalps, and 4 pairs of legs; opisthosoma usually lacks appendages; respiration by means of lung books, tracheae, or both, opening on the opisthosoma; gonopore always on lower surface of 2nd opisthosomal segment; length 0.25 mm–18 cm; almost 60,000 living species (see ARACHNIDA).

Class Pycnogonida (Pantopoda) (sea spiders)

Marine chelicerates with body divided into a cephalon trunk and abdomen; cephalon with a tubular proboscis and typically 3 pairs of appendages; trunk extremely narrow, consisting of 4 segments, each with a pair of jointed legs; abdomen reduced to a small oval stump; no respiratory organs; no segmental excretory organs; length 2 mm–6 cm; about 600 living species.

Class Pentastomida (*Linguatulida*)

Elongated vermiform parasites with secondary annulation, 2 pairs of claws at sides of mouth; without respiratory or circulatory systems; adults live in lung, gut, and coelom of reptiles, birds, and mammals; larvae free-living or encysted in an intermediate host, often a fish; length, up to 9 cm; about 70 living species known (see ONCOPOD).

Class Tardigrada (water bears)

Small and widely distributed; 4 pairs of stumpy legs end in claws, oral stylets, and a suctorial pharynx; no definite circulatory or respiratory systems; length, usually less than 1 mm; about 350 living species known (see ONCOPOD).

Critical appraisal. Although there is a close evolutionary relationship between the Annelida and the arthropods, it is not known whether the phylum Arthropoda is monophyletic–derived from a single source–or has evolved from different annelid types that have subsequently undergone parallel evolution. If the latter is the case, the phylum must be polyphyletic. Opinion is not unanimous in favour of either view. Because of such uncertainties, the classification of the arthropods and their close relatives is a matter of debate. In particular, the affinities of the Onychophora, Tardigrada, and Pentastomida are by no means clear. For this reason they are sometimes included together; for a discussion of these groups as oncopods, see ONCOPOD. Alternatively, the Onychophora are often elevated to the rank of phylum, while the Tardigrada and Pentastomida are regarded either as proarthropods or as orders of Arachnida. Some authorities treat all three groups as distinct phyla.

The classification adopted above is based on that suggested by Lord Rothschild in 1965, except that the Collembola are treated here as a separate class rather than as an order of apterygote insects; for a discussion of Collembola as apterygotes, see APTERYGOTE. The Xiphosura and Eurypterida are combined in the above classification into a single class (Merostomata) and not regarded as orders of Arachnida. Some authorities increase the number of arthropod classes by using Merostomoidea, Marellomorpha, and Pseudocrustacea and including a number of fossil forms, more usually regarded as Trilobita or Crustacea.

The ancestry of the myriapods and insects can only be guessed since there are no fossils to indicate their origins. These groups resemble Crustacea in that they are mandibulate and have antennae. On the other hand, the Merostomata, Arachnida, and Pycnogonida feed by means of chelicerae that terminate in pincers and are reduced in spiders. The division of classes into the subphyla Mandibulata and Chelicerata is a fundamental one. The chelicerates not only lack antennae, whose tactile function is taken over by the pedipalps, but they have no distinct head; rather, the region is fused with the thorax to form a prosoma, or cephalothorax. In the Onychophora, myriapods, and insects a whole limb has become a mandible, and the tip is used for biting. The Crustacea chew with the aid of a structure called a gnathobase, assisted by other segmental appendages.

Although the Symphyla share a number of characters in common with apterygote insects (*e.g.*, styles, coxal sacs, spinnerets, total cleavage, embryonic dorsal organ), they differ in location of the gonopore and in number of abdominal segments.

BIBLIOGRAPHY. P.P. GRASSE (ed.), *Traité de zoologie*, vol. 6, *Onychophores, Tardigrades, Arthropodes, Trilobitomorphes, Chélicérates* (1949), a standard reference work in French, with an article on the anatomy, ecology, habits, and classification of arthropods; J.L. CLOUDSLEY-THOMPSON, *Spiders, Scorpions, Centipedes and Mites*, 2nd ed. (1968), a general account of the biology and ecology of terrestrial arthropods other than insects; A. KAESTNER, *Invertebrate Zoology*, vol. 2, trans. and adapted by H.W. and L.R. LEVI (1968), an excellent modern account of arthropods other than insects; R.C. MOORE, C.G. LALICKER, and A.G. FISCHER, *Invertebrate Fossils* (1952), a comprehensive textbook; R.E. SNODGRASS, *A Textbook of Arthropod Anatomy* (1952), a standard work on morphology; V.B. WIGGLESWORTH, *The Life of Insects* (1964), an excellent popular account; H. WOODS, *Invertebrate Palaeontology*, 8th ed. (1946, reprinted 1966), a discussion of invertebrate fossil remains, including arthropods.

(J.L.C.-T.)

Artiodactyla

The mammalian order Artiodactyla, or even-toed ungulates, includes the pigs, peccaries, hippopotamuses, camels, chevrotains, deer, giraffes, pronghorn, antelopes, sheep, goats, and cattle. It is one of the larger mammal orders, containing about 150 species, a total that may be somewhat reduced with continuing revision of their classification. Many artiodactyls are well-known to man, and the order as a whole is of more economic and cultural importance than any other group of mammals. The much larger order of rodents (Rodentia) affects man primarily in a negative way, by competing with him or impeding his economic and cultural progress.

GENERAL FEATURES

Abundance and distribution. Artiodactyls were once the dominant herbivores (plant-eating mammals) of almost every continent. They are an important link in the chain by which the sun's energy, having been used by green plants, is made available to other forms of life. They tend to be medium- or large-sized animals. If they were any smaller they would compete with rabbits and the larger rodents, and if they were larger they would compete with elephants and rhinoceroses, the largest of terrestrial herbivores. The success of artiodactyls has depended on skeletal adaptations for running and on the development of digestive mechanisms capable of dealing with plant foods; none is adapted to flying, burrowing, or swimming. The individual species tend to be fairly narrowly adapted, in comparison with other mammals, but many of them nonetheless have broad distributions.

Native artiodactyls are absent only from the polar regions and from Australasia, but many have been introduced into Australia and New Zealand. In Australia, the position of medium and large herbivores is occupied by kangaroos. Through most of its evolutionary history, the order was absent from South America; only within the last few million years have some groups entered that continent. The occurrence of the majority of living artiodactyls in the Old World is a recent phenomenon; a considerable variety once inhabited North America.

Families of living artiodactyls

The order Artiodactyla contains nine families of living mammals, of which the Bovidae (antelopes, cattle, sheep, and goats) is by far the largest, containing nearly 100 species. There are five Eurasian and four African species of pigs (family Suidae) and two Central and South American species of piglike peccaries (Tayassuidae). The two hippopotamus species (Hippopotamidae) are African.

The more familiar large species were until recently widespread throughout Africa south of the Sahara and in the Nile Valley; the pygmy hippopotamus has a restricted distribution in West Africa. The camel group (Camelidae) was formerly abundant in North America, the now extinct North American stocks having produced the camelids of South America (wild guanaco and vicuña, domestic llama and alpaca) and the Old World dromedary and Bactrian camel.

The remaining artiodactyls (*i.e.*, the suborder Ruminantia) are all ruminants (cud chewers), the most primitive of which are the chevrotains (Tragulidae), with three species in Asia and one, the water chevrotain, in West Africa; the chevrotains are clearly remnants of a group that was once more numerous and widespread. Deer (Cervidae) are basically Eurasian and have not spread into sub-Saharan Africa, although they have reached the Americas. There are about 30 species, the greatest number being concentrated in South America and tropical Asia. The giraffe and the okapi (Giraffidae), two distinctive African species, are closely related to deer. The pronghorn (Antilocapridae), although sometimes called pronghorn antelope, is not a true antelope; it is the only survivor of a stock of ruminants that was very successful in the later part of the Tertiary Period in North America (about 2,500,000 to 65,000,000 years ago). The family Bovidae is primarily African and Eurasian, with a few members in North America. Bovids are advanced artiodactyls, many of which live in open grassland and semi-arid areas.

Importance to man. Artiodactyls have long been exploited by man for economic purposes. At Olduvai Gorge in East Africa there is clear evidence of the use of antelopes for food almost 2,000,000 years ago. In Europe during Paleolithic times (about 30,000 years ago) Cro-Magnon man depended heavily on the reindeer. By this time the use of animals other than as food had become established; skins were used as clothing and footwear, and bones were used as tools, weapons, and accessories.

The domestication of animals was a major advance in human history (see DOMESTICATION, PLANT AND ANIMAL). Domestication of herd animals probably arose gradually, perhaps before agriculture. Domesticated goats and sheep are first known from the Near East at some date close to 7000 BC. Cattle and pigs were domesticated at some subsequent date but certainly before 3000 BC. In South America the llama, now used for transport, and the alpaca, which provides a source of wool, were developed from guanacos by the Incas or their predecessors. The dromedary (*Camelus dromedarius*), domesticated in Arabia, was introduced into the Southwestern United States, southwestern Africa, and inland Australia in the 19th century. A large feral population now exists in Australia.

In addition to providing meat, milk, hides, and wool, artiodactyls have served man in a number of other ways. In Kashmir, the underfleece, or pashm, of the Siberian ibex (*Capra ibex*) and of local domesticated goats has been used as the basis for the manufacture of cashmere shawls. In southwestern France, pigs have been used to locate underground truffles (the fruiting bodies of certain edible fungi).

No group of mammals is more extensively hunted than the artiodactyls. Sport hunting of various deer supports a multimillion-dollar industry in North America and Europe. In many cultures hunting has been reserved for monarchs or the aristocracy. In the centuries after the Norman Conquest of England, the forest law provided severe punishment for the slaughter of deer and boars. Père David's deer (*Elaphurus davidianus*) of China now survives only because it was preserved first in the hunting park of the emperors of China and later by the Duke of Bedford after the slaughter of the Chinese herds at the end of the 19th century.

Wild ungulates were the primary source of meat for human populations long before the appearance of modern man. Prehistoric man hunted the large mammals of his environment with an ever increasing effectiveness that was certainly instrumental in his survival. The extent to which man was involved in the extinction of some of the larger Pleistocene animals (*i.e.*, those that were abundant 10,000 to 2,500,000 years ago) is still being investigated. There is now known to have been a wave of late Pleistocene extinction of large mammals, including artiodactyls; in North America this wave reached its zenith about 9000 BC. Many animals also became extinct in Africa, where long-horned buffalo and large relatives of hartebeests survived until very recently. More of the large mammals have survived in Africa than elsewhere, but the reason for their survival is not known. A second, probably final, wave of extermination of the larger mammals has taken place with the spread of European culture and firearms in the past 300 years. It has been marked by wanton slaughter and has ultimately produced an interest in conservation. It now seems, however, that the unprec-

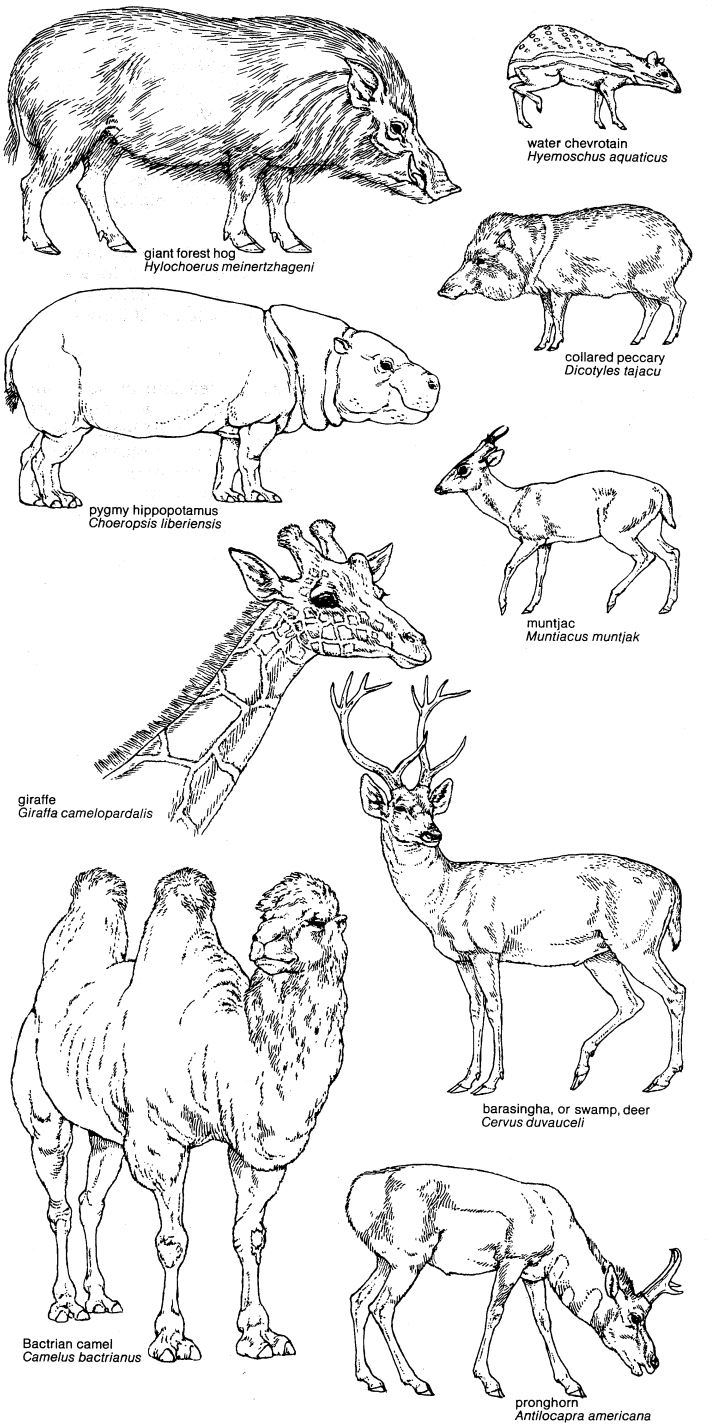


Figure 1: Representative non-bovid artiodactyls.
Drawing by R. Keane

edented demands on the environment being made by rapidly expanding human populations will result in a nearly complete extinction of large wild mammals.

Artiodactyls, like so many other animals closely associated with man, are a part of his religion and folklore, exemplified, for instance, by Aesop's *Fables* and the Persian worship of Mithras (see ANIMALS AND PLANTS IN MYTH AND LEGEND).

NATURAL HISTORY

Behaviour. *Migration.* Many artiodactyls undertake seasonal migrations between their breeding grounds and feeding areas or between different feeding areas. They can then take advantage of the seasonal changes in different areas. This means that larger populations, and hence a larger biomass (*i.e.*, the total weight of all individuals in an area), can be supported than if all passed their lives in one area. The North American mule deer (*Odocoileus hemionus*) comes from its summer pastures at high altitudes as the first snow falls and returns at the end of winter, several weeks after the snow has melted.

Seasonal changes

Social behaviour. Although the popular image of artiodactyls is one of great herds numbering thousands of individuals, some species are solitary, and many others form only small family groups. The maternal family unit, in fact, is the most cohesive one, providing the basis for herd formation. Most artiodactyls are more or less social, and grazing forms may be found in especially large aggregations. It appears that the practice of aggregating gives protection, favouring those members of the species that are the most active contributors to the gene pool (thus the most available to natural selection), since the individuals most frequently taken by predators are old, solitary males, males maintaining territories, and animals of either sex separated from the herd.

The advantage of herd behaviour

Social facilitation (the instigation of collective behaviour) takes place in herds. After one animal flees, all of the others flee, and the predator may thus not catch any. Social facilitation may also promote a restricted season for births; this helps survival of the young by denying these easy-prey individuals to predators through much of the year, and keeps the predator population lower than if young were available throughout the year. Another advantage of herding is that the older generation in a herd can guide migrations to water, feeding areas, or mating grounds.

Females and young are usually in herds separate from those of the younger males, but territorial (the older, proven) males may accompany the females. There are some variations of this behaviour. In the Eurasian roe deer (*Capreolus capreolus*), for example, the basic unit includes the doe, her litter of two, and often the young of the previous year. During the rutting (mating) season males associate with females in heat but do not gather harems. The female herds of red deer (*Cervus elephas*) are separate from the males except in the breeding season, when the stag will defend his female herd against other males. Among cattle and related species, the males associate with the females and young, but the bulls are ranked below a so-called master bull, each defending its place within the rank order. Female hippopotamuses and their young form a group in water and have a favourite resting and basking sandbank. The males have their resting places around this area. Each male's rank in the social hierarchy determines how close to the females he may be.

There can be some flexibility of social organization within a species. During the rutting season the male Rocky Mountain goat (*Oreamnos americanus*) makes little effort to herd females within a fixed area if there is little snow, but he does drive off other males. When there is much snow, he neither fights other males nor defends individual females.

Forest-dwelling artiodactyls often live singly, as does the okapi (*Okapia johnstoni*) of central Africa; individuals meet only for mating. Female moose (*Alces alces*) with calves are intolerant of their own young of the previous year and of adults, so even small herds do not form.

The territory of an animal is an area from which the possessor attempts to exclude other individuals of the same species (and occasionally other species). An animal in an area lacking its own scent is more timid and ready to flee. Among solitary artiodactyls the territory holder defends an area sufficient to meet his needs for food and shelter. Among social artiodactyls the territorial system is interwoven with breeding activities, and territories are normally defended only by certain males. Other males are driven off, and a percentage of males are prevented from mating.

Territoriality

The most simple territorial organization among artiodactyls is that of the common wild pig (*Sus scrofa*), which lives within a home range including resting, feeding, drinking, and wallowing places. There is little sign of territorial defense, and the herd (called the sounder) may move to a new area. At the other extreme, male Uganda

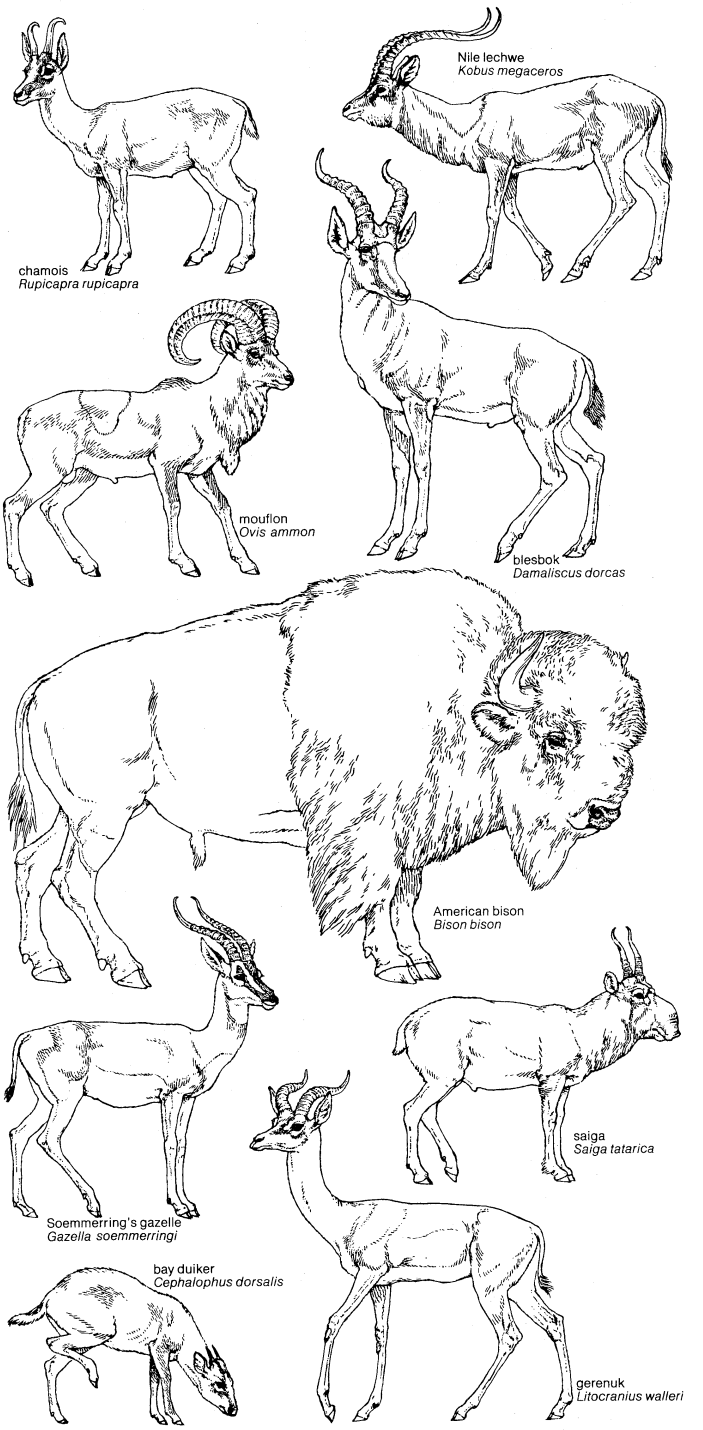


Figure 2: Body plans of bovid artiodactyls.
Drawing by R. Keane

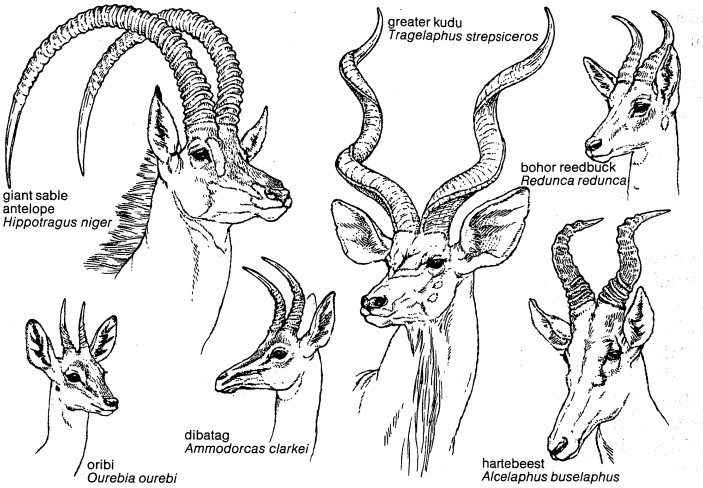


Figure 3: Horn variation in African antelopes.
Drawing by R. Keane

kob antelopes (*Kobus kob*) hold territories, for breeding only, that are as small as 15 to 30 metres (50 to 100 feet) in diameter. There are 30 to 40 territories on the breeding ground of a herd, and groups of females and young move about the territories despite the efforts of individual males to detain them. The semi-arid Serengeti plains of northern Tanzania contain nomadic aggregations of blue wildebeest (*Connochaetes taurinus*), males of which defend temporary territories only while an aggregation remains stationary.

In territorial defense an aggressive encounter between males is generally preceded by visual signalling of intentions. Chital deer (*Cervus axis*), for example, have several sorts of threatening displays. When sharp, potentially lethal horns appeared in early ruminants, intimidating displays rather than combats would doubtless have been favoured. Horns or antlers eventually functioned to maintain head contact during struggles rather than to bruise, slash, or gore. This stylized fighting, in which the competing males interlock horns or antlers and try to "outwrestle" each other, minimizes the danger of killing an opponent of the same species (conspecific). It evolved in two ways: further development of the wrestling, found in stags and some of the antelopes, and ramming, as in sheep. In sheep the horns are the sole organs of display. They increase in size throughout life and parallel the dominance order of the males, so that unnecessary fighting is minimized. Ramming may have intermediate forms; goats, for example, butt with a sideways hooking motion. In the fighting of hornless artiodactyls, such as pigs, the combatants may be badly mauled or even killed. The fighting behaviour of camels retains primitive elements of biting, kicking, and neck wrestling.

Reproduction. Many advanced artiodactyls have elaborate courtship behaviour, a regular component of which is for the male to sniff or lick the female's urine, and afterward to raise his head slightly with upcurled lips. This behaviour, which has been called *flehmen*, apparently enables the male to recognize females in heat. In the mating ceremonies of tragelaphine antelopes (kudus, bushbucks, and others) the male follows the female, nuzzling her neck several times. When he mounts, he lays his neck along hers so that their heads touch. In Thomson's gazelle (*Gazella thomsoni*), following the *flehmen* behaviour, the male runs close behind the female and finally taps her hindleg with his foreleg. Similar leg contact also occurs in some other antelopes. Its function could be to test the female's readiness to mate, to habituate her to contact, or to heighten her readiness to mate. It appears to be equivalent to the neck contact of tragelaphines. During mounting, the male Thomson's gazelle holds his head high and does not touch the female's flanks with his forelegs; the pair may continue walking. This is probably a more advanced pattern of events than that in tragelaphines. The kob antelope has elaborate displays after mating. These and the specialized sexual displays seem to be a consequence of this species' tightly clustered territories on the mating grounds. Another pattern occurs in the normally solitary Indian hog deer (*Cervus porcinus*); as many as 20 or 30 aggregate loosely in a certain area, then females and males leave in pairs and usually remain together until they have mated. Mating in artiodactyls often intensifies toward dawn and dusk.

Gestation periods

Gestation periods vary and are related in part to the size of the animal. They range from four months in the small chevrotain to 14 months in the Bactrian camel (*Camelus bactrianus*) and over 14 months in the giraffe. Females of normally gregarious species become solitary a few days before giving birth. The female chital, or axis deer, for example, remains near a patch of dense bush and high grass to which she can retreat if endangered. The female collared peccary (*Dicotyles tajacu*) withdraws to a burrow. The European wild pig gives birth in a rough nest.

In temperate regions, birth takes place in spring or early summer, and in tropical areas there are often more births during or just after the rainy season. The absence of a well-defined breeding season in a species may indicate less rigorous environmental conditions, which sometimes vary in different parts of a species' range. Warthogs have one restricted breeding season in most of eastern and southern Africa, while elsewhere two seasons or year-round breeding have been recorded. The breeding season of the waterbuck (*Kobus ellipsiprymnus*) is continuous in Uganda, but in Zambia there is a sharp peak at the height of the rains.

Most modern artiodactyls have one young at each birth, but there are some well-known exceptions among ruminants. The Chinese water deer (*Hydropotes inermis*) bears twins or triplets, but during gestation carries even more fetuses; early records (now known to be incorrect) of large litters were based on observations of dead pregnant females containing the large number of fetuses. The mule deer, white-tailed deer (*Odocoileus virginianus*), roe deer, pronghorn (*Antilocapra americana*), nilgai (*Boselaphus tragocamelus*), four-horned antelope (*Tetracerus quadricornis*), and saiga (*Saiga tatarica*) commonly bear twins. In the white-tailed and mule deer and in the saiga, a higher percentage of twins are borne by the older females; this is probably true in other species. The number of young is usually three in the warthog, five in the European wild pig, and two in peccaries.

The female wild pig almost ignores her young, which free themselves from their birth membranes and seek a teat. Female camels show comparatively little maternal attention and do not eat the afterbirth (the fetal membranes and placenta). Ruminants generally eat the afterbirth, as well as the dung and urine of the young, thus helping to prevent discovery of the young by predators. Licking of the young tends to facilitate its recognition

by the mother. An artiodactyl is normally precocious (well developed) at birth and may weigh one-tenth as much as its mother. An extreme example of precocity is the wildebeest calf, which rises within five minutes of birth, follows its mother within another five minutes, and can move as fast as an adult in 24 hours. Young deer fawns "freeze" during danger but rejoin the herd when the danger is long past or when retrieved by the mother.

Pigs and hippopotamuses are weaned after a few months, but among higher artiodactyls, lactation lasts longer. Wildebeest, for example, suckle for almost a year, although they start to eat grass when only a few days old. This may either maintain a bond between parent and offspring and form the base for larger social groupings or help to "develop" the four-chambered stomach. Higher artiodactyls eat soil when they begin to eat solid food, probably to establish a normal flora and fauna in the rumen (the first of the four stomach chambers).

Locomotion. Artiodactyls are preyed upon by carnivores and therefore need speed and agility to escape death. They have an added disadvantage in the sheer weight of their very large stomachs, which they need in order to digest plant food. Running ability reaches an extreme in advanced artiodactyls living in open country. The hippopotamus, with an adult weight of 2,500 to 3,000 kilograms (2¾ to 3⅓ tons), is the only living artiodactyl big enough to need heavy, pillar-like limbs for support.

In the normal walking of artiodactyls the legs move in the following order: (a) left front, (b) right rear, (c) right front, (d) left rear. This basic pattern is masked in faster walking or trotting by each foot being lifted off the ground before the one ahead of it in the sequence reaches the ground, resulting in telescoping the first (a and b) and second (c and d) pairs of movements. In galloping or fast running the two front legs leave the ground one immediately after the other, then the two back legs. The chief propulsive force in locomotion comes from the back legs, except in the giraffe (*Giraffa camelopardalis*), in which the front legs provide the main propulsive power.

Camels often amble, both legs of each side moving together, and the giraffe and the okapi always use this walking gait. Here the middle two (b and c) and the first and last (a and d) actions of the normal walking pattern occur together. The giraffe, having a short body and great height, could not adopt the normal ruminant gait without tripping. The long neck moves back and forth in time with the strides and helps smooth the movement. Galloping by the giraffe is of the normal ungulate type.

Artiodactyls living among bush or rocky cover may develop a bounding sort of gait in which the legs are pulled up very sharply during each stride. Deer and some antelopes are examples. When walking, species in such habitats are supported by the diagonally opposite legs for a greater length of time in each stride than are fast-running, open-country ruminants. This is a more primitive stable position and allows an easier leap from hidden danger. Some bovids, notably goats in Eurasia and the klipspringer (*Oreotragus oreotragus*) of Africa, are especially agile on rocky slopes and precipitous ground.

Maximum speeds of locomotion

The maximum speeds of some artiodactyls are: warthog, 48 kilometres (30 miles) per hour; camel, 14–16 kmph (9–10 mph); giraffe, a little over 48 kmph (30 mph); Cape buffalo (*Syncerus caffer*), 56 kmph (35 mph); Thomson's gazelle, 80 kmph (50 mph).

Ecology. *Food habits.* Most artiodactyls are closely tied to the resources of their environment. They are dependent, for example, on feeding areas not being covered by too much snow or shrivelled under a drought, and on the regulating effects of fire or other herbivores on the seasonal succession of vegetation. Various grazing species feed on grass at different heights. Browsers, those that feed on the foliage of shrubs and trees, show more extreme variation in feeding height, the maximum being that of the giraffe.

Herbivorous animals need less initiative and intelligence to collect food than do the meat-eating, hunting carnivores, but digestion is more difficult. Advanced artiodactyls have evolved the ability to bolt food and to ruminate it (chew it more thoroughly) at a later time or while resting in an area where they may be less obvious to predators and can conserve energy. Tropical artiodactyls frequently have adaptations for water conservation, having developed to a high degree internal physiological regulation (homeostasis).

Primitive artiodactyls were probably omnivorous but favoured plant foods, a characteristic still found in pigs. The latter dig with the snout and, to a lesser extent, with the front legs and upper tusks (canine teeth). The warthog of Africa (*Phacochoerus aethiopicus*) has a modified method of gathering food. When food is scarce it forages for young grass shoots under very low bushes; its tusks and localized thickening on its skin protect the eyes and muscles from thorn damage, and small incisors enable it to pluck food.

Hippopotamuses (*Hippopotamus amphibius*), although they spend a great deal of time submerged in lakes or rivers, do not feed in the water. They graze at night, wandering over well-used trails, sometimes far from water, often damaging crops.

Most members of the camel family are found in arid habitats. The vicuña (*Lama vicugna*) of the South American Andes lives at high altitudes where it grazes on soft grasses and herbs. It has much the same food requirements as domestic sheep.

Chevrotains live in dense undergrowth close to water or in marshes, where they browse on soft vegetation, roots, and tubers, following a way of life probably not unlike that of their ancestors.

The other ruminants browse or graze. They may take many plant species in the course of the year, but at any one season a large part of the diet consists of only five or six plants. Some ruminants are strongly specialized. The reindeer of the Arctic (*Rangifer tarandus*), for example, eats a variety of sedges, grasses, and herbaceous plants in summer but, as the long winter approaches, gradually shifts to a diet of lichens. It uses its front feet to scrape snow away from lichens to a depth of about 60 centimetres (two feet). The females are unique among deer in possessing antlers, which are thought to help them get scarce food in late winter by driving off the males that have by then shed their antlers. Reindeer may eat lemmings. The red deer, on the other hand, has catholic feeding habits. In woods it browses on lichens, berries, fungi, and the leaves of most deciduous trees; in open country it eats grass, heather, berries, and lichens. Shrubs and trees are used more in winter. When the red deer lives in the same areas as other ruminants it can be a serious competitor for food.

The importance of grasses in the ruminant diet

Grasses form a substantial part of the diet of many ruminants. Young grass consists of about 5 percent protein, 1 percent fat, 3 percent minerals, and 20 percent carbohydrates; the remaining percentage is water. The most noticeable changes as grass ages are an increase in carbohydrate content to 75 percent and a large decrease in the amount of water. Such food, especially when coated with silica, as are many grasses, or when covered with dust, would be impossible for nearly all nonruminant herbivores to eat or digest. The major evolutionary trend in ruminants has been to make use of grasses, and grasslands and the higher ruminants have evolved largely in adaptive balance with one another. This adaptive balance was shown during a study of the change from plains to thickets of scrub growth in an area in the eastern Congo over a period of about ten years. There was an accompanying decrease in numbers of antelopes and warthogs, no change in buffalo, and an increase in elephants and hippopotamuses.

There is not usually a one-to-one dependence of any artiodactyl species on one plant. The plant species that constitute the major part of the diet may vary with the season, and similar parts of different plants may be eaten in preference to other parts of the same plant. Food resources in an area are thus parcelled out among the various artiodactyls present. Sometimes behavioral differences minimize competition between closely related

species in the same area. A study has shown that in central Africa the roan antelope (*Hippotragus equinus*), a grazer, favours open areas with taller, ranker perennial grasses and is more or less sedentary within a small area; the sable antelope (*H. niger*), also a grazer, prefers savanna woodland or the edges of open areas, and herds follow a more or less cyclic annual route over an area of about 200 square miles. When pasturage is restricted, sheep will cut grass very short, and goats will damage trees and bushes. An American zoologist, George B. Schaller, has observed that, in Kanha Park in central India in the hot season, blackbuck (*Antilope cervicapra*) continue to graze on grass shoots in open areas; chital deer seek out tender grass blades, especially along forest edges, and also feed on leaves and fruits; barasingha (*Cervus duvauceli*) eat dry and moderately coarse grass along ravines; sambar deer (*Cervus unicolor*) browse on leaves and crop coarse grasses in the forest; and gaur (*Bos gaurus*) graze on tall, coarse grass and break down saplings to get at the leaves. The choice of habitat also varies: chital avoid steep terrain and forests with an unbroken canopy; blackbuck require less water than the others and thus remain in drier regions; sambar and gaur are less specialized in habitat requirements, and both are active primarily at night; barasingha prefer reed beds but also enter forests and climb hills.

It has also become evident that grazing successions are one of the mechanisms that enable the maximum use to be made of environmental resources. On the Serengeti plains, for example, the wildebeest grazes on ground already covered by the zebra and leaves the grazed grass in a condition suitable for the Thomson's gazelle. Interactions take place between artiodactyls and some plant species. It has been noted in the Tarangire area of northern Tanzania that *Acacia* seedlings germinate only where the impala (*Aepyceros melampus*) has left its dung. In parts of southern Peru plants growing on or close to the dung of the vicuña are different from those of the surrounding pasture.

Artiodactyls often favour the boundary zone between habitats. In Rhodesia, Lichtenstein's hartebeest (*Alcelaphus lichtensteini*) is usually found at the edge of clearings adjacent to woodland.

Areas of distribution. Some artiodactyls have surprisingly small ranges; Hunter's hartebeest (*Beatragus hunteri*) and the dibatag (*Ammodorcas clarkei*), for example, are found in two very restricted areas in eastern Africa. Others have extremely large ranges, such as the roe deer, which lives from the western shores of Europe to the eastern shores of Asia, or the red deer, which is found in a similar band across Eurasia and is regarded by many as conspecific with the North American wapiti or elk (otherwise called *Cervus canadensis*). Sometimes a considerable area may be occupied by a chain of related species, an example being the oryxes; the beisa and gemsbok (races of *Oryx gazella*) occur in South and East Africa, the scimitar-horned oryx (*O. dammah*) in West Africa, and the Arabian oryx (*O. leucoryx*) in Arabia.

The influence of climate on the distribution of artiodactyls

It is well known that climate is one of the factors limiting the ranges of artiodactyls. A number of South African antelopes differ, at the species level, from their ecological counterparts farther north in Africa. The bontebok and blesbok, races of *Damaliscus dorcas*, are found in the south and the sassaby (*D. lunatus*) farther north; the black wildebeest (*Connochaetes gnou*) occurs in the south and the blue wildebeest (*C. taurinus*) farther north. This probably is a result of climatic or climatically influenced factors; each species evidently functions best in a certain temperature and aridity range. Wide distributions can occur more easily along lines of latitude than they can by spanning the tropics to temperate or polar regions. Species that cross lines of latitude are often associated with mountain chains, examples being the Rocky Mountain goat, with its wide latitudinal range in western North America, and the goral (*Nemorhaedus goral*), found from Indochina to the Amur River. Climatic effects on distributions sometimes occur with regard to altitude. In Central Asia, the goa (*Gazella picticaudata*) is found in valleys from 10,000 to 12,000 feet above sea level, the chiru (*Pantholops hodgsoni*) and the yak (*Bos mutus*) are on the very high steppe between 18,000 and 20,000 feet.

South America has a more impoverished artiodactyl fauna than Africa, being limited to deer and camelids. This arises in part from the late arrival of the artiodactyls (deer in middle to late Pliocene, about 4,000,000 years ago, camelids perhaps a little later) and in part because a number of large rodents compensate for the shortage of large herbivores. The cervids in South America have not shown the same capacity for radiation in open country as have bovids in the Old World.

The areas of distribution and numbers of individuals are determined by complicated interweaving of effects not yet completely understood. Bloodsucking flies are thought to be the main reason that red deer in Scotland ascend to higher feeding grounds in June, and reindeer are afflicted by horse flies (*Tabanus*) and other dipteran pests. It is questionable whether the level of artiodactyl populations is controlled by predation, by availability of food, by reproductive rate, by disease, by climate, or by competition, insofar as these can be regarded as separate factors. It is known that undernourishment increases the susceptibility of an animal to the effects of parasites. If such an infected animal, say a pig, is caught by a leopard, it would be an oversimplification to assign a single reason for its death; it could have died from starvation, parasites, or predation. There is no evidence that artiodactyls are affected more than marginally by predators during most of their mature lives. Mortality is greatest among juvenile and aged animals. In a study of central African warthogs, it was estimated that a 60 percent loss occurred during the first six months of life in an expanding population and 95 percent in a declining one. Although predation was thought to be the main cause, another was the fact that the piglets had only limited control over their body temperatures and were thus more at the mercy of environmental temperature change. Food supply may sometimes be decisive, either directly or through the indirect action of intermediate agencies such as drought. The year 1961 lacked long rains, causing a severe shortage of forage in the Nairobi Game Park in Kenya. Many antelopes died of starvation, populations fell, and those of the blue wildebeest had not recovered nine years later, perhaps for reasons unconnected with the initial drought. Disease has generally been considered to have only a secondary importance in regulating numbers.

Thickness of the snow cover in winter is a very important factor for Asian artiodactyls. The saiga, for example, cannot move in snow deeper than about 40 centimetres (16 inches), and the wild sheep *Ovis ammon* in snow deeper than 60 centimetres (24 inches), at the most. The snow may have other effects; a layer of ice on top of snow may damage an animal's legs and weaken the animal to the extent that it is caught by a predator. Saiga may be unable to dig through even a shallow layer of compacted snow. Hoarfrost on vegetation is especially dangerous when prolonged or when it occurs in consecutive winters, though elk may escape the worst effects by feeding in winter on bark and high shoots. Massive periodic mortalities among Palearctic (Eurasian) ungulates in winter have been known since ancient times. The saiga has adapted to these crises by migrating great distances in a short time away from snowstorms or from areas where fodder is short. It also has a very rapid maturation to a reproductive state, ensuring that populations will build up after heavy mortalities.

Population density over the range of a species is affected by social behaviour, such as the effects of territoriality, dispersal of the young, and whether the species lives in herds. Fecundity may be reduced in overcrowded conditions by effects on reproductive control mechanisms, reduced viability of the young, or retarded maturation.

FORM AND FUNCTION

General structure. Artiodactyls have larger stomachs and longer intestines than carnivorous animals because

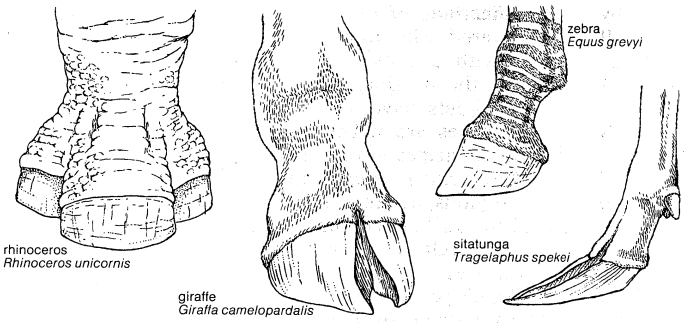


Figure 4: Comparison of even-toed (artiodactyl: giraffe and sitatunga) and odd-toed (perissodactyl: rhinoceros and zebra) ungulate feet.
Drawing by R. Keane

plant food is less easily digested than meat. The necessity of escaping predators and the handicap of a heavy digestive system have resulted in limb bone adaptations.

Comparison of odd- and even-toed ungulates

In all artiodactyls the main weight-bearing axis of the leg passes through the third and fourth toes together. This has been called paraxonic support and is contrasted with the mesaxonic limb support of the other great order of herbivorous mammals, the perissodactyls (rhinoceros, horse, tapir), in which the weight-bearing axis passes through the third or central toe alone. As artiodactyls evolved there was increasing development of the third and fourth toes and a parallel decline of the second and fifth toes flanking them. Progressive simplification of limb extremities has characterized their evolution, and even in the earliest known artiodactyls, the pollex and hallux (corresponding to the big toe and thumb of man) were already rare.

The other main morphological characteristic of artiodactyls is that the astragalus, one of the bones in the ankle, has upper and lower rounded articulations (areas of contact of bones) and no constricted neck, instead of simply one rounded articulation above a neck, as in other mammals. This character is so basic to artiodactyls that it has not developed very much within the known history of the order, having already been present in long extinct members. The artiodactyl astragalus also has an articulation on its rear surface for the calcaneum (heel bone). The three articulations are in nearly parallel planes, allowing the astragalus to rotate vertically.

Other features of the limbs, skull, and dentition distinguish artiodactyls. The ulna (posterior forearm bone) and fibula (posterior bone of the lower leg) have become reduced. The humerus, the upper bone of the foreleg, is large and has a large protrusion, the greater trochanter, to which muscles are attached. The femur, the upper bone of the hindleg, has a large greater trochanter and a second, lesser trochanter, but lacks the third trochanter characteristic of perissodactyls. There are typically 19 thoracic and lumbar (upper and lower back) vertebrae. The separate lumbar region of the spine is retained with its forwardly directed transverse processes (lateral projections on the vertebrae). There is no clavicle, or collarbone, in the shoulder girdle. The hip girdle shows fore-and-aft elongation and a well-developed ischium (upper anterior bone of the pelvis). There is never a penis bone.

The large tongue is very mobile and can be thrust forward. The brain is moderately developed, with folding of the surface of the cerebral hemispheres variably developed, often less in small artiodactyls than in large ones. The olfactory region of the brain is well developed and hearing is acute. The brains of earlier artiodactyls, such as the extinct entelodonts, were smaller than those of later forms. There are often scent glands on the head and body.

Specializations of the head. The skulls of pigs and peccaries lack a complete bony bar behind the eye (postorbital bar) as in most suiform artiodactyls and the early camels. The hippopotamuses, most camels, all ruminants, and two fossil suiform groups (entelodonts and oreodonts) have a complete postorbital bar. Any surface exposure of the periotic bone (bone around the ear) on the skull is called the mastoid, and skulls without such a surface exposure are described as being amastoid. Amastoid skulls are found in most suiform groups (including entelodonts, anthracotheres, and all living suiform groups); mastoid skulls occur in some early suiform groups, oreodonts, and all remaining artiodactyls that have lived since the end of the Eocene Epoch (about 38,000,000 years ago). Hippopotamuses have many modifications for aquatic life—large lungs, eyes and nostrils on top of the head, nostrils that can be closed by muscular control, and small ears. They are able to remain submerged for at least five minutes.

The distinction between horns and antlers

Horns and antlers. Pigs, peccaries, hippopotamuses, camels, and chevrotains have no horns or antlers. In the early Miocene, Old World ruminants related to giraffes and deer first developed such appendages. The majority of deer have antlers, defined as solid, bony, branched outgrowths of the frontal bones, present only in the males (but also in female reindeer) and shed seasonally. They are not covered by a horny sheath but, during a growth period of about four months, have a fine-haired skin or "velvet." The antlers have two basic branches, the anterior or brow tine, and the posterior branch or beam. The brow tine is unbranched, except in Père David's deer, in which both it and the beam are branched, the brow tine forming the dominant part of the antler. Antlers are specialized sex characters used for fighting by males in the rutting season and to scrape or slash at trees and bushes for territorial marking.

A study of the chital deer showed that antlers increase in size up to the seventh year, remain at a constant size until the ninth year, then decline. The horn of bovids consists of a hollow, unbranched horny sheath (formed of modified skin like fingernails and toenails) that fits over a bony core; horns are often present on both sexes. If such a horn is accidentally lost it is not regenerated; this is unlike the situation in deer, in which normal shedding is followed by regrowth. In the giraffe, but not in the okapi, horn growth is mainly from the parietal bone. The pronghorn has horns in both sexes. The sheaths are shed each year after the breeding season, and new ones develop under the old ones. The sheath is two pronged, but the underlying bony core is unbranched.

Teeth. There is a complete set of teeth in early artiodactyls and in modern pigs of the genus *Sus*, consisting on each side of three upper and lower incisors, an upper and lower canine, four upper and lower premolars, and three upper and lower molars. There has been a tendency toward reduction of the front teeth and development of a gap (diastema) between them and the back teeth. There has been very little tendency for the premolars to molarize, and the first premolar often disappears. Early forms had five-cusped upper molars, but the fifth cusp (protoconule) disappeared early.

Members of the suborder Suiformes have the full complement of incisors and canines, except for peccaries, which lack the lateral pair of upper incisors. Hippopotamuses have continuously growing incisors and canines, the lower canines being very large.

The canines of pigs grow continuously. In this group the canines are weapons for offense and defense, the sharp

cutting edges of the lower canines being maintained by wear against the uppers. Young camels retain the full complement of front teeth, with three incisors and one canine in the upper and lower jaws; the upper incisors are extremely small. In the upper jaw of the adult only the rear incisor and canine are present. The vicuña has continuously growing lower incisors.

The molars of pigs are low crowned (except those of the warthog) and have many cusps; those of peccaries are more simple. Peccaries have one less premolar than pigs; camels also have reduced premolars. Chevrotains have rather flattened lower premolars but have incipiently selenodont molars; *i.e.*, in which the cusps are drawn out into longitudinal crescents. Premolars of ruminants are wider, and the molars definitely selenodont. In many bovids and the pronghorn, but not in giraffes or deer, the molars are markedly high crowned.

Limb adaptations for fast running. Adaptations for fast running reach an extreme in advanced artiodactyls living in open country. In addition to the increased rotation of the astragalus, which increases the propulsive thrust at the ankle and enables a quicker recovery at the end of a stride before starting the next one, there are other features that help to increase the speed of striding. The legs of most camels and ruminants have lengthened, especially in the lower parts; the number of toes, or digits, in the feet is reduced from the original mammalian five, and ruminants walk on the tips of their toes. The muscles are inserted high on the legs; only tendons pass lower, so that a large mass is not concentrated near the tip of the limb, where its inertia would restrict speed of movement. Muscle contraction is fast. The movement of each leg is almost limited to a fore-and-aft plane. Emphasis on the fore-and-aft articulations between the limb bones is especially pronounced in many bovids, the alternating bones in the wrist (carpus) and ankle (tarsus) taking the strain of impact on uneven ground.

Pigs have four toes on each foot, but only two of them touch the ground. Their limbs are short and not very advanced. Peccaries have lost the outer accessory hind hoof in the back leg. All four toes of each foot of hippopotamuses touch the ground, and the terminal phalanges have nail-like hoofs. The toe bones of camels are completely enclosed in hardened, horny hoofs, and lateral toes spread across the broad pad which aids in walking on desert sands. Chevrotains have four hoofed toes on each foot; deer often retain the first and second phalanges (sections) of their lateral toes; but all bovids have lost the bones of their lateral toes.

The fibula bone in the back leg and the ulna in the front leg have been reduced in different artiodactyl lineages. Both are still complete in pigs and hippopotamuses, although the fibula is slender. In most other artiodactyls, the lower end of the fibula has survived, and the upper end is occasionally found, but always less noticeably. In camels the ulna has fused with the radius. Pigs, hippopotamuses, and camels have separate navicular and cuboid bones in the ankle, and magnum and trapezoid bones in the wrist; other artiodactyls have a fused naviculo-cuboid and magnum-trapezoid. In chevrotains and some deer, the adjacent ectocuneiform is sometimes joined with the naviculo-cuboid.

Modifications of the foot

The artiodactyl method of limb support through the third and fourth toes, with the attendant lengthening of lower limb bones, has frequently led to a fusion of the two principal metacarpal and metatarsal (midfoot) bones in the forelegs and hindlegs, respectively, forming cannon bones. The nearest approach to a cannon bone in the living Suiformes is the proximal fusion (*i.e.*, at the upper ends) of the two central metatarsals in peccaries. Camels have front and rear cannon bones, but the fusion does not extend right to the bottom, the lower articular surfaces being less pulley-like than in ruminants. There is a hind cannon bone in all chevrotains and, in addition, a front one in Asiatic species (*Tragulus*). All other living artiodactyls have front and rear cannon bones. Lateral metatarsals and metacarpals survive in chevrotains; splints of lateral metacarpals often survive in bovids; and either upper or lower splints of metacarpals in deer.

Modifications of the skin. *Hair and coloration.* Pigs are covered with rather sparse, coarse hairs, and peccaries with a denser coat of coarse hairs. Except for those of the warthog and the babirusa (*Babyrousa babirussa*), piglets have longitudinal stripes or flecks. Hippopotamuses are naked. Tragulids have light-coloured flecks and stripes in their fur. The coats of camelids and deer are much thicker in species living toward the polar regions, at great heights, or in deserts, but are not noted for striking colours or patterns. Many young deer and the adults of a few species have pale flecks and stripes, and some South American deer have reddish fur. Antelopes have a wider range of coat colours, and some are strikingly marked; *e.g.*, the oryxes, bontebok, and blesbok of southern Africa.

Scent glands. External glands occur in various places on artiodactyls. Preorbital glands, immediately in front of the eyes, are present in the giant forest hog (*Hylochoerus meinertzhageni*), in all cervids except the roe deer, and, among the bovids, in duikers, many neotragines, gazelles and their allies, and the hartebeest group. These glands are apparently required in small forest forms and have disappeared in many, but not all, open-country forms. In some, the glands are definitely connected with territorial marking; a firm object is marked by rubbing, soft vegetation by swinging the head gently from side to side. Foot, or pedal, glands are present in the African bush pig (*Potamochoerus porcus*), camels, tragulids, the pronghorn, some bovids, and on the back legs only of most American deer.

Inguinal (belly) glands are found in bovids, there being two in sheep, saiga, chiru, gazelles, duikers, and blackbuck, and four in members of the tribes Reduncini and Tragelaphini. Carpal (wrist) glands are present in some pigs, some gazelles and allies, and the oribi (*Ourebia ourebi*). Glands in other positions are rather less frequent, but postcornual ones (behind the horns) occur in the Rocky Mountain goat, the pronghorn, and the chamois (*Rupicapra rupicapra*), supraorbital ones in muntjacs (several species of *Muntiacus*). There are jaw glands in the pronghorn; neck glands in camels; dorsal glands on the back of peccaries, pronghorn, and springbok; and preputial glands (in front of the genital region) in several pigs, grysbok (*Raphicerus melanotis*), and the musk deer. Tail glands are found in musk deer, pronghorn, and goats; tarsal glands in pronghorn and American deer; and metatarsal glands in camels, some deer, and the impala. Pronghorn, blackbuck, gazelles, and oribi are thus particularly well equipped with glands. The use of such glands, apart from the use of preorbital glands in some species for territorial marking, is a matter for conjecture. Chital deer, when alarmed, thump the ground several times with their hind feet, which possess glands; the scent remaining on the ground may function as a danger signal. In general, mammals often mark with their glands when they are threatening other individuals of their own species.

Digestive system. The higher artiodactyls feed only on plant matter, which consists largely of cellulose and other carbohydrates and water. This necessitates adaptations of the structure and functioning of the stomach and intestines. Even pigs have enlarged stomachs—they have a pouch near the cardiac orifice (the upper opening) of the stomach—and in peccaries the stomach is more complicated. In hippopotamuses the stomach is divided into four compartments, and micro-organisms ferment food as part of the digestive process. Unlike pigs, hippopotamuses have lost the cecum (a blind pouch) further on in the gut.

The multichambered stomach

In the most advanced ruminants, the much enlarged stomach consists of four parts. These include the large rumen (or paunch), the reticulum, the omasum (psalterium or manyplies)—which are all believed to be derived from the esophagus—and the abomasum (or reed), which corresponds to the stomach of other mammals. The omasum is almost absent in chevrotains. Camels have a three-chambered stomach, lacking the separation of omasum and abomasum; the rumen and reticulum are equipped with glandular pockets separated by muscular

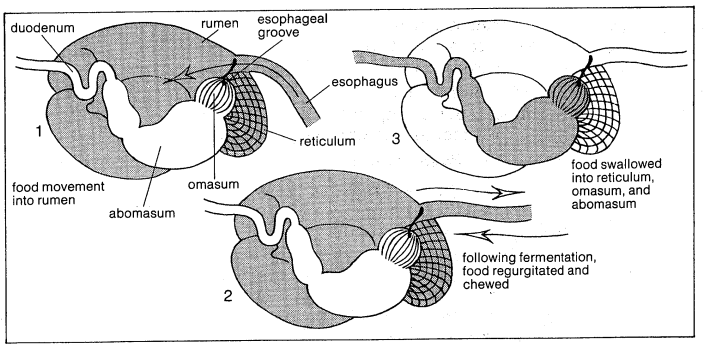


Figure 5: Stages of digestion in the four-part stomach of a representative ruminant.
Drawing by R. Keane

walls having sphincters (valves) and glands. The esophagus opens into the rumen, not into the area between rumen and reticulum; these and other differences suggest that camels evolved the ruminating habit independently of the true ruminants. The total stomach of the domestic ox (*Bos taurus*) occupies nearly three-quarters of the abdominal cavity, and, even in medium-sized cattle, the rumen alone can have a capacity of 25 to 75 gallons, having undergone a tremendous growth in early life, with the changeover from a milk diet.

Food taken into the rumen is later regurgitated into the mouth and completely masticated, then swallowed again and passed to the reticulum, omasum, and abomasum. The regurgitation and chewing in the mouth is called rumination.

In the rumen many different species of minute protozoans (animals) and bacteria live without free oxygen. The digestion of the cellulose of plant cell walls is the main function of the fauna and flora in the rumen, since mammalian digestive juices are incapable of digesting cellulose. The contents of the plant cells are thus released for digestion. Large volumes of saliva are secreted into the rumen to help digestion. Soluble products of microbial action, mainly fatty acids, are absorbed through the rumen wall. In the omasum, some fatty acids and 60–70 percent of the water are absorbed; in the abomasum gastric juice containing hydrochloric acid is secreted, as in an ordinary mammalian stomach.

In the rumen any ingested protein is degraded into fatty acids and ammonia; the ammonia and other simple nitrogen-containing substances are used by the micro-organisms for their own cell-protein synthesis. These organisms are ultimately digested in the abomasum and small intestine, thus providing the ruminant with protein.

Many artiodactyls are adapted to living in conditions of water shortage. The best known and one of the most spectacular examples of this is the camel. Its body temperature can fluctuate according to the outside temperature, thus minimizing water loss through sweating; it excretes rather dry dung and a concentrated urine (*i.e.*, high in urea and low in water) and is not seriously weakened by as much as a 25 percent dehydration in its body, since water is not withdrawn from the bloodstream and the continuing circulation avoids any buildup of excessive internal temperatures. The thick coat hinders the inward transference of heat from the environment (the temperature of which may often exceed the animal's body temperature); a thirsty camel can take in water very rapidly. Oryxes and gazelles are antelopes noted for needing little water, the dorcas gazelle (*Gazella dorcas*) in the Sudan depending on leaves of *Acacia* bushes for its water. The zebu (a form of domesticated cattle) needs less water than most temperate climate breeds.

Reproductive specializations. The testes of male artiodactyls descend outside the body cavity but may regress into the abdomen in the nonbreeding season. Female pigs have many teats, but ruminants have only two to four (although domestic cattle occasionally have as many as six). Among the bovids, the alcelaphines (hartebeests, wildebeests, and relatives), gazelles, and some caprines (sheep, goats, and relatives) have two, the rest have four.

The unborn mammal within its mother breathes, feeds, and excretes through an organ called the placenta, which is connected with the tissues of the mother's uterus (womb) wall. Hippopotamuses and pigs have an epitheliochorial placenta, a layer of fetal tissue merely pressed close against the uterus wall, but camels and ruminants possess a syndesmochorial placenta, in which the epithelium of the maternal tissues is eroded to facilitate intercommunication. This is an advance over the epitheliochorial placenta, but the artiodactyls are not particularly advanced, when compared with other mammals, in which there may be still closer association of maternal and fetal blood vessels (endothelial and hemochorial placentas). Even in many syndesmochorial placentas the uterus lining may be wholly or partly restored before the end of pregnancy. Although there is no erosion of maternal tissues in the epitheliochorial placenta, the capillaries beneath the fetal and maternal surface layers may pass just beneath the surface layers, making them thin. The actual fingerlike processes (villi), through which the placenta contacts the uterus, are evenly distributed ("diffuse" placentas) in hippopotamuses, pigs, camels, and tragulids; in higher artiodactyls they are in pockets or groups called cotyledons ("cotyledonary" placentas). It is interesting that there are few of these cotyledons in deer—for instance only five in Père David's deer—but many in giraffes and bovids (up to 160 or 180 in giraffes and goats). The musk deer (*Moschus moschiferus*) is exceptional among deer in retaining a diffuse placenta.

The epitheliochorial and syndesmochorial placentas

EVOLUTION AND PALEONTOLOGY

The artiodactyls can be traced back to a probable descent from a group of early generalized mammals called condylarths, and were certainly distinct by the Eocene Epoch, which ended about 38,000,000 years ago. Fossil artiodactyls can be more or less convincingly classified in three suborders; the more primitive Suiformes, centred around pigs, the Tylopoda, centred on camels, and the Ruminantia or ruminants. The most primitive artiodactyls are the suiform group Palaeodonta, which had four functional toes on each foot, primitive, low-cusped cheek teeth, and the typical artiodactyl astragalus. The artiodactyls became more prominent in the Oligocene (between about 38,000,000 and 26,000,000 years ago) with a decline of the then dominant perissodactyls, and the later history of artiodactyls appears as successive waves of groups, each better adapted than its predecessors to the changing environment. In the suiform line, the earlier palaeodonts are succeeded by other groups such as the entelodonts, giant "pigs" of the European and North American Oligocene, characterized by very large skulls (some nearly a metre long), very small brains, and a large, bony flange below the eyes. The functionally two-toed ruminants succeeded four-toed suiforms in the Miocene, and within the Old World ruminants of the bovid subfamily Caprinae, the zenith of the tribe Caprini, for example, followed that of the mainly Pliocene tribe Ovibovini.

The earliest artiodactyls

The artiodactyls had an interesting history in North America through the Tertiary Period. Some forms, such as the entelodonts, were shared with the Old World, but

others were characteristic of North America. One very prominent New World family was the merycoidodonts (or oreodonts), which lasted until the early Pliocene (about 6,000,000 years ago). They had somewhat piglike proportions, short faces, a large upper canine and a caniniform first lower premolar, and selenodont molars. A close relative, *Agriochoerus*, had clawed feet, the function of which remains uncertain.

Camelids evolved in North America and, at or toward the end of the Tertiary, spread into South America and into the Old World. By the end of the Pleistocene they all became extinct in their homeland, just as horses did. The hypertragulids were a mainly Oligocene group of chevrotain-like forms related to the Protoceratidae. The latter had horns above their noses, a position unique among artiodactyls, as well as in the usual position. The North American Miocene (26,000,000 to 7,000,000 years ago) produced some ruminants, such as *Blastomeryx*, that are hard to distinguish from the early palaeomerycine relatives of giraffes and deer in the Old World, which, with the North American groups, constitute the family Palaeomerycidae. Some developed horns, and the dromomerycine *Cranioceras* even had a third horn above the back of its skull. During the Miocene and Pliocene there finally appeared relatives of the surviving pronghorn, an example being *Merycodus*. Many of these North American groups have parallels with Old World groups, and the subject of North American artiodactyl evolution is of great interest. Only further finds will indicate whether *Blastomeryx*, the dromomerycines, *Merycodus*, and the pronghorns evolved from hypertragulids already in North America or sprang from some immigrant ruminant and, if the latter, whether the supposed hypertragulid *Leptomeryx* could be such an immigrant ruminant. It is uncertain whether the hypertragulids are nearer the tragulines or the camels, and how close the oreodonts are to the anthracotheres. Of the great New World radiation there survived after the Pleistocene only three or four camelid species and the pronghorn (deer and bovids in the Americas are immigrants), whereas in the Old World as little as 200 years ago, Eurasia and Africa had abundant deer and antelopes.

Until the Miocene there were some archaic artiodactyls in Europe, the xiphodonts, which have cautiously been taken as tylopods, and the cainotheres and anoplotheres, which are classified near anthracotheres.

A possible ruminant ancestor was *Archaeomeryx* from the upper Eocene of China, a small animal that already had a fused naviculo-cuboid bone in the ankle. Tragulids occurred in Africa and Eurasia back to the Miocene, and the more advanced gelocids are known from the upper Eocene and lower Oligocene. At the end of the Oligocene, the first ruminants began to appear with teeth more advanced than those of tragulids. From early in the Miocene they began to be recognizable as giraffes, deer, or antelopes, although the last were relatively uncommon before the late Miocene. Much remains to be learned about the detailed early history of these groups. Several different giraffids lived in later Miocene and early Pliocene times, but the group has since declined to only two species. Deer gradually acquired more complicated antlers, which became very large in some lineages. Different subfamilies of bovids originated in Eurasia and Africa, and it is of zoogeographic interest that representatives of African subfamilies have been found as fossils in northern India and Pakistan.

CLASSIFICATION

Annotated classification. The following classification is principally based on that of American paleontologist George Gaylord Simpson, with alterations in the bovid subfamilies, in the placing of early relatives of giraffes and deer in a giraffoid subfamily Palaeomerycinae, and in the placing of hypertragulids and protoceratids with camels. Groups indicated by the dagger (†) are known only as fossils.

ORDER ARTIODACTYLA

Cloven-hoofed ungulates, the major group of herbivorous mammals. Weight supported mainly through 3rd and 4th toes; astragalus with upper and lower articulations rounded. Stomach compound and, with intestines, enlarged for plant digestion. About 150 species.

Suborder Suiformes

Complete dentition, bunodont (low-cusped) molars, short legs, 4-toed feet are among their important characteristics.

†Infraorder Palaeodonta

Eocene to lower Miocene. Primitive, small-brained artiodactyls. Two superfamilies, Dichobunoidea and Entelodontoidea, with 5 and 3 families, respectively, and collectively about 30 genera.

Infraorder Suina

Lower Oligocene to present. Includes the living pigs, peccaries, and their likely ancestors and extinct relatives.

Family Suidae (pigs). Lower Oligocene to present; Old World. Small to moderate size; shoulder height to about 100 cm (39 in.). Coarse hair. Omnivorous, with sharp-edged tusks. Five Recent and about 22 fossil genera.

Family Tayassuidae (peccaries). Differ from pigs by having 1 fewer incisor and premolar, smaller canines, less advanced cheek teeth; hindleg with a cannon bone, more complicated stomach and more densely haired coat.

Infraorder Ancodonta

†Family Anoplotheriidae. Eocene and Oligocene; Europe; uncertain relationships.

†Family Anthracotheriidae. Eocene to Pleistocene. Mainly Old World, a few in North American Oligocene. Large, with cheek teeth showing beginnings of selenodonty.

Family Hippopotamidae (hippopotamuses). Middle Pliocene to present. Thought to be derived as late as the Pliocene from anthracotheres. Old World, now restricted to Africa. One large (shoulder height to 170 cm; weight to 3,000 kg) and one small species (height to 90 cm). Feed on land but frequently resort to water.

†Family Cainotheriidae. Mainly Oligocene; small European forms of uncertain relationships.

†Infraorder Oreodonta

†Family Merycoidodontidae (North American oreodonts). Eocene to early Pliocene. No suppression of upper incisors, an incisiform lower canine, selenodont cheek teeth, short faces and short limbs.

†Family Agriochoeriidae. Eocene to lower Miocene. Close to the above family but with clawed feet.

Suborder Tylopoda

Some reduction of upper incisors, reduced premolars, selenodont cheek teeth; cannon bones present. Feet became 2-toed early in the geological history of the group.

†Family Hypertragulidae. Upper Eocene to lower Miocene; North America. Like Old World Tragulina (see below) but with a canine-like first lower premolar. Fused naviculo-cuboid in the ankle.

†Family Protoceratidae. Oligocene to lower Pliocene; North America. Some with horns on the top and at the front of the skull. Later ones with hindleg cannon bone. Generally considered close to hypertragulids, but failed to fuse navicular and cuboid bones.

Family Camelidae (camels and lamoids). Upper Eocene to present; now a relict group, represented in southern South America, in Asia, and in North Africa. Red blood corpuscles oval. Gallbladder absent. The hump of the 2 Old World camels is composed of fibrous connective tissue and fat.

†Family Xiphodontidae. Eocene and lower Oligocene of Europe. Already 2-toed, despite their antiquity, and tentatively placed with the camels.

Suborder Ruminantia (ruminants)

Upper incisors lacking; lower canine incisor-like; cheek teeth selenodont. Fused magnum-trapezoid bone in the wrist. Two-toed feet evolved within suborder.

Infraorder Tragulina

†Superfamily Amphimerycoidea

†Family Amphimerycidae. European Eocene and Oligocene of Europe; poorly known. Asian *Archaeomeryx*, usually placed in the Hypertragulidae, but may fit here; it retained upper incisors and had a fused naviculo-cuboid in the hind leg.

†Family Gelocidae. Eocene-Oligocene of Europe and Asia.

Superfamily Traguloidea

Family Tragulidae (chevrotains). Miocene to present. Sabre-like upper canines in males; incipiently selenodont molars.

Bony carapace often develops above the pelvic girdle in males.

Infraorder Pecora

Mostly with horns or antlers and without upper canines. Hollowed auditory bullae. Four-chambered stomach.

Superfamily Giraffoidea

†*Family Palaeomerycidae*. Upper Oligocene to upper Pliocene; North America, Europe, Asia, Africa. Three subfamilies, the Palaeomerycinae (Old World basal stock for giraffes and deer), Blastomerycinae, and Dromomerycinae (the last two New World).

Family Giraffidae (giraffes and okapi). Miocene to present; Old World, now confined to Africa. Living giraffes long-necked and long-legged; the okapi more compact. Giraffes may be up to 5.5 m (18 ft) in total height. Extinct relatives including the large, short-legged, and grotesquely horned sivatheres. Living species have no gallbladder.

Superfamily Cervoidea

Family Cervidae (deer).

Subfamily Moschinae (musk deer). Pleistocene and present; Asia. One recent species, the musk deer (*Moschus moschiferus*) with sabre-like upper canines, gallbladder present (lacking in most other deer).

Subfamily Muntiacinae (muntjacs). Lower Miocene to present; southern Asia. Includes the living muntjacs (*Muntiacus*), with small, 2-pronged antlers above a long, skin-covered base; and tufted deer (*Elaphodus cephalophus*) with tiny antlers. Eurasian fossil genus *Dicrocerus* had larger, 2-pronged antlers. All have large, curved upper canines.

Subfamily Odocoileinae. Pliocene to present. Characterized by the persistence of the lower ends of the lateral metacarpals, thickly haired skin between the hoofs, and in all except the moose a large interdigital gland in at least the hindfoot. The vomer is fused far back posteriorly with the palate in the American deer and in the reindeer. Includes moose (Eurasian elk), reindeer, roe deer, perhaps the Chinese water deer (which has long canines but no antlers), and the deer of North and South America other than the wapiti (American elk), a cervine.

Subfamily Cervinae. Pliocene to present. Top ends of the lateral metacarpals persist. Smooth skin between the hoofs; interdigital glands lacking. Branching of beam of antler differs from that in New World deer; red deer and wapiti, fallow deer, chital, sika, Père David's deer.

Superfamily Bovoidea

Family Antilocapridae (pronghorn and merycodonts). Miocene to present; North America. Smooth, branched horns consisting of hollow sheaths over bony cores; only sheaths shed annually; in fossil Merycodontinae, 1 or more burrs at base of cores. Teeth high-crowned.

Family Bovidae (cattle, sheep and goats, antelopes). Wild cattle, sheep, and goats were ancestral to domestic livestock not differing in any fundamental characters from antelopes. Great variety in horn shape. Many with high-crowned teeth.

Subfamily Bovinae (cattle and some antelopes). Miocene to present. African tribe Tragelaphini, with keeled, spiral horns and not very advanced teeth, includes eland, kudu, nyala, and bushbuck. Tribes Boselaphini and Bovini, mainly Eurasian, former including the Indian nilgai, the 4-horned antelope, and some extinct forms, the latter including cattle, bison, and buffaloes. Bovini are descended from extinct Boselaphini; large size, up to 180 cm at the shoulder. Teeth specialized for grazing. Animals wallow frequently.

Subfamily Cephalophinae (antelopes). Mostly small, with tiny horns set toward the back of the head. Gallbladder lacking. Generally forest-living. African duikers.

Subfamily Hippotraginae (antelopes). Moderate-sized, stocky, mainly grazing antelopes; shoulder height 60–160 cm. High-crowned teeth. Tribe Hippotragini includes the roan, sable, oryx, and addax antelopes; and Reduncini the reedbuck, kobs, lechwes, and waterbucks. Oryx and addax inhabit arid areas; others near water with adjacent cover or high grass. All except nearly extinct Arabian oryx are now native only to Africa, but the subfamily formerly occurred in India.

Subfamily Alcelaphinae (antelopes). African wildebeests, hartebeests, topis, and several extinct lineages. Long-faced, plains-living, grazing antelopes. Sometimes included in the Hippotraginae.

Subfamily Antilopinae (antelopes). Tribe Neotragini includes some small African antelopes, and Antilopini include gazelles, springbok, and the Indian blackbuck. Graceful, long-legged antelopes of arid, open country. Subfamily may also include the saiga and the Tibetan chiru (tribe Saigini), hitherto classified with Caprinae.

Subfamily Caprinae (sheep, goats, and relatives). Mainly Eurasian. Moderate-sized, shoulder height often 90–100 cm. High-crowned teeth. Agile animals; many species in mountains or on steep, rocky slopes. Tribe Caprini comprises sheep and goats; Ovibovini the musk-ox (*Ovibos*), the Chinese takin, and some bizarre extinct forms; Rupicaprini the chamois, serow, goral, and Rocky Mountain goat.

Critical appraisal. The great 18th-century classifier Carolus Linnaeus recognized the camels and ruminants as associated but placed some nonartiodactyls with them. It was the French naturalist Henri de Blainville who, at the beginning of the 19th century, first recognized the complete order of artiodactyls as it is accepted today. Nine discrete groups exist among the living forms: pigs, peccaries, hippopotamuses, camels, chevrotains, deer, giraffes, pronghorn, and bovids; their classification presents no great problems, apart from a few genera. Fossils, however, bring confusion to various schemes.

The relationship of North American Tertiary artiodactyls to those of the Old World is a basic question in the study of their zoogeography, history, and classification. It can be agreed that the camels evolved in North America and are as old as the tragulines, which in the Old World were ancestral to ruminants. Most paleontologists today believe that the hypertragulids, protoceratids, and oreodonts were related to the camels, but others have linked the first two groups with the tragulines and the oreodonts with the anthracotheres. Other questions affecting the higher levels of artiodactyl classification are the placing of the North American native ruminants and whether the earliest Old World pecorans should be taken as giraffoids or cervoids.

The subdivisions of the Bovidae remain controversial. Modifying a classification proposed by German zoologist Max Schlosser, some authorities have grouped the Bovinae, Cephalophinae, and Hippotragineae as the Boödontia and the Alcelaphinae, Antilopinae, and Caprinae as the Aegodontia, to indicate phyletic lines believed to have arisen early in bovid history. Boödonts and aegodonts have evolved differently in Africa and Eurasia, and there is much to be said for developing a classification that reflects these geographical relationships. Until additional evidence of phylogenetic relationships is available, however, the modified version of Simpson's classification above will remain favoured by most taxonomists.

BIBLIOGRAPHY. I.W. CORNWALL, *Bones for the Archaeologist* (1956), on the morphology and identification of bones including artiodactyls; J. DORST and P. DANDELOT, *A Field Guide to the Larger Mammals of Africa* (1970), summarized descriptions, habits, ecology, and distribution maps for all African artiodactyls; R.F. EWER, *Ethology of Mammals* (1968), a comprehensive text with much recent information on artiodactyl behaviour; V. GEIST, "The Evolution of Horn-Like Organs," *Behaviour*, 27:175–214 (1966), on the functional implications of horn shape; G.G. SIMPSON, "The Principles of Classification and a Classification of Mammals," *Bull. Am. Mus. Nat. Hist.*, vol. 85 (1945), forms the basis for most modern classifications of artiodactyls; T. HALTENORTH in W. KUKENTHAL and T. KRUMBACH, *Handbuch der Zoologie*, vol. 8, pp. 1–167, *Klassifikation der Säugetiere: Artiodactyla* (1963), a weighty classification of artiodactyls (in German); V.G. HEPTNER, A.A. NASIMOVIC, and A.G. BANNIKOV (eds.), *Die Säugetiere der Sowjetunion*, vol. 1, *Paarhufer und Unpaarhufer* (1966), a massive work on Eurasian ungulates, most of it dealing with artiodactyls, originally published in Russian in 1961; A. KEAST, "Comparisons of the Contemporary Mammalian Faunas of the Southern Continents," *Q. Rev. Biol.*, 44:121–167 (1969), a review of zoogeography and adaptations, with further references; P.S. MARTIN and H.E. WRIGHT (eds.), *Pleistocene Extinctions: The Search for a Cause* (1967), a collection of essays on Pleistocene extinctions involving artiodactyls; D. MORRIS, *The Mammals* (1965), a general account with illustrations; G.B. SCHALLER, *The Deer and the Tiger* (1967), a study of the life of Indian artiodactyls; C.A. SPINAGE, *The Book of the Giraffe* (1968), much information well presented for the general reader; W.P. TAYLOR (ed.), *The Deer of North America* (1956), all aspects of the life of North American deer; J.Z. YOUNG, *The Life of Vertebrates*, 2nd ed. (1962), contains a chapter on artiodactyls; F.E. ZEUNER, *A History of Domesticated Animals* (1963), a useful source for much information difficult to find elsewhere.

(A.W.G.)

Arts, Classification of the

As long as the term "arts" is applied to a realm so vast and indefinite as to embrace literature, music and dance, theatre and film, the visual and decorative arts, and other equally diverse activities, their classification—the ways in which each is regarded as being either unique or like others—will remain a controversial but necessary undertaking. Classification is a useful approach to the organization of knowledge in any field: the classification of plants and animals in the 18th century led to the discovery of evolution in the 19th. In the arts, classification can be of immense help in understanding the interrelations between the arts and in drawing attention to characteristics of each that might otherwise go unnoticed. Whether it is consciously devised or largely unconscious, some sort of classification is implicit in any serious exploration of the arts. Related to it is an appraisal of the importance and value of each of the arts. Admirers of a particular art often feel that it is unique and deny that it is like any other. Logically, however, anything correctly described as a work of art, whether a poem, a picture, or a sonata, belongs to that class and hence resembles other members of the class to some extent. Some works or types of art are more nearly unique than others in resembling fewer others or in resembling others in fewer respects. On the other hand, no type or example is exactly like any other, but the differences may be negligible for all practical purposes.

Classifying the unique

There is no one correct way of making a classification. The aims and interests of the person making the classification, and his philosophical orientation, are usually significant factors. The emphasis may be placed on sensory qualities, for example, or on moral and religious ideals. Regardless what classification is used, however, as it is subdivided into smaller groups, detailed similarities and differences are brought out in the emotional, intellectual, and social features of the works or art involved.

The classification of the arts is closely related to subjects that receive more extensive treatment in the articles AESTHETICS; ART, PHILOSOPHY OF; and ARTS, CRITICISM OF THE. More extensive discussions of certain specific systems of classification may be found in articles such as ARTS, STYLE IN THE; PRIMITIVE, FOLK, AND POPULAR ARTS; and ARTS, SOCIAL AND ECONOMIC ASPECTS OF THE. The relationship between the arts within a given culture is discussed in such articles as SOUTH ASIAN PEOPLES, ARTS OF and comparable articles on the arts of other peoples.

The article that follows will attempt only to draw the reader's attention to the importance of the subject of classification in the study of the arts and to offer him means of applying thoughtful scrutiny to a process that has too often remained subconscious.

BASES OF CLASSIFICATION

To classify things is to arrange them in groups or sequences according to a plan, especially on the basis of some characteristic that they are thought to have in common. Thus, the names in a directory may be arranged according to the letter of the alphabet with which each last name begins. A biologist classifies the kinds of animal or plant within a certain area. A museum of art may place its sculptures in one gallery, its paintings in another. It may also classify them according to historical period, such as Italian Renaissance or 18th-century French.

A base or basis of classification is a concept of what the members of a certain class are thought to have in common, a factor that can be used either for grouping them or for separating them from other classes. "Subject represented" is one such base: it may be used to separate "paintings" into three or more groups, such as portraits, landscapes, and narrative scenes. This involves classification in a narrow sense of the term; that is, as grouping certain items or subgroups together under certain headings. It also involves division, which can be regarded as the opposite of classification, but the latter term is usually made to cover both processes. A large, complex art such as literature can be divided into smaller ones, such as poetry and prose. Each of these also can be subdivided, as into epic, dramatic, and lyric poetry. A relatively complex, detailed division and classification of the arts from a philosophic point of view is called a system of the arts.

The study and scientific practice of classifying phenomena in an extensive field is called taxonomy. In biology and other exact sciences, it is highly complex and consistent, with precise definitions. Thus a "genus" in biology comes between a "family" and a "species." In classifying arts, however, such terms are used more loosely. Thus, genus can mean, in aesthetics, any fairly large, inclusive group, capable of including subgroups and also of being subsumed under a still larger group. A species, in the broad sense, is one of the main divisions of a genus, and the traits or characteristics of a genus are generic traits. Differentiae are ways in which individuals or subgroups in a larger group are significantly unlike. Aristotle distinguishes between epics and tragedies in poetry in terms of differentiae.

Taxonomy in the arts

To define or locate a particular example or subgroup in a system of classification, it must be ascertained, first, what genus (inclusive group) or genera it belongs to and then how it differs from others in that group or groups. A species is one of the main divisions of a genus, but it may also act as a genus (in the broader sense of that word) by including smaller groups.

Problems of definition. What are the arts, and what is art in general? This is still a controversial question after centuries of debate. No particular definition commands universal assent. Several meanings are still frequently used, of which the oldest is the broad, technical sense. In this sense, the English term art and its equivalents in Greek and Latin covered not only what are now called "fine arts" or "aesthetic arts" but any kind of transmitted, useful skill, such as agriculture, medicine, and war. This sense of "art" survives in such terms as Bachelor of Arts, a degree that is often awarded for a course of study that involves no aesthetic arts at all.

In the 18th century, the so-called beaux arts, the beautiful or fine arts (also called "elegant" or "polite" arts), were distinguished from the merely useful arts on the ground that they were aimed at giving aesthetic pleasure to the beholder. In the 19th and 20th centuries, there has been a tendency to abandon the term art in speaking of the purely utilitarian skills and to call them instead "industries," "technics," "branches of engineering," or "applied sciences." Without the prefix fine, the word art alone is now commonly understood to mean the fine or aesthetic arts. To produce an experience of beauty or aesthetic satisfaction is said to be their distinguishing function or characteristic but not necessarily their only one. In this moderately broad, technical sense, some, but not all, architecture, furniture, and clothing can qualify as arts in spite of their useful purposes. They can be called useful arts, rather than merely useful skills, because they combine aesthetic and utilitarian forms and functions. On the other hand, such technics as coal mining and placing metal pipes underground do not qualify as arts at all, since they do not ordinarily involve aesthetic perception.

In psychology, anthropology, and other sciences, a particular product or performance does not have to be beautiful or aesthetically satisfying to qualify as art. The same can be said of much primitive art, children's art, and art produced by the insane. They may be classed as art if they belong to types of product or performance that have been socially recognized as having an aesthetic function. In this sense, any picture, clay figure, dance, or traditional song can be accepted as a work of art, whether beautiful or not. Being nonevaluative, this conception makes it possible for the scientist to study the arts as a field of cultural phenomena for investigation, without having to show in advance that they are pleasant, good, or beautiful.

In another sense of "art," called the expressionist theory, art has been defined as the expression and transmission of remembered emotion. This definition is not inconsistent with the technical meanings, but it puts an emphasis on the artist's procedure rather than on the effects and functions of the product.

In a third sense (an extremely narrow one), the concept of art is limited to painting and drawing alone, sometimes to the visual arts alone. This definition has the disadvantage of excluding music, poetry, dance, and many other arts that have long been recognized as "fine" or "aesthetic." It is confusing in that painting and the other visual, manual arts were themselves long excluded from the category of "liberal arts."

Decorative and industrial arts

The decorative arts are a species of visual art whose main function is to combine utility with beauty or aesthetic satisfaction. They tend to emphasize visual ornamentation and design along with fitness for some useful end or ends. Although Western painting and sculpture in the past traditionally tended to emphasize representation, the decorative arts used both abstract and representational design. Utility, design, and representation appear with varying degrees of emphasis in such arts as medieval book illumination, jewelry, Greek or Chinese vase painting, Persian rugs, and French rococo furniture. Some styles of decorative art are comparatively plain and simple in order to achieve an effect of visual design without superficial ornament.

Ironworking is usually classed as an industry not an art, because most of its activity is devoted to products that have no aesthetic intent. A small branch of the industry, however, may be devoted to making decorative designs in iron and steel, perhaps for use in architecture. This branch may be called a useful art or a decorative art; the two categories overlap. If old-fashioned hand methods are used there, the term handicraft is also applicable.

The term industry, or industrial art, is now applied chiefly to arts in which machinery and mass production are commonly employed, with many specialized workers cooperating, as in motion pictures. It may be applied to the process of making large numbers of colour-print reproductions from paintings. Such art used for advertising, however, as in newspaper layouts and street posters, is called "commercial art." The industrial arts include many types of large-scale manufacture in which an aesthetic appeal is sought, as in the manufacture of books, magazines, refrigerators, typewriters, furniture, television sets, automobiles, airplanes, or appliances. The aesthetic factor in such work is sometimes called "styling."

Bases in status. Until the 18th century, the most common basis for classifying the arts was in terms of their social and psychological status. This system, which was devised by the ancient Greek philosophers, involved a wide separation between the so-called liberal and servile arts. In ancient Greece and Rome, the liberal arts were conceived as: (1) those that befitted a freeman and were thus comparatively noble, aristocratic, or genteel; (2) those requiring the exercise of superior mental ability, rather than mere hand labour, however skillful; and (3) those tending to elevate the minds of the artist and of his patrons, rather than merely providing material comforts, pleasures, and conveniences. The servile arts were those befitting only a person of a lower class, at least as far as the work itself was concerned. It was beneath the dignity of a lady or gentleman to work in them but not to use and enjoy such products. For the wealthy aristocrat, even to practice painting as a hobby exposed him to ridicule. Aristotle notes that even if the gentleman practices music for his own pleasure, he should not do it too well.

Liberal and servile arts

From the medieval period through the Renaissance, the class distinction between liberal and servile was retained in modified form. The "seven liberal arts" of the Middle Ages were composed of the trivium, or literary group—comprising grammar, dialectic, and rhetoric—and the quadrivium, or mathematical group—comprising arithmetic, geometry, music, and astronomy. Music was theoretical in emphasis, with little attention to its sound and much to the numerical relations among the tones. In the 12th century, literature, including poetry, tales, and dramas, was classed as a mere supplement or aid to philosophy. Modern taste admires the visual arts of the Middle Ages, but many churchmen and philosophers of the time condemned them as sensuous, perhaps idolatrous. The artist in a material medium such as stone or metal was low in social status until well along in the Renaissance.

In and after the Renaissance, art was conceived as being more hedonistic, more devoted to providing aesthetic pleasure through the sensuous perception of beautiful forms. This conception partly replaced the emphasis on moral and intellectual qualities in the ancient Greek and Roman traditions. The manual arts of decoration and design gradually rose in prestige, along with the arts of music, poetry, ballet, and theatre. Visual artists in such media as architecture, landscape, design, and pageantry gained in respect and in financial rewards. High status in them depended more on success in pleasing a luxury-loving aristocracy than on moral and intellectual virtues. But some philosophers urged that art at its best combines both sets of values.

Elite and popular arts

A hint of social status persists today in the distinction between "elite" and "popular" art, which includes products of the "mass media" such as film, newspaper and magazine illustration, radio, and television. But this distinction refers more to the level of taste and education supposedly required to appreciate these arts, and to their different publics, than to social or financial status.

Certain arts have been grouped from time to time as "lively," suggesting animation and gaiety, in contrast with the supposed solemnity of classical, "highbrow" art. Such arts are, on the whole, suited to popular taste, though often appealing also to the scholarly. They often are easy to grasp and involve a simple narrative. The comic strip is static, but it suggests movement, often playfully exaggerated into mock violence. Some, but not all, lively arts involve directly presented motion, as in the film, which has rapidly ascended from a popular to a serious art. At its best, it can achieve all types of aesthetic value, including emotional expression, character, and plot, but it does not always try to avoid banality.

Major and minor arts

The antithesis between "major" and "minor" arts is partly dependent on the social status and cultural context of the arts concerned, though it is often mentioned as if certain arts were inherently and permanently greater than others. The ground of superiority of an art, according to some philosophers, is its greater ability to express thoughts and feelings of universal, lasting value. Poetry and other forms of literature have been regarded as superior in this respect to the decorative arts, which were associated with superficial, thoughtless luxuries. As the decorative and useful arts rose in critical esteem, however, a tendency to reject these evaluations has become evident. A Persian rug or a sonata, it is said, is not necessarily inferior in aesthetic, moral, or spiritual values to a poem, play, or novel. Its value depends on what it does in its own medium. What seems at first to be mere trivial ornament may have deep, symbolic meanings for the culture in which it originated. Certainly the relative size of the products provides no sure measure of value, as, for instance, between a badly designed cathedral and a well-designed miniature painting. The major–minor antithesis may be useful if applied to the relative status of an art in its own cultural context. The status of an art may change within a culture and differ widely from one culture to another. Tattooing has been a major art in some primitive cultures, notably among the Maori people of New Zealand. Mosaic was a major art in Byzantine culture. Poetry, for the present, has declined in importance in Western civilization. The creative energies of man flow at different times through different channels.

Mode of presentation as a base. One way to avoid questions of value in a descriptive classification of the arts is to use as a base the concept of the sense to which the work is primarily addressed, instead of fineness or beauty. Painting, sculpture, and architecture are said to be addressed primarily to the sense of vision; hence they are now increasingly known as "visual arts." Music is an "auditory" art. Opera and sound film with colour are "audiovisual." An armchair is addressed not only to vision but also to the sense of touch and to muscular sensations of comfort and fitness for the posture or range of postures desired. Thus, a throne, a theatre seat, and a dentist's chair will have somewhat different forms and functions. From the aesthetic standpoint, the visual characteristics of physical objects are usually emphasized.

Some philosophers have pursued this approach into the realm of "lower senses," providing for the "gustatory" art of cuisine and the "olfactory" art of smell, as in perfume and ritual incense.

Painting is one species of visual art, and it can be subdivided in various ways. One is according to style, such as baroque or romantic. Another is based on the amount of representation, whether the work is realistic, abstract, or nonobjective. The words abstract and nonobjective are sometimes used interchangeably, to mean nonrepresentational or intentionally devoid of resemblance to any outside object. At other times, an abstract painting is said to be one that started with a representational conception but omitted some or all its representational details, while a nonobjective picture is one conceived and executed from the beginning in terms of lines and colours without any definite outside reference.

Abstract and nonobjective art

Painting can also be divided according to the subject represented; that is, into portraiture, landscape, fantasy, and so on. Some paintings are colouristic in style; others are linear. Some emphasize perspective, with imaginary vistas into deep space. The other arts can be similarly divided: narrative literature can be in prose or verse; prose narrative can be divided into novels, short stories, and so on.

Poetry is sometimes presented aurally, as in speaking it aloud; at other times, visually, as in reading it silently; at still other times, tactually, as in reading Braille type for the blind. In opera, poetry is presented audiovisually. Originally, when few persons could read, it was presented to the sense of hearing. Now, it is more often read silently. A poem is fundamentally the same whether read silently or heard. In reading it silently, the word sounds are imagined rather than heard. For purposes of classification, the art of poetry may be described as primarily auditory, but now, often visual or audiovisual; in short, it is variable.

The "performing arts" are so designated because of the ways in which they are presented to the attention of observers. Most of the visual arts perform automatically, so to speak. The artist who created the painting or statue did his performing once and for all time, when he made it. There is a temporal sequence in perceiving any complex, three-dimensional work of art, especially a large one such as a cathedral, since the observer must walk around it and perhaps inside it in order to see it fully as a three-dimensional form. The form itself may change in appearance from moment to moment, through changes in sunshine, atmosphere, and shadows. Such changes are usually considered to be indeterminate and superficial, however, not enough to characterize the work of art itself as mobile. In mobile sculpture, such as that of the modern American Alexander Calder, the form as a whole changes in accord with air currents and other pressures from outside. Unless moved in a definite way, as by a motor, the succession of arrangements in mobile sculpture is somewhat indeterminate.

Mobile and static arts

Arts in which the objects do not ordinarily have to move or change radically in order to be properly observed have been called "static arts," "space arts," or "arts of rest," in contrast with "mobile," "dynamic," or "time" arts such as music and dance. Some arts, notably cinema and ballet, unfold actively in both space and time. In cinema, the product—the film projected on the screen—is made to perform automatically; the operator does not have to influence the showing unless something goes wrong. The main performance was done permanently in the photographing, editing, and other processes involved in producing it. Much the same can be said of a phonograph record or tape, since the performance was completed by the musician before the record was made or marketed. Slight adjustments in tone, balance, volume, and the like may be required at the start; after that, the record performs automatically.

What is meant in speaking of "performing arts," however, is, for example, what the actor does in speaking his lines and moving or gesturing as directed, what the musician does in playing a certain sonata on his piano or violin, and what the dancer does in moving his body as required by the choreography. In these examples, there is usually a set of directions, prepared by the designer, the creative artist or group of artists, and made available to the performer. As a rule such directions allow some scope for original interpretation or minor departures from the score, so that the performance is not purely automatic. Different performances of the same script or printed score may vary considerably as to nuances of expression and yet be recognizable as renditions of the same composition. Such variations help to distinguish an original interpretation (good or bad) from a merely mechanical reproduction. Of course, a creative performer may at times compose his own score or improvise without the aid of any directions.

Early theories contrasted "arts of motion" with "arts of rest" and put painting in the latter category. With the advent of the film, especially such works as Walt Disney's "animated cartoons" in colour, photography and painting began their rapid development as arts of time. They are now arts of both space and time, rest and motion. This development illustrates the need for flexibility in classification, to allow for the accelerating evolution and unpredictable variation of the arts.

Arts of notation. Arts in which works are shown or sounded in a definite temporal sequence tend to acquire an appropriate notation to guide that sequence. Chief among these are the performing arts of spoken literature (especially poetry and drama), music, dance, pantomime, and other temporal arts of the theatre. Stage design, costume, and lighting are sometimes classed as theatre arts but are not necessarily shown in temporal order.

The first notation of art

Poetry was the first art for which a definite notation, that of writing and literature in general, was developed. As sung or recited by the ancient bards, poetry left a great deal to be supplied by tradition and individual taste. Indications of rhythmic variation—by separating lines, sentences, and phrases with the aid of punctuation, capital letters, and italics—were slowly developed. Some of the earliest musical notation consisted of a few conventional marks above the words of a poem. Dance steps, which were simple and regulated by tradition with the aid of music in songs and rituals, remained long without notation.

The notations of music, from the Renaissance to the mid-20th century, became very precise in directing pitch, melody, chord structure and progression, rhythm, metre, phrasing, dynamics (loud and soft), and many nuances of expression. Because of its scope, precision, and adaptability to many styles of music, the printed score was widely used as a framework for forms that combined music with words, such as song, cantata, oratorio, or opera. It was found impossible, however, to record Oriental, primitive, and other exotic styles of music and dance in Western musical notation. Such music includes pitches that cannot be precisely indicated in Western scales; its rhythms tend to overflow European bar lines, and the timbres of its instruments and voices cannot be exactly described in conventional Western notation. Such limitations were even more keenly felt as Western avant-garde composers sought to use a great variety of sounds, instrumental and otherwise, that had not been previously used in serious music—sounds of nature, city life and countryside, birdsongs and traffic noises, flowing water, thunder and lightning, and many new sounds derived from electronic machinery. In music of the 20th century, timbre and tone quality tended to be emphasized rather than conventional melodic and chordal progressions. Timbre, rather than rhythm or pitch, became the principal component of experimental music. Physical scientists cooperated with composers in adapting the new knowledge of electronic sound to music.

For the dance also, conventional musical notation came in the 20th century to seem less than adequate for guiding movements. Traditionally, the mobile designs of dance had been based on a set of conventional postures and movements, mostly devised in the 18th century. Light and graceful whirls, leaps, and glides on tiptoe could be fairly well symbolized in a simple choreographic notation, which is still used to some extent in traditional

solo dance and ballet. As in music, however, style leaders felt the need of a more flexible notation, capable of recording new types of emotional expression.

Another motive for breaking with the past in this regard was the feeling that dance should not be a handmaid to music or limited to expressing musical forms and feelings in bodily movement. Experiments were tried in dancing without music or with rhythm alone, and new bases were sought for recording the moods and movements of the dance in visual symbols.

Imperfections of notation

It would be unreasonable to expect a perfect notation in so complex an art as modern ballet. Even in such a long established notation as the printing of poetry, many qualities—*e.g.*, tempo—cannot be specified. Verbal notation is always incomplete, partly because both poet and reader usually prefer to have much of the word-sound content, along with other meanings, left to the imagination. Printed words are enough to start the reader on an approximate path toward understanding. The sound film and the phonograph offer new, partial substitutes for printed notation as guides to performing the temporal arts.

Other types of base. Three main types of base for classification and division of the arts may be distinguished. They may also serve as aids in defining particular arts or be combined into one system of the arts.

The first is that of medium or material. It is used in the names of such arts as painting, metalwork, woodcarving, jewelry, and ceramics. These refer to the physical materials out of which the work of art is made. Sound waves in music and light waves in cinema are also physical media. Tools and instruments, such as pianos, voices, brushes and canvas, chisels and marble, the human body (as in dance), are all distinctive materials of the various arts. Perceptual qualities such as rhythm, pitch, and colour are parts of the medium of sound film. Poetry also uses rhythm.

The second type of base is that of the process or technic employed, for example, the hands or other parts of the body, or instruments or machines, such as the cameras and projectors used in motion pictures. Shaping, sounding, and verbalizing are three main types of artistic process. They often overlap and combine in various arts.

When photography is conceived merely in terms of operating a camera, it is a medium of art rather than an art itself. It can be used either for purposes of art, as a scientific instrument, as a means of recreation, or in other ways. It approaches art when used as a device for giving someone an aesthetic, visual experience by means of the resultant picture.

Some technics are overt, as in dance, employing the whole body; others are mental and inner, as in memorizing a dramatic role. Some artistic skills are professional or managerial, some mechanical or manual. An architect plans and oversees the construction of buildings and the making of plans and specifications. A landscape architect works with plants, roads, levels of ground, hills and valleys, roads, lawns, bridges, and sprinkler systems. An author may do little that is overt muscular work but much dictating or typewriting.

Some art production is solitary, some cooperative. Some emphasizes visual shaping; some sounding; some verbalizing, usually organizing words and meanings approximately in accord with the rules of grammar and syntax in the language employed.

The third type of base is that of the form, design, and functions of the product, including its function as an aesthetic object; that is, as a stimulus to aesthetic perception and imagination. Arts can be grouped or distinguished as to the amount of emphasis on the following: (1) Presentative (*i.e.*, directly perceptible) and suggestive factors; (2) modes of suggestion (*e.g.*, mimesis symbolism, common association); (3) components in aesthetic form (*e.g.*, melody, harmony, rhythm, perspective, plot, characterization); (4) modes of composition: utilitarian (*e.g.*, as church or palace), representational (*e.g.*, narrative, dramatic, lyric), expository (*e.g.*, essay, symbolic picture), thematic design (*e.g.*, fugue, sonata, Persian rug, sonnet).

Types of formal emphasis

HISTORICAL DEVELOPMENT

In European philosophy, the classification of the arts, or "system" of the arts as it is usually called, has formed an integral part not only of the philosophy of art but of philosophic systems in general. This has been especially true since the work of the late 18th-century German philosopher Immanuel Kant (*q.v.*). European philosophers have not limited themselves to making superficial, verbal arrangements but have undertaken to show the role of each art in the mind of man, in world history, and in civilization. A system of the arts may also be used in an attempt to evaluate the arts and to list them in a hierarchy according to their metaphysical and moral roles. In the Western democracies, the classification of the arts has usually been given a more modest, secondary role in philosophy, limited to showing empirical relations.

Thinkers as different as Saint Augustine of Hippo (AD 354–430), one of the chief influences in Christian thought, and Francis Bacon (1561–1626), the English philosopher who was instrumental in the development of modern science, felt the need of a systematic survey of the aesthetic arts as part of a general survey of human knowledge and experience. Separated by more than 1,000 years, Augustine and Bacon disagreed on the value of worldly knowledge obtained through empirical science. Augustine disapproved of such knowledge, while Bacon approved of it. Looking back over his youth, Augustine found much to regret and repent in the pleasures afforded by the various arts to the five senses. Bacon, in his monumental survey the *Advancement of Learning*, admired the progress of the sciences and foresaw their future benefits to man; to poetry, painting, and music he gave a fairly high place, but in the "voluptuary arts," which appeal to the lower senses, he found little to praise.

Kant, writing toward the end of the 18th century, covered the field in a tolerant, empirical way, avoiding moral and metaphysical dogmatism. Beginning with "art" in general (in the broad, technical sense including all transmitted human skill), he distinguished it from nature, science, and paid handicrafts. He then distinguished between aesthetic art and mechanical art, the first of which he subdivided into fine, or beautiful, art and agreeable, or pleasant, art. His division of fine art into the arts of speech, the shaping arts, and the arts involving the beautiful play of sensations led him eventually to further subdivisions into poetry, music, landscape gardening, the art of colour, buildings, and furniture. Under agreeable art, he assigned a place for dinner music, table arrangement, and entertaining narrative.

Georg Wilhelm Friedrich Hegel (1770–1831), another German philosopher who was vitally concerned with the arts, offered a system combining the ancient Greek philosophical conception of a cosmic mind (which embraced all man recognizes as reality) with the theory of evolution. In the world process envisaged in his scheme, he assigned an important role to the arts: architecture is most capable of expressing the early, symbolic stage in world history; sculpture, the classic stage; painting, music, and poetry, the romantic stage.

The symbolic, classical, and romantic stages

The article on the Fine Arts, in the 11th edition of the *Encyclopædia Britannica* (1910–11), by the English man of letters Sir Sidney Colvin, is one of the few attempts in English to deal with the subject in a detailed, systematic way. He proposed three main divisions: the first, into shaping, moving, and speaking arts; the second, into imitative and nonimitative arts; and the third, into serviceable and nonserviceable arts. This triple division expressed Colvin's belief that no one formula could adequately describe the manifold interrelations of the arts.

On the other hand, Max Dessoir, a leading German aesthetician in the first part of the 20th century, tried to combine several bases in one pattern. Space and time arts are arranged in two vertical columns, with sculpture, painting, architecture, and plastic arts under "space" and mimicry, poetry, music, and poetic arts under "time." A third parallel column contrasts the arts "of imitation and definite associations" with the "free arts of indefinite associations." He commits the same error that many previous writers made in supposing that the space arts coincide

with the arts of rest and the time arts with the arts of movement and succession and that sculpture and painting are necessarily imitative (*i.e.*, representational). He did not recognize the extent to which the traditional arts changed, making new modes of classification necessary.

Instead of rectangular diagrams, Étienne Souriau, a French aesthetician of the mid-20th century, offered a wheel-shaped one. By means of this form, he showed how seven basic types of perceptible data (lines, volumes, colours, sounds, etc.) were developed into complex arts. Using concentric circles, he showed how each type of datum developed into a nonrepresentational and a representational art. Though no such diagram can offer a complete picture, Souriau presents a large number of interrelations in a simple pattern.

BIBLIOGRAPHY. ST. AUGUSTINE, *Confessions*, trans. by E.B. PUSEY, pp. 223–239 (1838), one of the earliest extant classifications of the arts, made in disapproving of them on moral and religious grounds; FRANCIS BACON, *De Augmentis Scientiarum*, an enlargement in Latin of *The Advancement of Learning*, bk. 2, ch. 1, vol. 4 of *The Works of Francis Bacon* (1858), a profound, far-reaching survey of the state of science and technology in the early 17th century, including a short classification of the arts; SIDNEY COLVIN, "Fine Arts," in *Encyclopædia Britannica*, 11th ed., vol. 10, pp. 355–375 (1910–11), one of the few attempts in English at a detailed, systematic classification of the arts on various grounds; MAX DESSOIR, *Ästhetik und allgemeine Kunstwissenschaft* (1906; Eng. trans. by S.A. EMERY, *Aesthetics and Theory of Art*, 1970), a system of the arts reduced to a short, neat pattern, by a leading German aesthetician; G.W.F. HEGEL, *Vorlesungen über die Aesthetik*, trans. by F.P.B. OSMASTON, *Philosophy of Fine Art: Introduction* (1920), a system that is remarkable for its cosmic breadth; IMMANUEL KANT, *Kritik der Urteilskraft*, part of this work trans. by J.C. MEREDITH as *Critique of Aesthetic Judgment* (1911), a pre-evolutionary empirical system, based on common sense and broadly tolerant toward the many manifestations of art in his time; THOMAS MUNRO, *The Arts and Their Interrelations*, rev. ed. (1967), a critical survey of the various classifications of the arts up to the mid-20th century, with reference to over 400 arts and types of art; *Oriental Aesthetics* (1965), on types and classifications of the arts in India, China, and Japan; ETIENNE SOURIAU, *La Correspondance des arts* (1947): a wheel-shaped diagram exhibits the system of the arts from a modern French point of view.

(Th.M.)

Arts, Criticism of the

Criticism of literature, of music, of the visual arts, or of any other of the arts can be a controversial enterprise. For one thing, disputes arise about the purpose and nature of the judgments made by critics; for another, disputes arise about the nature and properties of what it is that critics discuss—the works of art themselves; and for still another, disputes arise about the possibility of generalizing about *all* the arts, each of which differs significantly from the others in many respects.

Each of these disputes is really a complex cluster, involving nearly all of the persistent issues involved in the practice of criticism. Although this article cannot pursue all of these issues, it will attempt to provide an orientation that will facilitate their investigation.

It may be safely said that criticism in all the arts is concerned with the description and assessment of particular works. Although some hold that the critic's proper function is only the appraisal or evaluation of works of art, it is certain that nothing can be evaluated unless it is describable beforehand. Accuracy of description can in fact be so extremely difficult, particularly for complex or unfamiliar works and traditions, that some critics devote their efforts largely to matters of a descriptive sort.

It may also be said that all critics of the arts are concerned with the description and evaluation of works of art *as works of art*, even if they are more interested in moral, political, religious, ideological, or other considerations. For example, the moral criticism of works of art presupposes that the actual properties and features of a given work first may be fixed and then may be examined further according to moral considerations. Similarly, religious, political, or other considerations can only follow some relatively objective determination of what it is that is being examined. It is in this sense that criticism of the arts is primarily aesthetic. Although the objectivity of a description or an evaluation may be argued, as well as the propriety of concentrating on only the aesthetic aspects of a work, the central point is not open to dispute; *if* given works of art are to be appraised on any grounds, they must be independently describable beforehand. One cannot judge what one cannot identify and describe.

Primacy of aesthetic issues

Quarrels also arise about what may rightly be construed as the properties of works of art. For example, to say that a novel *expresses* a certain Gothic longing or that a painting *symbolizes* the end of feudalism or that a work of sculpture *represents* a bird in flight presupposes some theory of what a work of art is. The object cannot be described truthfully unless it is of a sort that can exhibit the properties ascribed, and quarrels often arise about what may rightly be construed as the properties of works of art as such. To see this is to see the sense in which descriptive statements are theory-laden. The theory of criticism and the theory of the nature of a work of art are largely aspects of the same question.

Provisionally, then, the criticism of works of art may be considered, at least minimally, as aesthetic criticism—that is, centred on the actual properties of the works considered. Correspondingly, appreciation of art may be said to be aesthetic when the art is savoured or enjoyed in terms of the properties that may be discriminated in it. Appreciation, in this sense of the term, may be informed by criticism, since both are focussed on the same properties. This point should be emphasized because it has been held by Tolstoy and others that a genuine appreciation of art must be a naïve, direct, or uninformed response. If serious thought and labour go into the creation of works of art, however, there is no reason why a similar effort will not be needed to appreciate or criticize them.

IDENTIFICATION, DESCRIPTION, AND INTERPRETATION

The crucial conceptual issues in criticism of the arts are what is meant by a work of art and how its properties may be ascertained. The difficulties involved in these issues are principally of two sorts. One concerns the identity and individuation of works of art. For example, different performances of the same piece of music may be perceived to differ from one another. A given piece may be transposed for different instruments, or it may be performed on instruments that have been modified since the date of composition, and performances by different artists will regularly exhibit noticeably different qualities. Nevertheless, it is normally considered one and the same piece of music as long as it is identified by reference to a particular score, even though aesthetically there may be more interest in the subtle variations in the performances than in what they have in common. Some critics hold that all distinct performances must be compatible with some ideal performance or that all admissible performances must correspond to the fundamental score, and deviations from it result in serious logical paradoxes. It does seem possible, however, to admit as instances of the same work performances so different from each other as to appear incompatible with any idealized performance.

The same conclusion may be drawn for all the arts that rely on a notational scheme. Variant versions of a poem, for example, may count as instances of the same poem, and analogous instances may be found in theatre, dance, film, and in view of the increasing importance of the blueprint, architecture. Analogies may be found in the plastic arts as well.

It is important to emphasize that an entire range of questions regarding the objectivity of criticism rests on puzzles respecting the referent of criticism. Both in describing and in evaluating particular works, it must be made clear whether different versions of the same work of art or several distinct works of art are being discussed. Identity may sometimes be established by detailed analysis, for example, by providing evidence on notational and

The referent of criticism

cultural grounds that what seemed to be different folk songs may be construed as variant versions of the same song.

The principal issues of criticism concern not the identification of works of art but their description and evaluation. It is not possible to segregate entirely what may be correctly said about a given work of art and what it is that individuates that work of art—that is, makes it a distinct entity within a class of related entities.

To illustrate this point, imagine that the central image of a given work may be construed in two radically different ways. Some of Anton Chekhov's plays, for example, may be construed as either comedies or tragedies, depending on whether the views of the author or those of the director Konstantin Stanislavsky are preferred. If the possibility of such competing views be admitted (even though the mere admission of them presupposes a certain theory of criticism), then any account of what a critic does must accommodate this possibility.

When critics say what a certain image means and their (defensible) accounts conflict with each other, it is clear that they cannot be *describing* the work at hand. It should be possible to confirm a description by an independent study of the object described. Obviously, it is not possible to confirm descriptions that are incompatible.

It might be argued that the differences in conflicting critical accounts are merely apparent, as a coin might be described as circular from one view and elliptical from another. To resolve such an apparent difference in a physical object like a coin, one could invoke some commonly accepted canon such as saying that when it is seen "under normal circumstances" the differences disappear. In criticism of the arts, however, there is no such common canon in terms of which the competing views may be sorted. Moreover, in some schools of criticism, such conflicting views would not be ruled out as impossible.

Distinction between description and interpretation

The solution is to distinguish between *describing* works of art and *interpreting* works of art: a critic's description may be true or false of the work in question, but his interpretation of the work can be only plausible or implausible. Each of several incompatible interpretations may be plausible, but they cannot all be true. In rendering an interpretation of a work of art, then, the critic must be imputing to it properties that cannot with certainty be found in it. This view of critical practice implies a theory of the nature of a work of art that allows some properties to be considered either truly in the work or only plausibly imputed to it.

A substantial body of criticism is intelligible only if the critics who purport to be describing given works—incorporating analysis, comparison, historical or biographical explanation, and the like—are, to some extent at least, interpreting those works. The enormous body of criticism of the greatest poetry, drama, and fiction shows how dissimilar the apparently descriptive efforts of critics may be. The mere existence of this vast body of diverse views would appear to refute the claim that interpretation is merely the unearthing of what is descriptively true of a work of art. It may include this but it must be more.

Interpretive criticism would be unintelligible without a counterpart theory of the nature of a work of art. The key to that theory lies in not confusing works of art with the physical objects associated with them. Physical objects having different physical properties cannot be the same object. As has been noted, however, works of art with different physical properties, and even critical aesthetic differences, may count as instances of the same work of art. This fact is most easily seen in literature, drama, music, and, to some extent, dance and architecture, arts in which the identity of a given work depends on some notational system. It is less easily seen in the visual arts, even though different tokens of the same work, as in etching, may be admitted there too. Clearly, then, works of art are individuated in ways that substantially depart from those of physical objects and in ways that vary from art to art.

A further consideration is that a work of art is a purposive system, whose internal organization may be seen to reflect the systematic decisions of the artist. Criticism explicates this purposive organization, and critics are sometimes obliged to impute a design to a work of art rather than merely to find it inhering in the work. Under those circumstances, the critic turns from description to interpretation. Thus, the very nature of art leads to interpretation as a legitimate, and even inescapable, practice of critics.

An even more decisive difference between physical objects and works of art lies in the sorts of properties that are normally attributed to them. For instance, *expressive* qualities may be assigned to works of art: a novel such as Flaubert's *Madame Bovary* may be said to express a certain "bourgeois consciousness." Or *symbolic* qualities may be assigned: the castle in Kafka's novel of that name has been said to symbolize divine redemption. Or, *representational* qualities may be assigned: Picasso's painting "Guernica" is called a representation of the horrors of war. Also, *meanings* may be assigned in various ways: in *Hamlet*, for example, the meaning of Laertes' exchanges with Polonius and Claudius is said to be in their contrast with Hamlet's self-doubt; and the meaning of Bigger Thomas' dilemma in Richard Wright's novel *Native Son* is said to derive from the plight of the Negro in America.

As has been noted, it is difficult to say whether properties such as these are to be found *in* works of arts or only may be imputed *to* them. Out of respect for the artist, critics may assume their own role is subsidiary—*i.e.*, that of identifying properties of a work that might not otherwise be appreciated. Because it cannot be said with certainty, however, whether expressive, symbolic, or representational qualities or meanings and the like are actually *in* works of art, criticism must go beyond the severer canons of description to those of interpretation. This issue of the autonomy of criticism has been debated in the works of T.S. Eliot, Oscar Wilde, and many others.

Thus, as has been shown, critics may well offer incompatible accounts that purport to describe the same work. Even in cases in which the identity of a work of art is closely associated with its physical properties, as a work of sculpture might be associated with a mass of cast bronze, pure description may be adequate only for its physical properties but not for the properties that may be ascribed to a work of art. If a line in a painting is called "humorous," it is not always clear whether the line is being described or interpreted. What is clear is that critical interpretation must conform to certain minimal constraints. It must be compatible with whatever is descriptively true of the work in question; any interpretation that is compatible only with what is descriptively false is inadmissible. Interpretation, therefore, depends on some indisputable range of critical description.

Constraints on interpretation

The limits of this range, however, are not easy to establish. To do so requires deciding which among competing critical practices is the correct one. Such a comparison of critical hypotheses leads to metacriticism—that is, to the evaluation of the critical norms themselves, an extremely difficult matter. Without becoming entangled in this problem, some pragmatic rules of critical practice may be laid down: it should be possible to formulate its canons, these canons should be used by a community of practitioners, they should be effective in confirming or denying relevant claims, and it should not be necessary to distort them to examine new works or works that were not anticipated when they were adopted.

These rules permit a number of critical approaches and, therefore, allow alternative and incompatible interpretations of given works of art. It would be difficult, however, to justify the restriction of critical practice to any highly specialized way of proceeding. In literary criticism, for example, it would be difficult to reject out of hand the legitimacy of many diverse but fruitful approaches, such as the archetypal criticism of the Canadian critic Northrop Frye or the Marxist criticism of the Hungarian statesman and writer György Lukács or the Freudian

criticism of the British psychiatrist Ernest Jones or the neo-Aristotelian criticism of the Chicago critic Ronald S. Crane or countless others.

An extreme view of criticism holds that the interpretation to be preferred is the one that maximizes the value of the work in question. This view, however, does not recognize the rigour and independence of criticism or its obvious analogies to other endeavours, such as science, that are concerned with claims to truth and validity.

The objectivity of critical practice may be conceded not only to description but also to interpretation. Since interpretation is concerned with plausibility rather than with truth, plural approaches to works of art may reasonably be tolerated. The appreciation of a given work may entail a canvass of all plausible ways of construing it. Moreover, different schools of criticism may prove variously fruitful for each of the arts or for the movements within them. A Freudian approach to music or architecture, for example, promises less than formalist approaches; without question, however, the Freudian approach would be more fruitful in literature than in architecture. To examine the diversity of tenable critical practices (and there are new ones all the time) is to doubt that all but those consistent with some supreme critical principle can be disqualified.

VALUE JUDGMENT IN CRITICISM

In addition to the identification, description, and interpretation of works of art, the problem of their evaluation has been the subject of a number of disputes. The issues are those of value theory in general: there seems little reason to suppose that value judgments of works of art differ significantly from those of conduct, for instance. Immanuel Kant held that aesthetic and moral judgments are of fundamentally different logical kinds, but he was chiefly concerned with taste (which, despite Kant, appears in the moral, as well as the aesthetic, domain, in judging such matters as tact, decency, and personal ideals) and he failed to consider the implications of professional criticism and connoisseurship.

Nature of values

Although all the basic issues of the nature of values are not within the scope of this article, two fundamental issues are vital to a survey of the theory of criticism. For one thing, expressing one's tastes, preferences, likes, and dislikes must be carefully distinguished from judging merit. And for another, the principal varieties of value judgments must be sorted out. Quarrels about the critic's appraisal and evaluation of works of art are linked to these issues.

There is no prevalent philosophical view regarding value judgments that can be formulated in a simple and straightforward way. The account that follows puts forward one viable thesis, which may clarify how disputes about values can or cannot be resolved in a domain dominated by considerations of taste.

Value judgments are such in virtue of their predicates —that is, what they affirm or deny about the subject at hand. The predicates themselves, however, are not inherently valuational or nonvaluational; they are construed either way on the basis of governing theories. For example, to judge that Peter is tubercular may constitute a value judgment if good health is accepted as a norm. If tuberculosis is not regarded as a normative concept, however, but as a purely biological phenomenon, the judgment remains factual though not valuative. Factual judgments may be assigned truth values—that is, they may be determined to be true or false. Value judgments, on the other hand, are judgments that entail reference to a norm, or standard of merit. Some value judgments may also be factual judgments as well, as in the case of the judgment about Peter's tuberculosis.

Statements about one's likes or dislikes—one's tastes, in short—are usually not considered normative statements but statements of fact. They concern an aspect of a person's disposition and behaviour.

In making value judgments, however, the person puts a value on things by judging them in accord with norms that usually lie outside his own tastes. Of course it is possible that the values he places depend entirely on his tastes—his likes, dislikes, preferences—but it is not necessary. Herein lies a crucial distinction, for value judgments that depend on personal tastes are strikingly different from those that do not. Value judgments that depend upon taste may be called appreciative judgments and those that do not may be called findings.

Findings differ from factual judgments only insofar as they are linked to standards of merit; otherwise they are alike. In principle, critical judgments of the merit of a work of art, like other value judgments, may be as rigorously objective and confirmable as judgments of fact. Logically, findings are judgments of fact, in that truth values may be assigned to them.

It should be emphasized that questions of the relativity of values, or of their variability, are not relevant here. The logical properties of findings are not affected by the nature of socially institutionalized values. Value judgments about beauty or fashion may well derive from the prevailing tastes of a community, but, as findings, they will be made in accordance with those norms and not with one's personal taste (if we disregard coincidence).

In contrast, appreciative judgments are based entirely on one's own tastes. The well-known Latin maxim *de gustibus non est disputandum*, "there is no disputing concerning taste," applies to appreciative judgments only; it does not bear at all on findings.

Distinguishing between these two fundamentally different kinds of value judgments permits some rational debate respecting the merit of a given work of art. To the extent that a critical tradition is standardized (as the standards of wine tasting or of dog breeding have been established), critical findings may be regarded as objective relative to such a tradition. (The entire system may be challenged, of course, but that is another matter.) But appreciative judgments cannot be objective in that way; they cannot be simply true or false. Appreciative judgments depend on personal tastes, which vary greatly among individuals, so that seemingly contradictory judgments are common. If these judgments were treated as findings, such contradictory judgments might be found to be jointly true, a logical impossibility. If such a contradiction were resolved by regarding each judgment as true relative to given taste, then appreciative judgments could not be distinguished from mere statements about one's tastes. Since this distinctive kind of value judgment is prominent in aesthetic criticism, appreciative judgments may be conceded much weaker confirmation than findings. In a finding, one holds that a given work has (or lacks) a certain merit; the matter can be decided by attention to relevant criteria and evidence, as factual matters are. In an appreciative judgment, one assigns a certain value to a work in accord with one's personal likes, dislikes, and preferences. In findings, the supporting reasons must be publicly compelling; in appreciative judgments, the supporting reasons need only be relevant and reasonable to a critical public. The public may attest to the coherence of an appreciative judgment and its supporting reasons, but, because of differences in taste, it is not bound to share that judgment, as it must in the case of findings.

The problem of conflicting judgments

Incompatible findings, like incompatible descriptions, are inadmissible. Seemingly incompatible appreciative judgments, however, are admissible insofar as they are coherent (in accord with informal canons of such judgments) and compatible with the actual properties of the given work. In the context of an academic tradition, judgments of beauty may well be findings; but, in the context of personal taste, judging beauty is rendering an appreciative judgment. One may (appreciatively) say that a certain woman is beautiful and offer supporting reasons that may seem forceful or not. But, if one claims, in the mode of findings, that a certain woman *is* beautiful relative to a particular tradition of values, the evidence will serve flatly to confirm or deny the claim.

Disputes may arise about judgments of both sorts, and there are procedures for appraising each type. The precept about the pointlessness of arguing about taste should

therefore be amended to say that the kind of rational dispute possible for findings is not possible for appreciative judgments, or judgments of taste. Nonetheless, rational dispute of a different sort is possible for appreciative judgments as well.

The two types of judgment—findings and appreciative judgments—cannot be segregated on the basis of their respective range of predicates; an assertion that something is beautiful, for example, may be construed as either type of judgment. The conditions under which relevant claims may be confirmed is the key factor in distinguishing between the two types.

Other aspects of value judgment

There are, of course, deeper questions about values than can be treated in this article. Especially pertinent is this question: are values discoverable in nature, or the real world, or can they be construed only as an expression of the interests of a particular community? If it were possible to prove the existence of "true" or "real" values, there would be no need to account for appreciative judgments; practicing critics would be concerned only with findings—that is, with making judgments that are in accord with those true values. In fact, however, no such true values have been proved, and the critic is chiefly concerned with appreciative judgments, in which his personal tastes are systematically articulated. In any culture there may be a number of critics operating at cross purposes, each imprinting his own distinctive taste on a community of enthusiasts. Revolutionary tendencies in taste appear, however, resulting in significant alterations in critical practice. Thus, the judgments of critics tend to be informal and variable; because of the logical properties of their appreciative judgments and the conventional nature of the norms that support their findings, it could not be otherwise.

These, then, are the principal conceptual issues bearing on the systematic practice of criticism of the arts. The critic's judgments may be either valuational or nonvaluational. With respect to the former—that is, his value judgments—findings must be distinguished from appreciative judgments; with respect to the latter, the nonvaluational judgments, descriptive judgments must be distinguished from interpretive judgments. Because appreciative and interpretive judgments do not allow truth values to be assigned to them in the way that findings and descriptions do, the critic must tolerate logically weak judgments at the heart of his enterprise. This tolerance, however, should not be misconstrued as a lack of rigour in applying critical canons. On the contrary, rigorous critical practice requires such tolerance.

VARIETIES OF CRITICISM

Internal and external criticism

Actual critical practice raises certain technical questions that tend to divide considerations that are internal from those that are external to aesthetic criticism—that is, those that bear on the work of art as such from those that involve ulterior appraisals or characterizations. In spite of disputes regarding the proper scope and orientation of criticism, the diverse schools of criticism exhibit broadly parallel practices in both the internal and the external considerations to which they address themselves. For example, Marxist critics such as György Lukács and Christopher Caudwell use the socioeconomic categories of their philosophy to analyze works of art as art, to characterize the artist's life, and to study the role of the work within the dynamics of social change. Clearly, it is not always easy to discriminate between internal and external criticism.

The issue here is conceptual; it is not open to ideological quarrel. For instance, it has been said that the ancient Greek dramatist Aeschylus in his trilogy the *Oresteia* was presenting an idealized version of emerging notions of political justice. This external interpretation presupposes an internal interpretation of the work that does not violate any canons of aesthetic criticism, or it would not be tenable. Precisely the same consideration applies to the attempt by the Freudian critic Ernest Jones to psychoanalyze Shakespeare on the basis of *Hamlet* or Freud's attempt to analyze Leonardo da Vinci from his drawings and paintings. The principle that the external and the internal criticism must match remains the same whether the criticism is concerned with libidinal conflicts as are the Freudians, with archetypes as are the followers of the Swiss psychologist Carl Gustav Jung, with class conflicts as are the Marxists, or with theological and moral resolutions of modern poetry as is the French Thomist philosopher Jacques Maritain. In the light of the pluralism of viewpoints that has already been shown inevitable, it is not surprising that internal and external criticism cannot be sharply opposed to each other. Although adherents of some particular approach may disagree, it is difficult to see how any one approach can be exclusively vindicated. In a very real sense, the preference for any given approach may be construed as an appreciative judgment in itself.

Advantages of pluralism

It is not at all necessary that a theory about some aspect of human society be true in order for it to be fruitful and defensible in criticism. In describing, interpreting, and evaluating a work, the critic requires only that the theory is relevant to it. For instance, the critic may appropriately characterize or appraise a work in terms of Marxist, Freudian, Jungian, or Thomist doctrines to the extent that the imagination of either the artist or the audience is informed by such doctrines. Such doctrines need only be recognizable and socially significant to be useful in criticism; scientifically valid theories, on the other hand, may be entirely uninstructive when applied to the description and interpretation of a work of art. False theories may be so rewarding in discussing particular works that they cannot be dismissed solely because they are not true. To understand Wordsworth in terms of some version of Platonism is not to subscribe to Platonism itself, and to read the *Oresteia* in Marxist terms need not presuppose that Aeschylus held similar sociological convictions. The truth, or at least the plausibility, of the descriptive, interpretive, and evaluative judgments of critics must be carefully distinguished from the truth of the doctrines advanced by them or by artists. The validity of Thomism, for instance, is an entirely separate issue from the validity of interpretations of Dante from the Thomist point of view.

Those critics who believe they subscribe to a science of values, whether it may be Freudian, Thomist, Marxist, or some other, would quarrel against such a use of diverse doctrines. Inasmuch as the foundations of their beliefs do not compel the belief of all others, however, the use of doctrines of doubtful validity in criticism must be tolerated.

A doctrine may be applied by a critic if it is compatible with his own practice of criticism and if it provides an account of the design of that work, especially those elements that seem to call for interpretation.

In sum, the discipline of criticism has been shown to exhibit significant rigour, despite the logical weaknesses of the judgments of interpretive criticism and the uncertainty of the doctrines it applies. Contributing to this rigour are the requirements that an interpretation not deny any descriptive elements; that it be plausible; and that it deal with the traditionally salient problems of the work—that is, the interpretive puzzles that obstruct the assignment of a coherent and comprehensive internal order to the work.

A final issue may serve to round out this account of the criticism of the arts. As has been seen, an interpretation must be compatible with what is descriptively true of a given work of art. It is sometimes supposed that both description and interpretation should be similarly constrained by historical and biographical considerations. In fact, some versions of this thesis hold that interpretations prove to be no more than description provided under relatively difficult conditions of discovery. In the historical thesis, the "meaning" (*i.e.*, the correct characterization) of a work of art is the one that accords with its critical reception in the culture in which it was produced. The biographical thesis holds that "meaning" is what accords with the artist's intention.

Historical and biographical criticism

A number of things may be said about both theses. For

one thing, the two theses may be construed as the same wherever the artist's intention is interpreted in terms of the conventions governing such works rather than in terms of his own avowals of what he intended. It is extremely difficult, however, to suppose that one and only one construction can be put on a given work, especially in complex societies in which artists often intentionally depart from academic or other conventions. The problem of plausible alternative interpretations haunts critical practice even when the problem is confined within historical limits, even though the elimination of such alternatives is often claimed to be an advantage of historical criticism.

Furthermore, plausible interpretations that are put forward long after the work of art is produced may also be construed as the work of historical criticism; the restriction of interpretation to the period in which the artist worked represents a somewhat arbitrary constraint on the practice of criticism. Some works of art attract critics and audiences over generations, or even ages, without requiring close attention to historical or biographical origins. The force of accurate historical criticism cannot be denied, but criticism that ignores the *evolving* historical reception of a work of art unjustifiably restricts the scope of criticism.

A historical hypothesis, particularly one concerned with cultural significance, bears a striking similarity to interpretive criticism; like interpretive criticism, history as a discipline must rely substantially on criteria of plausibility. It should be noted that the same arguments that apply to historical and biographical criticism also apply to criticism based on cultural anthropology or the history of ideas.

Historical misunderstandings and inaccuracy about a given work of art must be distinguished from the historically changing significance of that work. A Freudian interpretation of *Hamlet*, for example, cannot be dismissed solely because the Freudian system was unknown to Shakespeare. If audiences that share Shakespeare's cultural and linguistic traditions are deeply influenced by the Freudian outlook, then a Freudian reading of his work, if independently plausible, cannot be regarded as a historical mistake. Interpretations that are constrained by *special* historical and biographical considerations must be distinguished from those that are not. Interpretations of both sorts may be welcomed.

Intentional criticism

Further objections may be raised against biographical criticism that narrowly construes the artist's intention as what may be documented in conversation, letters, and the like. Quite often, there is simply no such documented knowledge of the artist's intention. Even when there is, it may prove vague and ambiguous or even irrelevant to the appreciation of the work. In any case, any interpretation based on a statement of the artist's intention is subject to precisely the same sort of confirmation that other interpretations would be.

Intentional criticism may be viewed as a form of historical criticism that assumes certain types and modes of artistic endeavour within a tradition (to which the artist's intentions are said to correspond).

Some critics opposing what the American critics W.K. Wimsatt and Monroe Beardsley called the "Intentional Fallacy" would not admit any criticism based on the artist's intention; they would rule it out on the grounds that it violates the proper aesthetic concern of the critic. In view of what has already been said in this article about the nature of a work of art and of its boundaries, however, intentional criticism may be viewed as simply another selective mode of critical practice, one that is by no means exclusively correct but not for that reason inadmissible.

In conclusion, then, the principal issues affecting critical practices in the arts may be said to be (1) the identity and individuation of works of art; (2) the nature and characteristic properties of works of art; (3) their description and interpretation; (4) value judgments about them: and (5) the admissible varieties of, and constraints upon, criticism of the arts.

BIBLIOGRAPHY. Perhaps the most comprehensive canvassing of the current philosophical literature respecting criticism of the arts is to be found in M.C. BEARDSLEY, *Aesthetics: Problems in the Philosophy of Criticism* (1958); and J. MARGOLIS, *The Language of Art and Art Criticism: Analytic Questions in Aesthetics* (1965). Each of these two books (with extensive bibliographies) proposes its own theory of criticism; the first reduces interpretation to description, while the second opposes such reduction. Of relatively recent theories of art bearing on the problems of criticism, the following are among the most penetrating: N. GOODMAN, *Languages of Art: An Approach to a Theory of Symbols* (1968), primarily addressed to notational problems in identifying works of art; R. WOLLHEIM, *Art and Its Objects* (1968), pursues the paradoxes of numerical identity regarding works of art; S.K. LANGER, *Feeling and Form* (1953), attempts a systematic ordering of the different arts in terms of symbolic form; and H. OSBORNE, *Aesthetics and Criticism* (1955), construes criticism primarily in evaluative terms and the work of art as the object of such practice. Recent anthologies of the most widely discussed papers include: W. ELTON (ed.), *Aesthetics and Language* (1954), a linguistically oriented attack on the Italian philosopher Benedetto Croce and on Idealist aesthetics; and J. MARGOLIS (ed.), *Philosophy Looks at the Arts: Contemporary Readings in Aesthetics* (1962).

(J.Ms.)

Arts, Fraudulence in the

The most common type of fraudulence in the arts is forgery—making a work or offering one for sale with the intent to defraud, usually by falsely attributing it to an artist whose works command high prices. Other fraudulent practices include plagiarism, the false presentation of another's work as one's own, and piracy, the unauthorized use of someone else's work, such as the publication of a book without permission of the author; both practices are generally in violation of copyright laws.

Types of fraudulence

Forgery most often occurs with works of painting, sculpture, decorative art, and literature; less often with music. Plagiarism is more difficult to prove as fraud, since the possibility of coincidence must be weighed against evidence of stealing. Piracy is more often a business than an artistic fraud; it frequently occurs in the publication of editions of foreign books in countries that have no copyright agreement with the nation in which the work was copyrighted. A stage production, the reproduction of a painting, the performance of a musical composition, and analogous practices of other kinds of works without authorization and royalty payments also fall into this category.

The fundamental consideration in determining forgery is "intent to deceive." The act of copying a painting or other work of art is in itself not forgery, nor is the creation of a work "in the style" of a recognized painter, composer, or writer or of a particular historical period. Forgery may be the act not of the creator himself but of the dealer who adds a fraudulent signature or in some way alters the appearance of a painting or manuscript. Restoration of a damaged painting or manuscript, however, is not considered forgery even if the restorer in his work creates a significant part of the total work. Misattributions may result either from honest errors in scholarship—as in the attribution of a work to a well known artist when the work was in fact done by a painter in his workshop, a pupil, or a later follower—or from a deliberate fraud.

Excluded from the category of literary forgeries is the copy made in good faith for purposes of study. In the matter of autographs, manuscripts in the handwriting of their authors, forgeries must be distinguished from facsimiles, copies made by lithography or other reproductive processes. Some early editions of Byron's work, for example, contained a facsimile of an autograph letter of the poet. If such facsimiles are detached from the volumes that they were intended to illustrate, they may deceive the unwary.

Motivations for fraudulence

The commonest motivation for fraudulence is monetary gain. Fraudulence is most likely to occur when the demand for a certain kind of work coincides with scarcity and thus raises the market prices. Unprincipled dealers

have encouraged technically skilled artists to create forgeries, occasionally guiding them to supply the precise demands of collectors or museums. This is by no means a modern phenomenon: in the 1st and 2nd centuries AD, sculptors working in Rome made replicas of Grecian works to satisfy the demands for the greatly admired Grecian sculpture of the preceding five centuries. These copies or adaptations apparently were not offered as contemporary work, but as booty from Greece at the extraordinarily high prices paid for such works in imperial Rome. Similar circumstances may account for the "discovery" of a manuscript or autograph by a dead author or composer, although many such finds are quite legitimate and have been authenticated.

The history of the arts reveals instances of persons who have used forgery either to gain recognition of their own craftsmanship or to enjoy deceiving the critics who had rejected their genuine work. A legend told about Michelangelo illustrates this point. At the age of 21, he carved in marble a small sleeping Eros, or Cupid, based on ancient Roman works that he admired. Some time later this carving was sold as an antique to the well-known collector Cardinal Riario, who prized it highly. When Michelangelo stepped forward and claimed the work as his own he won immediate fame as a young man who could rival the work of the greatly venerated ancient sculptors.

Two further motivations behind forgery must be noted: forged documents have been produced from time to time to exalt or malign some religion, political party, or race; and forgeries are sometimes created as a hoax. Some hoaxes are intended to confound or ridicule the experts; others are intended to parody or burlesque an artist or genre.

Three methods of forgery

There are basically three methods of producing a forgery: by an exact copy, by a composite of parts, and by a work done in the style of an artist or period and given a deliberately false attribution. These methods apply most directly to the visual arts but can be discerned in literature and music as well. (Ed.)

LITERARY FORGERY

Financial gain is the most common motive for literary forgery, the one responsible for the numerous forged autographs that appear on the market. The popularity of such authors as the Romantic poets Burns, Shelley, and Byron led to the fabrication of numerous forgeries of their autographs, some of which remain in circulation. These forgeries were usually made by men who had access to only one or two genuine specimens, which they began by tracing. Their forgeries are stiff, exaggeratedly uniform, and lacking in the fluency and spontaneity of genuine autographs.

Instances of literary forgery

Occasionally a forger appears with a certain specious glamour like Constantine Simonides (1824–67), a Greek adventurer who varied his trade in perfectly genuine manuscripts with the sale of strange concoctions of his own. Maj. George de Luna Byron, alias de Gibler, who claimed to be a natural son of Byron by a Spanish countess, successfully produced and disposed of large quantities of forgeries ascribed to his alleged father and to Shelley, Keats, and others. More commonplace is the Edinburgh forger A.H. ("Antique") Smith, who was responsible for forgeries of Robert Burns, Sir Walter Scott, Mary Stuart, and other persons from Scottish literature and history—a feat that ultimately earned him 12 months' imprisonment.

Particularly notorious was the case of the Wise forgeries. Thomas James Wise (1859–1937) was one of the most distinguished private book collectors on either side of the Atlantic, and his Ashley Library in London became a place of pilgrimage for scholars from Europe and the U.S. He constantly exposed piracies and forgeries and always denied that he was a dealer. The shock was accordingly the greater in 1934 when John W. Carter and Henry Graham Pollard published *An Enquiry Into the Nature of Certain Nineteenth Century Pamphlets*, proving that about 40 or 50 of these, commanding high prices, were forgeries, and that all could be traced to Wise. Subsequent research confirmed the finding of Carter and Pollard and indicted Wise for other and more serious offenses, including the sophistication of many of his own copies of early printed books with leaves stolen from copies in the British Museum.

No forgery to attain recognition is better known than the "Thomas Rowley" poems of Thomas Chatterton (1752–70), which the youthful author attempted to pass off as the work of a medieval cleric. These poems, which caused a scholarly feud for many years, were influential in the Gothic revival. Chatterton, however, enjoys a place in English letters as a creative genius in his own right. The more conventional forger William Henry Ireland (1777–1835) cheerfully manufactured Shakespearean documents until his forged "lost" tragedy *Vortigern and Rowena* was laughed off the stage at the Drury Lane Theatre, London, in 1796. More fortunate was Charles Bertram, who produced an account of Roman Britain by "Richard of Westminster," an imaginary monk. Bertram's dupe, the eccentric antiquary Dr. William Stukeley, identified the monk with the chronicler Richard of Cirencester, known to have resided at Westminster in the 14th century. Bertram's forgery (cunningly published in a volume containing the works of two genuine ancient authors, Gildas and Nennius) had an enormous influence upon historians of Roman Britain, lasting into the 20th century. Equally influential were the Ossianic poems of James Macpherson (1736–96), which influenced the early period of the Romantic movement. To what degree Macpherson's poems are to be regarded as spurious is not certain. Denounced in his own day they were possibly, as he claimed, based upon a genuine oral tradition of Scottish Gaelic poetry; but there can be little doubt that they were carefully edited and interpolated by their collector.

Among the forgers who have tried to make the experts look foolish is George Psalmanazar (1679?–1763). A Frenchman, he went to England where he pretended, with great success, to be a native of Formosa (Taiwan), and published a book about that island, which he had never visited. Another is William Lauder, who attempted to prove Milton guilty of plagiarism by quoting 17th-century poets who wrote in Latin, into whose works he had interpolated Latin translations from *Paradise Lost*. A forgery made as a joke but taken seriously was the "Ern Malley" poems, offered to an Australian magazine in 1944 as the work of a recently dead poet. Actually it was composed by two young soldiers who wished to ridicule certain aspects of contemporary poetry.

The pure fabrication is a kind of forgery that defies classification, often because there is no false attribution and the motives are difficult to ascertain. An example of this is the *Historia regum Britanniae* (*c.* 1135) of Geoffrey of Monmouth (d. 1155), a pseudo-historian who compounded stories from Celtic mythology and classical and biblical sources into a fictitious history of ancient Britain. The book became one of the most popular of the Middle Ages and was the basis for some Arthurian legends recounted in medieval romance and epic.

Detection of literary forgeries. The scientific examination of a forged document may demonstrate its spurious character by showing that the parchment, paper, or ink cannot belong to the period to which they pretend. A skillful forger takes care, however, to secure appropriate materials; and in any case, scientific examination will not avail against the contemporary forger, living in the same age as his victim. Accordingly, other tests must be employed.

Forgeries may be detected by the methods of examination formulated by Jean Mabillon in his great work *De re diplomatica* (1681), for determining the authenticity of a document by the writing and the style of the terminology. These techniques have developed during three centuries into the modern sciences of paleography and diplomatic, by which various scripts and formulas can be assigned to particular ages and localities, and effective comparison can be made between two examples of handwriting pur-

porting to come from the same pen. Thus it is possible to state that a particular document could not have been written at the date that it bears. In dealing with printed texts, analogous methods are employed.

Nevertheless, a forgery may pretend to be no more than a copy of a genuine original. It then becomes necessary to examine the language and style in which it is written and to look for anachronisms or for statements that conflict with known authorities. This is the method of textual criticism brilliantly employed by Richard Bentley in his *Dissertation upon the Epistles of Phalaris* (1699), which proved that these letters, far from being written by a Sicilian tyrant of the 6th century BC, were, in fact, the work of a Greek sophist of the 2nd century AD.

While the detection of the careful forger may require an expert, forged literary autographs can often be detected by anyone taking the trouble to compare them with an authentic example. Many collectors have been deceived by their own credulity, because they wished to believe that they were getting a good bargain and subconsciously suppressed their critical faculty. A classic case is that of the French forger Vrain-Denis Lucas, who sold a collection of forgeries including a letter of St. Mary Magdalene, written in French on paper made in France. (Ge.B.)

FORGERY IN THE VISUAL ARTS

Any art object—paintings, sculpture, jewelry, ceramics, fine furniture, and decorative pieces of all kinds—can be forged. The difficulty of forging, however, is as important as market price in determining what is forged. Probably fewer than 1 percent of stone sculptures are false because they require so much labour to make and their market is limited, but as many as 10 percent of modern French paintings on the market may be forgeries. The technical difficulties in making a convincing imitation of an ancient Greek vase are so great that forgeries are almost nonexistent. In contrast the forgery level of tiny archaic Greek and Cretan bronze statuettes, which are simple to cast, is possibly as high as 50 percent. A forger is most likely to succeed with a mediocre piece in the middle price range because such a piece probably will never be subjected to definitive examination. Although the price should be low enough to allay suspicion, the object can still yield a fair return for the effort expended by the forger.

The copy is the easiest forgery to make and is usually the easiest to detect. When a duplicate has appeared the problem is merely to determine which is the original and which is the copy. At least a dozen excellent replicas of Leonardo da Vinci's "Mona Lisa" exist, many of them by his students. Various owners of these copies have at various times claimed that they possess the original. The Louvre is satisfied that it owns the painting by Leonardo because close examination reveals slight changes in the composition underneath the outermost layer of paint, and because this painting has an unbroken record of ownership from the time that the artist painted it.

Copies

A monumental sculptural forgery was a copy based on a Greek bronze statuette of a warrior of 470 BC, only five inches high and located in the Antikenabteilung, Berlin. The forgers made an eight-foot-high reproduction of it in terra-cotta and offered it as an Etruscan masterpiece. The resemblance was noted by the experts, who thought it to be an example of an Etruscan artist borrowing a Greek design motif. In 1961, after it had been in the Metropolitan Museum of Art in New York for 40 years, an analysis was made of the black glaze that covered the figure. It was found that the glaze contained as a colouring agent manganese, which never was used for this purpose in ancient times. Finally, Alfredo Adolfo Fioravanti confessed that he was the sole survivor of the three forgers.

Fine examples of pottery and porcelain have always commanded high prices, which have, in turn, encouraged the making of forgeries and reproductions. Since many European factories tried to imitate Italian majolica during the 19th century when it was especially popular, forgeries are common. The work of Urbino, Castel Durante, Faenza, and Gubbio was copied freely, and, to a lesser extent, so were the wares of Orvieto and Florence. Most of these forgeries are not close enough to deceive a reasonably expert eye. Potters used natural deposits the impurities of which, for good or ill, often affected the final result; until recently it has been impossible to procure materials in a pure state. In all but a few isolated instances (some German stoneware reproductions, for example) the forger no longer has access to these original deposits and he has to imitate the effect of the impurities as best he can. Although the best forgeries are often remarkably close to the originals, they are not very numerous.

In the composite fraud, or *pastiche*, the forger combines copies of various parts of another artist's work to form a new composition and adds a few connecting elements of his own to make it a convincing presentation. This type of forgery is more difficult to detect than the copy. Such a combining of various elements from different pieces can be very deceptive, because a creative artist often borrows from his own work. In fact, the similarity of a figure or an object in a forgery to that in a well-known work of art often adds to the believability of the new creation.

Composition fraud

The Dutch forger Hans van Meegeren employed a combined composite and stylistic procedure when he created seven paintings between 1936 and 1942 based on the work of Jan Vermeer. In "Christ at Emmaus" he combined figures, heads, hands, plates, and a wine jar from various early genuine Vermeers; it was hailed as a masterpiece and the earliest known Vermeer. Ironically, Van Meegeren never was detected as a forger. At the end of World War II he was arrested for having sold a painting attributed to Vermeer to one of the enemy, and was accused of being a collaborator. He chose to reveal himself as a forger, which was a lesser offense, and proved his confession by painting another "Vermeer" while in prison.

A variation in composite forgery, quite common with inlaid French furniture, involves the use of parts from damaged but genuine pieces to create a single complete piece that may or may not resemble one of the pieces from which it has been made. These made-up pieces are still considered forgeries. In composites of archaeological material only one part may be ancient, the balance being made up to complete the object. The head of a small terra-cotta figure may be ancient, the body and limbs of modern workmanship. A single ancient element in a composite forgery will help to deceive the buyer.

Most difficult of all to detect is the forgery done in the style of a particular artist or age. If the forger is skillful and is able to absorb the attitudes, conventions, and techniques of the period, he can often create a very successful piece of duplicity.

Stylistic fraud

The work of the Italian Alceo Dossena belongs in this class. He very competently forged works that were acquired by collectors and museums throughout the world. From 1916 to 1928 he produced hundreds of forgeries created as original expressions of archaic Greek, medieval, and Renaissance sculptors.

A newly discovered type of art inevitably brings on a flood of forgeries. At the end of the 19th century, when the first small, attractive Tanagra figurines were found in Greece, the market very shortly was flooded with a myriad of fraudulent Tanagra terra-cotta statuettes. In the mid-20th century, African primitive art became very popular, and woodcarvers from Italy to Scandinavia responded to supply the demand. Later, a very early civilization was discovered in Turkey, and the few genuine Anatolian ceramic pieces that appeared on the market were followed immediately by very competent forgeries apparently made in the same location as the ancient pieces. The lack of knowledge about genuine pieces made detection extremely difficult.

Detection of forgeries in the visual arts. The key to detecting forgery of unique objects lies in the fact that

every object has within itself evidence of when and where it was made. The two main approaches, stylistic and technical analysis, are complementary and are best used together.

Stylistic analysis

Stylistic analysis is subjective: it rests on the astute eye of the art historian. Each artist has a style, a flair, a verve unique to himself, and this can be recognized. His style will undergo change throughout his career, and this, too, can be stylistically analyzed and documented from his known works. When an unknown work purporting to be by a certain artist is discovered, the art historian attempts to fit it into the overall body of works by this artist. The subject matter, the brushwork, the choice of colours, and the type of composition are all consistent elements in a given artist's production. Any variation immediately arouses suspicion. When the idiosyncrasies of an artist's brushwork are studied, a fraud can sometimes be detected in much the same way a handwriting forgery is proven. In ancient works, particularly in antiquities, the scholar must examine the iconography of a piece. Forgers rarely have the scholarly background to combine iconographic elements correctly, and their errors often betray them.

An object must also be studied for its purpose. Ancient works were made for functional purposes. A forger usually makes an attractive piece often inconsistent with that purpose. As they were used most ancient pieces developed signs of wear. These rubbed and worn areas should appear in logical places on the object.

Documentation is also an important area of investigation. The apparent authenticity of many spurious pieces is bolstered by false documents to attest to the point of origin, former owners, and expert opinions concerning the pieces. A careful examination of these records often detects the forgery.

The hardest deception to detect is usually one that has been made recently. The forgery is a product of the time in which it was made, and the forger is closer to current understanding of the artist or period forged. The forgery therefore is often more appealing than a genuine work of art. As a forgery ages, viewpoints and tastes shift, and there is a new basis of understanding. Consequently, a forgery rarely survives more than a generation.

Technical analysis

Technical analysis, an objective approach, rests on an arsenal of equipment and tests. The fundamental principle is the comparison of a suspected work with a genuine work of the same artist or period. The suspected piece must show the same pigments or materials used and comparable age deterioration. Inconsistencies automatically cause the piece to be suspect. Oil paintings dry out and develop a crackle, bronzes oxidize, and ancient glass buried in the ground develops iridescent layers. The microscope is the most useful basic tool: a close examination of the physical condition often will show if the aging is genuine or has been artificially induced. The type of tools used by the artist can be detected from an examination of their telltale traces.

Ultraviolet rays readily reveal additions or alterations to a painting, since the varnish layers and some of the paint layers fluoresce to different colours. Ultraviolet is also used in the examination of marble sculpture. Old marble develops a surface that will fluoresce to a yellow-greenish colour, whereas a modern piece or an old surface recently recut will fluoresce to a bright violet. Infrared rays can penetrate thin paint layers in an oil painting to reveal underpainting that may disclose an earlier painting on the same canvas, or perhaps a signature that has been painted out and covered by a more profitable one. X-rays are used to examine the internal structure of an object. A carved wooden Virgin supposedly of the 15th century but revealing modern machine-made nails deep inside is obviously a fraud. A forger usually works for the surface effect and is not concerned with the internal structures.

Sometimes it is necessary to remove small bits of materials from a work and subject them to various analyses. Chemical analysis is particularly valuable in determining the pigment used because many of the paints available to the modern forger were unknown in earlier times. Today titanium, a 20th-century product, is used to make the white pigment in most oil paints, whereas white lead was the element used in the time of Rembrandt. Many ancient colours were manufactured by grinding natural minerals such as lapis lazuli for blue and malachite for green. Today cheaper synthetic chemicals are used. Some chemical tests, however, require the removal of more ancient material than is desirable. In that event a speck as small as the head of a pin can be analyzed spectrographically. From the burning of a minute sample a photographic record of the spectrum of the light emitted is analyzed to reveal the elements present and their relative percentages.

The dating of an object by the study of radioactive decay of carbon-14 has had little application in the detection of art forgery because of the large quantities of material that must be destroyed. Thermoluminescent dating is based on the slight damage to all matter, including clays, by the faint nuclear radiation present in the earth. Magnetic dating of ceramic objects is based on the slow but perceptible shift of the earth's magnetic field over the centuries.

Considerations of aesthetics and risk. One may logically question the real meaning of the difference between a genuine and a spurious work of art when in many cases it requires such expert study to detect the difference between the two. Or to phrase it another way, what is the difference in value of a work of art that has been on view in a museum for 40 years, after it has been proved to be false? This is a somewhat philosophical point in that the object itself has not changed, only our opinion of it. Its monetary value has been reduced from that of a rare, expensive, original piece to that of an attractive but spurious imitation. Its aesthetic quality has become a real danger, as it is a perversion of the truth. The forgery presents a false understanding of the work of an artist or an ancient culture, one which has been perverted in its modern translation. To appreciate the work of ancient artists their work must be studied alone and not be diverted by forgeries, or one will be inexorably misguided.

Despite all the studies and technical tests available, forgeries will still be made. The 20th-century art forger is far better equipped and much more knowledgeable than his predecessor. The demand for rare works of art has increased, and he will attempt to supply them. In collecting, whether by the private collector or by a museum, there comes a point when, after all the studies and all the tests are conducted, a decision has to be made as to whether or not to purchase a piece in question. The element of risk can be minimized but not eliminated. At this point, the collector should be ready to back his opinion with the purchase price. In order to acquire great pieces, particularly from newly discovered and relatively unknown cultures, it is necessary to take a calculated chance. The collector who has never bought a forgery probably has never bought a great piece of art. (J.V.N.)

BIBLIOGRAPHY

Literary forgery: J.A. FARRER, *Literary Forgeries* (1907, reprinted 1969), provides a good introduction, which may be supplemented by H.T.F. RHODES, *The Craft of Forgery* (1934); and S. COLE, *Counterfeit* (1955). For individual forgers and forgeries, see E.H.W. MEYERSTEIN, *A Life of Thomas Chatterton* (1930); W. ROUGHEAD, *The Riddle of the Ruthvens, and Other Studies*, rev. ed. (1936), on "Antique" Smith; T.G. EHRSAM, *Major Byron* (1951); A.N.L. MUNBY, *Phillipps Studies*, vol. 4 (1956), on Constantine Simonides; W.G. PARTINGTON, *Thomas J. Wise in the Original Cloth* (1947); and B.A. MORRISSETTE, *The Great Rimbaud Forgery* (1956). E.J. GOODSPEED, *Modern Apocrypha* (1956), gives an authoritative account of modern forgeries of Christian writings. On medieval forgeries in general, see the classic essay by T.F. TOUT, "Medieval Forgers and Forgeries," *John Rylands Library Bulletin*, 5:208–234 (1919); and for an example of forged charters, R.W. SOUTHERN, "The Canterbury Forgeries," *English Historical Review*, 73:193–226 (1958). On Geoffrey of Monmouth, see J.S.P. TATLOCK, *The Legendary History of Britain* (1950); and on the detection of forgeries: W.R. HARRISON, *Suspect Documents: Their Scientific Examination* (1958), and *Forgery Detection: A Practical Guide* (1964); and J.V.P. CONWAY, *Evidential Documents* (1959).

Forgery in the visual arts: SEPP SCHULLER, *Fälscher, Händler und Experten* (1959; Eng. trans., *Forgers, Dealers, Experts*, 1960); and HEINRICH SCHMITT (pseudonym FRANK ARNAU), *Kunst der Fälscher, Fälscher der Kunst* (1959; Eng. trans., *Three Thousand Years of Deception in Arts and Antiques*, 1961), are both standard anthologies. DIETRICH VON BOTHMER and JOSEPH V. NOBLE, *An Inquiry into the Forgery of the Etruscan Terracotta Warriors in the Metropolitan Museum of Art* (1961), is the exhaustive technical and art historical study of an important group of clever forgeries. A similar article is JOSEPH V. NOBLE, "The Forgery of Our Greek Bronze Horse," *Bulletin of the Metropolitan Museum of Art*, 26:253–256 (1968). Stories of frauds told from the viewpoint of the forgers are given in LAWRENCE JEPPSON, *The Fabulous Frauds* (1970).

(J.V.N./Ge.B.)

Arts, Practice and Profession of the

The practice of the arts goes back to the remote prehistory of mankind; the profession of the artist is of more recent origin. The oldest surviving works of art such as the cave paintings at Lascaux, France, and Altamira, Spain, dating from the late Paleolithic Period between 10,000 and 15,000 BC, were presumably made by men who never thought of themselves as artists. Though these works can be viewed as art today, it is probable that their creators were intent only on executing a magic ritual designed to aid them in the hunt.

At some point in the development and differentiation of human society, the professional artist emerged. The monumental buildings and stone carvings of ancient Egypt or Assyria and the intricately decorated pottery and elegant murals of Crete were undoubtedly done by trained and experienced craftsmen. They were no longer occasional practitioners of art, as the cave painters had been, but full professionals—*i.e.*, men skilled in a specialized occupation, practicing it full-time and probably earning all of their livelihood from it.

Professional and amateur artists. But even the relatively simple definition of a professional as a person receiving pay for carrying out a specialized occupation on a full-time basis is difficult to apply to the arts. In no other field do the categories of professional and amateur so overlap. One of the 20th century's major poets, Wallace Stevens, and one of its most innovative composers, Charles Ives, earned their living as insurance executives. By every criterion of proficiency and accomplishment, both deserve to be ranked as professional artists, yet in a real sense, both were amateurs. Similarly, it is difficult to classify Thomas Jefferson as a professional architect, even though he designed some of the most beautiful buildings of his time.

Historical and cultural dichotomies in the arts. The phenomenon of amateurism in the arts represents only one of the key problems in an attempt to discuss the professional artist. There is also the question as to whether the arts are professions, as distinct from crafts or skills. This aspect of the concept involves consideration of matters of the status or prestige of the arts and decisions on what constitutes proper preparation for careers in them. On these issues, attitudes have varied enormously from one art to another, from one historical period to another, and from one culture to another. Nearly everywhere and always, music, architecture, and poetry have been regarded as professions, while pottery making has been regarded as an art and granted the dignity of a profession only in some non-Western cultures.

Diverse evaluations of the arts

Since antiquity, there has been a continuing debate as to whether painting and sculpture demand skills of eye and hand alone or intellectual grasp and training as well. The painter has often spoken contemptuously of the sculptor, describing him as a mere stonecutter, and the sculptor, in turn, has looked at the painter as simply an illusionist and trickster.

The attempt to gain for some of the arts the status of learned or quasi-learned professions resulted in a distinction between "the fine arts" and "the applied arts," a distinction that has done harm to both. The notion that a painter has a profession, but a cabinetmaker or silversmith only a craft, helped to isolate the former from everyday life and to limit the creative enterprise of the latter. The gradual breakdown of the attitude that permitted such exclusive categories to be created and the general rejection of the false distinctions between art and utility may be among the most encouraging developments in the arts in recent times.

Of even greater importance has been the lessening of the separation that exists between the professional and the practitioner. The "cultural explosion" following World War II not only made the arts more readily accessible to the general public but also stimulated artistic activity in great numbers of people. Never have there been so many professional artists active in so many fields, and never have there been so many nonprofessional practitioners of the arts.

Preparation of the artist

Formal education for a profession in the arts is a relatively new development. The school of art or architecture, the conservatory of music, the academy of dramatic art, and the university's creative-writing program have emerged in modern times, some within the last generation or two. To most artists of the past it would have seemed incomprehensible that they should go to school to learn their profession. Neither would any of the great educational institutions before this century have thought it part of their function to teach artists. Indeed, serious doubts continue to be expressed as to whether academic schooling is the best training for most branches of the arts.

Formal instruction in the visual arts—painting, drawing, sculpture, and architecture, as well as the many arts of decoration and design—has the longest and most diversified history. A consideration of the patterns of development in these related fields will reveal the general directions of training in all the arts as well as many of the basic problems running throughout the history of art education.

THE VISUAL ARTS

The ideal of the universally trained artist

Throughout much of history, the insistence that the artist be learned in many fields has constituted a deliberate attempt to raise his status, to rank him as a practitioner of the liberal rather than the mechanical arts. The Roman architectural theorist Vitruvius (fl. 1st century BC) declared:

> [The architect] should be a man of letters, a skilful draughtsman, a mathematician, familiar with historical studies, a diligent student of philosophy. acquainted with music; not ignorant of medicine, learned in the responses of jurisconsults, familiar with astronomical calculations . . . (From *De Architectura*, F. Granger [tr.], Harvard University Press, 1962.)

Such a prescription represented the ideal of the universally trained artist, and, so far as is known, no institution existed in antiquity to provide this or any other kind of formal training in the arts.

The traditions of apprenticeship. Until the 16th century, when the earliest academies of the arts appeared, the artist acquired his skills mainly through various systems of apprenticeship. He learned his trade as he practiced it under the instruction and supervision of a master. Sometimes the apprenticeship was regulated by a craft guild. Typically, a boy was bound to a master at the age of 14 and served for 7 years. This system was in force throughout most of the Renaissance, and it was under such rules that Michelangelo entered the workshop of Ghirlandajo and Leonardo da Vinci that of Andrea del Verrocchio. In the 17th century, the beginning artist came to be considered a pupil rather than an apprentice. He lived and studied in the home of a master for an indefinite period and was free to leave when he felt he had learned enough.

Functioning of the systems. Whether as apprentice or more informally attached pupil, however, the young artist was trained on the job by his master. As an apprentice, he began by doing the most menial jobs: grinding the colours, cleaning the brushes, or preparing the wood panel or the plastered wall. Gradually, he was trusted

The student, from menial to master

with more responsible tasks, given instruction in the technique of the art itself, and allowed to paint in some of the decorative details of an altarpiece or fill in large neutral areas of a fresco. Finally, he might be assigned the subordinate figures or landscape background of a painting by the master. It is said that Leonardo as an apprentice painted so beautiful an angel in a "Baptism of Christ" by his master that the master resolved to renounce painting and devote himself to sculpture.

Strengths and weaknesses of the system. The training of an apprentice in a Renaissance workshop was not confined to a single art. The workshop of Antonio and Piero Pollaiuolo, perhaps the most active art establishment in Florence of the mid-1400s, accepted commissions in sculpture as well as in painting of all sorts. The graduate of such a workshop, then, if he had kept his eyes open and his hands busy, could have developed a wide range of artistic skills. The apprentice system, however, had serious weaknesses. In a good workshop run by a conscientious master, an apprentice could receive excellent training. Far too many apprentices, however, undoubtedly were condemned to years of repetitious drudgery, receiving little systematic instruction and, at best, being trained purely as craftsmen. Moreover, the guilds deliberately limited the number of apprentices.

Academies and the artistic elite. The founding of the first academies of art, in the late 16th and early 17th centuries, was a step forward. Although later such institutions became restrictive and rigid, they were designed to break the monopoly of the guilds, to regularize the training of artists, and to lift artists from the category of mere craftsmen by liberalizing their education.

The French Academy as model. The most famous and influential of all academies of art was the Académie Royale, commonly known as the French Academy, founded in 1648 in imitation of the Accademia di S. Luca, which had been set up in Rome in 1593. Under the patronage of Louis XIV's great minister, Jean-Baptiste Colbert, the French Academy by the 1660s had become a major force. Under the directorship of the painter-decorator Charles Le Brun after 1683, it exerted a virtual dictatorship over French art. Its success inspired the creation in other countries of official academies that aped its organization and program.

Curriculum and rewards. The curriculum of the French Academy, fixed under Colbert and Le Brun, centred around drawing. The student drew from the drawings of his professors, then from casts, and finally from life. In addition, he attended academic lectures, analyzing pictures from the royal collection. These discourses were designed to instill in the student certain aesthetic criteria and a fixed hierarchy of values in which subject matter played a leading role. In the scale, historical paintings were regarded as the noblest, still lifes as the meanest.

The academy's students competed for a series of prizes culminating in the Prix de Rome, a four-year scholarship in Rome that assured its holder a successful career at the top of his profession. Even those who won no prizes could anticipate a secure future in the service of the state or of individual patrons, for the steady production of uniformly trained and indoctrinated artists was of great importance to the regime.

Academicism as a negative force

The academy performed a great service in breaking the stranglehold of the guilds on the training of artists, in raising the standards of art instruction, and in improving the status of artists. On the other hand, it locked the artist into a closed system, made conformity a virtue, and treated individualism and originality of style as sins to be avoided at all costs. By standardizing methods of instruction, it also standardized the bases of critical judgment. By admitting only 200 students and thus creating a small artistic elite, it neglected the need for trained artists and designers in the decorative and applied arts. Academicism established above all the principle that the artist succeeded not through native genius but through a correctness of technique that could only be acquired through proper training.

The English Royal Academy. In England, a number of schools of art were established during the early 18th century, notably one founded by the painter and caricaturist William Hogarth. In the employment of female models, these schools went beyond anything yet attempted in France. It was not until 1768, however, that the patronage of the crown was obtained for a Royal Academy. By the end of the century, this academy had achieved immense prestige and influence, and a painter who could sign the coveted "R.A." after his name was certain of a prosperous career. The *Discourses* delivered over a period of 15 years by the painter Sir Joshua Reynolds, the academy's first president, remain the best statement of the academic philosophy.

As a teaching institution, however, the Royal Academy fell far short of its French prototype. The limitation to 40 on the number of academicians increased the social and economic desirability of membership but narrowed the scope of the academy's teaching functions. The restrictive curriculum and the discouragement of "eccentricity" of style were not calculated to foster true talent. Of the outstanding English artists of the early 19th century, only J.M.W. Turner was trained and supported by the Royal Academy. The painter-poet William Blake despised it, and landscapist John Constable became an R.A. only late in his career.

Counter movements: the applied arts and crafts. It was not only its self-perpetuating and oligarchical exclusiveness that provoked criticism of the Royal Academy. With the growing impact of the Industrial Revolution, the academy was attacked by reformers in and out of Parliament for its failures to encourage the applied arts and to remedy the growing shortage of trained artisans in such fields as china manufacturing, ornamental metalwork and plastering, and carving in wood and stone. Parliamentary hearings sharply critical of the Royal Academy led in 1837 to the establishment of a government School of Design and an accompanying museum.

Schools and museums of design. The School of Design and its branch schools were succeeded, after 1852, by a network of art schools under a government Department of Practical Art. By 1864 there were 90 such schools instructing about 16,000 students. The founding, in 1852, of the Victoria and Albert Museum, which came to have vast collections of design in all fields, was an important step forward in art education. This combination of schools and museums of applied art was soon imitated in many of the leading cities of Europe, especially in Austria, Germany, and The Netherlands. Despite attacks on the basic principle of separating schools for the applied arts from those for the fine arts, such schools continued to flourish in the 20th century. Most of them, however, modified their curricula to include such studies as art history and aesthetics.

Waning of academic dominance. In painting and sculpture, however, the academic idea lost some of its vitality and dominant influence in the 19th century. By the mid-18th century, the French Académie Royale was under attack from the philosophers of the Enlightenment, and in 1793, at the height of the Reign of Terror following the French Revolution, it was dissolved and its functions handed over to a Commune des Arts led by the painter Jacques-Louis David. The Institut de France, set up in 1795 to supervise the arts, revived the academy in everything but name. The academic establishment, under different forms, remained a powerful though no longer controlling force in the French art world. As the École Nationale Supérieure des Beaux-Arts, it continues today to train artists to compete for the Grand Prix de Rome.

Continued influence of the masters. Even in the heyday of the academy, the system of training the artist in the master's studio had continued to flourish. In 1771 the Irish painter James Barry estimated that 5,500 persons were receiving some kind of art instruction in Paris alone and that of these, 1,500 were being trained to work in such industrial enterprises as the Gobelin tapestry works. Thus most artists were still trained in the studios of established artists or in elementary technical institutes, such as an industrial school founded by Jean-Jacques Bachelier in 1762. France's dominant position in the decorative arts during the 18th and 19th centuries can be attributed in

large part to the effectiveness of the Bachelier school and similar training centres.

During the second half of the 19th century, the demand for training in painting and sculpture grew so great that large numbers of beginners congregated in such studios, or ateliers, as the famous Académie Suisse in Paris, where they painted from a model under the often cursory supervision of a well-known teacher.

Copying the old masters

Another training method became important during the 19th century. Much earlier, copying of frescoes such as those by Masaccio in the Brancacci Chapel in Florence had been a major source of instruction for later painters. With the opening to the public of such great museums as the Louvre, the practice of painting from the old masters again became a basic feature of the artist's training. Even so sophisticated a painter as Paul Cézanne was devoted to this method of teaching himself.

Synthesis of art and craft. Despite 19th-century advances in the formal training of artists in all fields, sharp criticisms of prevailing art education were made, notably by the critic John Ruskin and the poet-craftsman William Morris. They opposed the separation between training for the applied arts and education for the fine arts. Morris saw the division as creating a group of designers who turned out standardized patterns for machine-made objects and a group of fine artists who were constrained from producing things needed by people for an aesthetically pleasing environment in daily life.

As a direct result of Morris' teachings, a number of art schools were established in England to revive handicrafts and to bridge the gap between the applied and the fine arts. The most successful of these schools was the London Central School of Arts and Crafts, founded in 1896. After 1900, however, the leadership in the movement to reform art education passed from England to Germany, where there was a decisive change in emphasis. Contrary to Morris' rejection of the machine, an attempt was made to realize the aesthetic possibilities of machine-made objects and also to create a new architecture.

Impact of the Bauhaus. The major figures in 20th-century art education were the Germans Bruno Paul and Walter Gropius. It was Paul's strong conviction that art schools had a responsibility to artists and to society: to train artists in fields in which they could both earn a living and be socially useful. After World War I, he combined the Berlin Academy of Art with the School of Decorative Art. In a pamphlet on art education published in 1918, he declared that all students, whether they intended to become fine artists or applied artists, should receive basically the same training. He stated also that no one should be admitted to an art school without first having learned a trade in a workshop or trade school.

Walter Gropius, in his early career in Germany and his later work in the United States, exerted an incalculable influence not only on art education but also on the whole development of modern art and architecture. Already a well-known architect in 1914, he became principal of the Weimar School of Arts and Crafts. After service in World War I, he returned to Weimar and in 1919 founded the Staatliches Bauhaus ("the State Architecture House"), a school merging the Weimar art school with the arts and crafts school. In 1925 the Bauhaus was moved to Dessau.

Unison of arts and crafts

The Bauhaus ideal was to unite all arts and crafts to create a new architecture that comprised a living environment; to break down the false separation between the applied arts and the fine arts, between art and utility; and to train artists in the creative possibilities of machine design.

The Bauhaus curriculum. To these ends, the Bauhaus curriculum emphasized the use of different tools and materials, as well as academic instruction in geometry, principles of construction and design, and the history of art. Instruction was divided into three stages. The first, lasting for six months, covered an elementary survey of Bauhaus principles and methods. The second, taking three years, was the basic practical and academic course in which the student, in addition to the general training he received, was required to specialize in a trade under the supervision of a particular master. At the end of this course, the student had to pass one of the regular city trade examinations before, finally, he entered the third phase, *Baulehre* ("building instruction"), in which he took an active part in one of the Bauhaus' projects.

During its short life, which ended in 1933, the Bauhaus achieved remarkable results and assembled an extraordinary faculty that included such leaders and innovators as painter-photographer László Moholy-Nagy, architect Marcel Breuer, and painters Wassily Kandinsky and Paul Klee. It had created the ideal of a new type of art education and proved that it was workable. John Ruskin, as Slade Professor of Art at Oxford in the 1870s, had taken his students out to repair roads, but it remained for the Bauhaus to transform Ruskin's experiments into a coherent program.

Acceptance of the arts in higher education. Though the Bauhaus had no immediate successor, it affected the teaching of art in many schools in following decades. Such education achieved notable growth in university-based or university-sponsored schools of art, especially in the United States. Such schools attempt, in many instances, to combine a four-year academic education with studio work in drawing, painting, sculpture, and other media. Many offer specialized work in such fields as industrial design, book illustration, advertising art, and costume design. The graduates of these schools usually are granted bachelor's or master's of fine arts degrees. None of the schools, however, has attempted to realize the Bauhaus ideal of a student completely trained in both the fine and applied arts.

The same is true of schools of architecture, most of which are connected with universities. In these, the academic curriculum is generally more rigorous than in schools of art. A degree is usually granted after a five-year course, which includes considerable work in traditional liberal arts subjects. A number of schools offer special concentrations in city planning. Few if any, however, see architecture as Gropius did, as a discipline unifying all of the crafts.

THE OTHER ARTS

Music. Next to schools for the visual arts and architecture, the most widespread of contemporary schools for the arts are those for the training of musicians, whether composers or performers. Like the school of art, the school of music has a long history.

The Paris Conservatoire model. The first full-fledged music school of modern times was the Conservatoire de Musique, founded in Paris in 1795 by the revolutionary National Convention as a successor to the earlier École Royale de Chant et de Déclamation and Institut National de Musique. In 1797 it had 125 professors and 600 pupils. It has had a continuous and distinguished history to the present day.

Development of the music curriculum

The curriculum developed at the Paris Conservatoire has been followed with slight variations in schools of music elsewhere. Originally, classes in composition, harmony, singing, and instrumental performance were given. Later, classes in music history were added, and an increased emphasis was given to training in sight reading. Still later, instruction in aesthetics and musical analysis became part of the standard curriculum.

The success of the Paris Conservatoire and the greatly increased demand for trained musicians during the 19th century led to the creation of conservatories in major European and American cities. In the United States, the Boston Conservatory of Music was founded in 1867, the Peabody Conservatory in Baltimore in 1868, and, later, others in New York, Philadelphia, and elsewhere. Some of these schools have remained independent, whereas others have become part of universities.

Private and institutional study. The conservatories of music have not altogether supplanted the older system under which a student worked under a particular master, as Ludwig van Beethoven did under Joseph Haydn at the end of the 18th century. Such a famous teacher of composition as Nadia Boulanger attracted students from all over the world, and singers and instrumentalists often

prefer to attach themselves to the studio of a favourite teacher. Increasingly, however, the conservatories have absorbed the master's classes.

Literature. The idea that the college-bred writer is superior to the one who has not had a liberal arts training goes back at least as far as the Elizabethan Age, when Robert Greene, one of the so-called university wits, sneered at Shakespeare as an upstart crow. The notion that creative writing, as distinguished from the ancient discipline of rhetoric, is something that can be taught in an academic environment is a recent development.

In the United States, especially, many creative writing programs have developed in higher education, usually in conjunction with the appointment of a well-known writer in residence as a member of the faculty. The creative-writing student typically receives more broadly based, and less technical, training than the student at a school of art or music.

Theatre and dance. The training of dramatists has taken a direction somewhat different from that of other writers. It has been combined with practical work and experience in the crafts of the theatre.

Academic teaching and apprenticeship. The workshop set up at Harvard by George Pierce Baker and the Yale School of Drama, which he founded in 1925, served as models for later drama schools. In recent years, some university-based schools of theatre arts have developed very elaborate programs, with professional courses in scenic design, lighting, and direction, as well as in playwriting and acting. They often sponsor highly trained repertory companies that present new and often controversial plays as well as the classics.

Training the performer

There also exist separate schools for actors outside the university establishment. Among the best known of these is the Royal Academy of Dramatic Art in London, which has trained many successful actors. The Actors Studio in New York City has enrolled fledgling actors as well as established actors who wish to keep their skills sharp.

The old apprenticeship system remains more a vital factor in the training of actors than in any other of the performing arts. Particularly in countries in which repertory companies are well-established, as in England, France, Germany, and the Soviet Union, actors can learn and gain experience in their craft by progressing from walk-on roles to more demanding ones. Often the actor can start in a provincial repertory company and move to one of the national companies. For the experienced actor, the repertory company provides the opportunity to do a variety of roles, and to go from a major to a minor part from week to week, as a member of an integrated and continuing ensemble. The system goes back to the beginning of stable acting companies in the Europe of the 16th and 17th centuries, when actors might enroll as boys and live out their lives as members of the same company.

Formal education for the dance in modern times dates from the establishment by Louis XIV of the Académie Royale de Danse in 1661. In 1669, the Académie Royale de Musique was founded, and a school to train professional dancers was added in 1672. These institutions made Paris, and specifically the Paris Opéra, the world centre of ballet training and performance in the 19th century.

Formal demands of the dance. Training for the dance has made less headway in universities than that for other of the performing arts. A few colleges in the United States offer a fully developed dance curriculum, and some schools of fine arts include programs in dance, but the professional instruction of the dancer, whether in ballet or the modern dance, usually is carried on in the 19th-century tradition of teaching in schools associated with ballet companies or opera houses. It has changed little from the time when Edgar Degas painted the young dancers of the Paris Opéra at practice. With few exceptions, training for the dance has remained a relatively narrow and physically demanding regimen.

The technologically based arts. Relatively little attention has been paid to the development of professional curricula in photography. More than any other art, it has been treated as a hobby to be learned without formal instruction. Where it is taught, the focus is on photographic technique rather than on photography as an art. There are, however, signs of a growing acceptance of photography as an important and expressive art form, and it is now becoming part of the curriculum not simply of schools of journalism but of schools of fine arts as well.

The new arts of the 20th century—film, radio, and television—still depend largely upon people trained in the older arts. The foundation in recent decades of film institutes associated with universities, notably in New York state and California, has been a significant step. A few radio and television programs exist, sometimes as parts of a department of communications arts or outside academic settings. Most of the directors, actors, and writers in film and television either have developed in the theatre or the literary world or have grown up in the industry without formal instruction.

THE SELF-TAUGHT ARTIST

With all of the many systems and institutions for formal training in the arts that have come into being over the centuries, there always have been self-taught artists. This is especially true of writers, the overwhelming majority of whom, from Homer to the present day, have had no specific training in their art, and many of whom, in fact, like Shakespeare or George Bernard Shaw, had comparatively little schooling of any sort. They have learned their trade by studying other writers and life itself, and by proceeding from imitation to original creation. Many people have turned to writing after education for careers in other fields. Medicine alone has produced such outstanding writers as the Englishmen Sir Thomas Browne and W. Somerset Maugham, and the American William Carlos Williams.

Art without schooling

Self-trained artists have also achieved prominence in other fields, though far less frequently than in writing. The history of architecture, perhaps of all the arts the one that seems most to demand strict technical training, contains two spectacular instances of the self-trained professional. Sir Christopher Wren, among the greatest of English architects, was a mathematician and professor of astronomy at Oxford before he took up architecture. But whereas Wren was grounded in the principles of the exact sciences, Sir John Vanbrugh, the architect of Blenheim Palace, had been a soldier and a dramatist before he turned to architecture at the age of 35.

The completely self-trained professional artist is a rarity in painting and sculpture. More than any other kind of artist, the painter especially must learn from and with other painters. Although such painters as Paul Cézanne, Vincent Van Gogh, and Paul Gauguin were in many respects self-taught, they absorbed a great deal from studies of, and interactions with, their fellow artists. Occasionally a true "primitive" such as Henri Rousseau, unhampered by conventional instruction, has produced a style with extraordinary power and directness as well as naïveté.

TRAINING OUTSIDE THE WEST

In non-Western cultures also, professional training in the arts has had a long history. Much of it has been carried out under systems of apprenticeship similar to those found in the West. Both in the Orient and in Africa, the profession of the arts has often been a hereditary one, sometimes, as in the case of bronze casters and ceramics workers in China, with secret techniques passed on from one generation to another. Similarly, Japanese Nō and kabuki actors long passed their craft from father to son, with the aristocratic Nō being regarded for centuries as a secret tradition. In West Africa, an entire village frequently has specialized in the production of a particular kind or style of art object.

Oriental and African art practices

Education in the arts has also been diffused through written treatises. In China a number of practical manuals on painting are extant, the two most famous ones dating from the 17th century. In India there were noted teachers of arts and crafts as early as the 1st century. Most recently, art schools similar to those in the West have been founded in a number of Eastern countries, notably in

Japan, where a conscious attempt has been made to combine ancient traditions and modern styles. During the period of French colonial rule in Southeast Asia, an influential École des Beaux Arts was set up in Hanoi. In many primitive and folk societies, however, such arts as dance have been passed on by imitative learning.

Conditions of work in the arts

In no way can artists be stereotyped as shivering recluses, painfully pursuing their craft in dimly lit garrets; as well-fed hangers-on to the coattails and whims of the wealthy; or as irresponsible wastrels, lurking in the outskirts and subcellars of respectable society. The conditions of their work and life have been shaped in large measure by the regard of their society for the arts in general or for their art in particular. High regard has not always brought success or satisfaction, however, and artists have turned into many different avenues in search of tangible or other rewards for their art.

THE STATUS OF THE ARTIST

The social and intellectual status of artists has differed considerably from one field to another, from one culture to another, and from one historical period to another. They have been regarded as entertainers, artisans, seers and prophets, and, often, as dangerous cranks or misfits. The history of each of the arts has been marked by a determined effort on the part of its practitioners to raise their status and to be accepted as honoured members of their society.

Prehistory. In preliterate cultures, the poet preserved and transmitted the beliefs and traditions that gave the culture its sense of identity and purpose. Less an original creator than a storyteller and singer, he kept alive the works that had grown up by gradual accretion and refinement, and often he contributed to the process. The Homeric epics of ancient Greece probably developed in this way over generations, though in the form in which they have survived they almost certainly were shaped by one or more supreme geniuses.

The primitive bard

The primitive bard seems to have occupied an anomalous position, held in awe because of his seemingly divine powers but without a fixed place in the social hierarchy—existing as a dependent of a tribal ruler or as a wandering beggar. It is no accident that the Greeks represented Homer as blind, thereby symbolizing not only the poet's inward-directed vision but also his separation from ordinary men. The peculiar contradiction that poets have complained of bitterly, that their gifts are venerated but that they themselves are neglected, has its most famous statement in the couplet, "Seven cities warred for Homer being dead,/Who living had no roof to shroud his head," by the English poet Thomas Heywood.

Antiquity. In the great age of classical Greece and, later of Rome, the artist lost some of his legendary quality as seer, but he gained considerably in social acceptance.

Greek poets, sculptors, and designers. The skill of the dramatist, like that of the athlete, was tested in competition and rewarded by prizes. The bard gave way to the historically defined individual. Sophocles, in the fifth century before Christ, was no blind beggar but a respected citizen, a soldier, and politician, as well as a tragic poet. Above all, the intellectual role of the writer was recognized and his claim to respect was validated by the celebrated dictum of Aristotle that poetry is more philosophical than history. Aristotle's defense of poetry was perhaps a deliberate response to the celebrated attack by Plato in *The Republic* on the creative artist as a potential threat to the stability of the state.

Emergence of the known artist

In other fields, too, artists began at the same time to emerge from anonymity and to achieve social acceptance. The nameless artisans and stonecutters who erected the first crude Doric temples and carved the cult images that they housed were succeeded by renowned architects such as Ictinus (flourished 5th century BC) and Callicrates, and by master sculptors such as Phidias (flourished 475–430 BC), who supervised the rebuilding of the Athenian Acropolis. Phidias is alleged to have represented himself and the Athenian tyrant Pericles on the shield of Athena. Such self-glorification would have been beyond the imagination of earlier generations.

The great mass of Greek sculptors of the classical and Hellenistic periods, however, continued to be regarded as craftsmen, paid standard wages for a day's work or for a specific piece. To Plato, for example, sculptors were common workmen. Even the recognized masters were expected to turn their hands to whatever their patrons demanded.

Rome and service to the state. Under the late Roman Republic and the empire, the artist continued to enjoy a favoured status. The great Roman dramatist Terence (flourished 2nd century BC) began his career as a slave, but his achievements earned him freedom and admission to the leading intellectual circles of his time. The organizing genius of the Romans enlisted poets, architects, and sculptors in the service of the state. Gaius Maecenas (died 8 BC), the first of the wealthy patrons of the arts, encouraged and supported such poets as Virgil and Horace. Vast building and engineering projects throughout the empire gave employment and prestige to many architects.

This was the age, too, that saw the emergence of the art connoisseur and the tourist in search of cultural treasures. The esteem for artists in Rome is reflected in the eagerness with which Pliny the Elder sought out and recorded biographical information on the great artists of Greece. For the first time the artist was felt to merit the same attention as the statesman or soldier. A similar rise in the prestige of the writer was signalized by the founding of such great libraries as the one at Alexandria, devoted to preserving the literature of the past.

The musician of antiquity. Much is known about the musician in antiquity, although practically none of the music itself has survived. There is considerable testimony to show that music itself apparently was considered among the highest of human accomplishments; in fact, of divine origin. The legend of Marsyas, the human being who dared to challenge Apollo to a musical contest and was flayed for his pains, was a favorite subject of Hellenistic sculpture. The myth of Orpheus, whose playing on the lyre could move even inanimate things to wonder and delight, reveals the almost superstitious awe in which the Greeks held the powers of the musician. Musical theory, with its close relationship to mathematics, was regarded as a branch of philosophy. The philosopher Pythagoras (flourished *c.* 530 BC), who saw the whole universe as a harmony of the spheres, was only one of many who gave the highest intellectual ranking to the study of music. The special status given to music in antiquity continued into the Middle Ages, when music alone, of all the arts, was ranked among the seven branches of learning.

Music as the highest art

In Greece, there emerged for the first time the distinction between the gentleman amateur and the paid professional that was to play so large a role in the later history of the arts. Every educated Greek was expected to be able to play an instrument, to sing acceptably, and to discourse on the theory of harmony. On a much lower intellectual and social level were the professional entertainers who competed for prizes at great public concerts, who performed in the dramas, and who supplied the musical accompaniment for athletic games.

Actors, dancers, and rhetoricians. The distinction between the creative artist and the entertainer or interpreter, which continues to be made even today, seems already to have existed in ancient civilization. The accomplishments of the actor-dancer were admired and applauded, but, aside from the citizen-dancer of the Greek festivals of tragedy, he had a relatively low intellectual and social status. The names and other details of some of the famous actors of this period are known, but references to them are in a tone altogether different from references to a Phidias or a Virgil, and everything indicates that actors were regarded as belonging to an inferior class.

The lost art of persuasion

One art virtually unpracticed today, that of the orator or rhetorician seeking to persuade, by speech or writing and according to a detailed formulistic pattern, occupied an especially favoured place in antiquity. The principles

of the art were expounded in treatises by Aristotle, Quintilian, and others, and its greatest practitioners, Demosthenes among the Greeks and Cicero (first century before Christ) among the Romans, were accorded the same respect given to the poets. In the twilight of the Roman Empire, the young St. Augustine supported himself as a teacher of rhetoric. The prestige of rhetoric as a noble and useful art reached its highest point during the Renaissance, when every court had its master rhetoricians and orators.

The Dark and Middle Ages. With the collapse of the Roman Empire in the West in the 5th century, the professional artist virtually disappeared in Europe. Until the revival of Western culture under Charlemagne about 800, the practice of the arts was confined mainly to scattered monastic enclaves, which kept alive the traditions of a written literature and began to develop a new architectural style. Some of the arts, such as the written drama, were lost altogether, only to be revived later on from church ritual and folk ceremonies. The sophisticated poet, a product of the stable and assured civilization of antiquity, vanished from the scene to be succeeded by anonymous bards, such as the creator of the Anglo-Saxon *Beowulf*, who, as in Homeric days, put into final form the results of centuries of oral shaping of legend and history. Only in the Byzantine Empire and, after the 7th century, in the rapidly growing world of Islām, was the professional artist able to flourish.

The rich burgeoning of medieval civilization following the early 9th-century Carolingian renaissance produced a magnificent harvest in the arts, but for the most part the creators remain unknown. The superlatively beautiful so-called Book of Kells from 9th-century Ireland was decorated, page by page, by a monk or group of monks, who saw the task as their gift to God, their way of living their vocation. It is doubtful that they were conscious of themselves as artists or that they were so regarded by their fellows.

The builders and sculptors of the great Romanesque and Gothic churches were journeymen masons and carvers who worked under the driving direction of a forceful churchman, such as the remarkable Frenchmen Hugh of Semur, Abbot of Cluny from 1049 to 1109, and Abbot Suger, who built the Abbey Church of St. Denis from 1140 on. The radiant stained-glass windows of the Gothic cathedrals were made in the workshops of Chartres and other centres. Only rarely did a particular craftsman, such as the sculptor Giselbertus of Autun and the sculptor-decorator Nicholas of Verdun, emerge from anonymity.

The Renaissance and Baroque. An enormous increase in the prestige of the artist came about with the beginnings of the Renaissance in 14th-century Italy. The veneration for antiquity that so dominated intellectual life was in large measure a veneration for the arts of antiquity, for a life-style in which the written and spoken word, the visual image, the public monument, and those who created them were seemingly central elements.

Reverence for the artist

New stance of the poet. The crowning with laurel of the poet Petrarch at Rome in 1341 was a ceremony intended to establish a link with ancient civilization. Almost immediately after the death of Dante, the *Divine Comedy* became the object of an intense scholarly and critical study equal to that given the great epics of antiquity.

Throughout Italy, but especially in Florence, the humanist of the 15th century was dedicated to the purification and preservation of Latin as a living universal language. He gave to the profession of writer and scholar a distinction it had not had since the collapse of Rome nearly a millennium earlier. Men of letters were held in an esteem that—in the case of that shown the Dutch humanist Desiderius Erasmus—approached reverence.

Regnancy of the visual arts. The new Renaissance attitudes were most striking in the visual arts. The painter and architect Giorgio Vasari, in his *Lives of the Most Eminent Architects, Painters, and Sculptors* (1550), saw the development of the arts in Tuscany as the working out of a divine plan. According to Vasari, God sent Michelangelo to the world endowed with so great a universality of power in each art that he might be considered of a divine rather than human nature. A century earlier, no one would have considered artists deserving of extended biographical treatment.

The artist had fought for this new status, however, from Filippo Brunelleschi, who stoutly defended his independence as an architect from the meddling of the Florentine government, to Michelangelo, in his defiance of Pope Julius II. Leonardo resented bitterly the placing of painting among "the mechanical arts" and attempted to raise the status of painting by deliberately downgrading sculpture, which he called a sweaty and fatiguing job for workmen. Even after he had become a world-famous artist, Michelangelo felt it necessary to defend his social and intellectual position. Forbidding his nephew to address letters to him as "Michelangelo the Sculptor," he insisted he had never been a painter or sculptor such as those for whom it was a business.

Status and the academy

The first academies of fine art were founded not primarily as teaching institutions but as means of enhancing the standing of artists. Florence's Accademia del Disegno, founded by Vasari in 1563, was headed jointly by a prince and an artist—the Grand Duke Cosimo and Michelangelo. In 1564 the academy publicly demonstrated the honour being paid to artists in its elaborate funeral ceremonies for Michelangelo. Such leading Venetian artists as the painters Titian and Tintoretto and the architect Andrea Palladio were glad to apply for membership in the Florentine Academy, and its advice on the design of the Escorial Palace was requested by King Philip II of Spain.

Particularly in northern Europe, however, the average painter, as distinct from the outstanding genius, continued to be regarded as a craftsman throughout the 16th century. In 1590 the Guild of St. Luke in Haarlem included not only painters, wood-carvers, and goldsmiths but also printers, slate layers, plumbers, and lantern makers. This extreme situation was bitterly complained of by a leading Dutch painter and theorist of the period, but it illuminates the attitude toward painting still held by many.

In the 17th century, the intellectual and social status of the artist reached perhaps the highest level it has ever attained. Two of the greatest Baroque artists, the Flemish painter Peter Paul Rubens and the Italian sculptor Gian Lorenzo Bernini, lived like princes. Both were accepted in leading intellectual circles and were completely at home in this environment. The Spanish master Diego Velázquez enjoyed lifelong security as the favourite painter and friend of King Philip IV, who, despite the grumbling of many aristocrats, awarded him the Noble Order of Santiago.

Writers and performers. The change in the status of the writer during the Renaissance and Baroque periods was not nearly as spectacular. The poet and rhetorician were treated with respect, but it was still necessary for Sir Philip Sidney about 1581 to write *An Apologie for Poetrie* (first published in 1595), and the poet Edmund Spenser, like many poets of lesser stature, had to beg for patronage from powerful nobles. Shakespeare published his narrative poems with obsequious dedications to his patron, the Earl of Southampton.

Practitioners of such newer literary forms as the novella and the popular drama never achieved full social or intellectual acceptance in this period. The dramatist, in particular, linked as he was with the day-to-day commerce of the theatre, and often himself an actor, was considered simply as a hack writer. His plays were regarded lightly as literature, and when Ben Jonson dared to publish his plays as his *Workes* in 1616, he was jeered at for his pretensions. Actors, though they were loved and admired as entertainers, had escaped only recently from being classed as "rogues and vagabonds," and technically they were regarded as household servants of the nobility. An occasional actor, such as Edward Alleyn or Shakespeare, was able to attain the position of a gentleman if he was a manager as well or had powerful patrons.

The theatre as mere entertainment

Only in the France of Louis XIV did the dramatist at-

tain the status of an honoured man of letters, and then only by adhering to the rigidly defined neoclassical standards of dramaturgy. The drama, like all the arts, was closely integrated into the political structure, and powerful ministers of the crown carefully supervised the work of the playwright. Leading dramatists such as Jean Racine were victims of political intrigues, but if they survived them, they could look forward to the ultimate reward of election to the Académie Française. Molière, however, who was an actor and theatre manager as well as a dramatist, was regarded socially as a member of Louis's household.

Revolutionary impetus. The great social and political changes of the late 18th century, climaxing in 1789 in the French Revolution, were as decisive for the artists as for other groups of men. The emphasis on personal freedom and the desire to break loose from mind-forged manacles as well as social shackles were key elements in the Romantic movement that dominated the arts for almost 100 years. A wide gulf separated the attitudes of Haydn from those of Mozart, though the two were only a generation apart in age. Haydn was content to serve as music master for the Esterházy family in Hungary from 1760 to 1790. Mozart could easily have made a similar career, but he rebelled against being a servant to the Archbishop of Salzburg. He determined to earn his living as an independent musician, and for a time his career flourished. The day of the free-lance composer had not yet dawned, however, and he died early, in poverty and loneliness.

For the first time in history, under the impetus of the French Revolution, there emerged the phenomenon of the revolutionary artist committed to an active political role. The most conspicuous example was the Neoclassical painter Jacques-Louis David, who not only expressed the ideology of the Revolution in his works and organized its great public festivals but was a political leader and a friend and ally of the Revolutionary leaders Marat and Robespierre. The ambivalent attitude expressed in the 20th century toward political activity by artists was illustrated earlier when David was caught up in Robespierre's downfall in 1794. Unlike other of Robespierre's followers who were executed summarily, David, after a stay in prison, was pardoned and restored to favour as virtual dictator of the arts. A similar situation befell the Spanish painter Francisco Goya, who had partially cooperated with the French puppet government in Spain. When the Napoleonic armies were driven out in 1814, he regained his position as court painter. There may have been in both cases an unwillingness to dispense with the services of a great artist. In addition, however, there also may have been at work the feeling, expressed even in this present century, that the artist is not to be taken altogether seriously as a politician.

The artist as political activist

The developing setting of modern art. A characteristic feature of artistic activity in the 19th and 20th centuries is that no generalizations about the status of the artist can be made. Trends are divergent and often completely contradictory. On the one hand, the artist became one of the representatives of the acquisitive society, a successful businessman satisfying the demands of the market. Whether he was a man of genius, such as the English novelist Charles Dickens and the French novelist Honoré de Balzac, or a talented and industrious hack, he wrote for a vastly expanded audience created by mass literacy and growing leisure. The English writer Anthony Trollope, sitting in his club and turning out his fixed quota of words, finishing a novel one day and beginning another the next, was a perfect embodiment of the respectable, punctual, middle class writer.

Similarly, the official artist was recognized by the national academy and hung at the salon exhibitions. He created a standardized and approved product—whether portrait, landscape, or sentimental illustration—to meet the great demands of a newly rich clientele eager for culture but unsure of its tastes. The American portraitist John Singer Sargent not only answered completely the artistic needs of his upper class patrons but identified with them as a person. At this point the social distinction between patron and artist virtually has disappeared.

On the other hand, the 19th century was above all the period of the alienated artist, the deliberate exile from society. The tone was set early in the century by the English poets Lord Byron and Shelley, both born aristocrats, who flouted the standards and conventions of their class. The contempt for accepted social behaviour and the adoption of the pose of the outcast are seen at midcentury in the French poet Baudelaire and in the closing decades in the French painter Gauguin. *Épater la bourgeoisie*, "to dumbfound the middle class," became the slogan of a whole group of French Romantics. The doom-ridden artist propelling himself toward destruction by drink, drugs, or a furious excess of behaviour became a characteristic figure of modern times, from Edgar Allan Poe to Jackson Pollock and many "pop" singers of the 1960s and 1970s.

Emergence of the alienated artist

Somewhat related to this phenomenon was the distrust with which the middle class came to view artists, above all painters and musicians. This is especially true of attitudes toward the avant-garde artist, whose rejection of academic conventions within his own art makes him suspect as an enemy of society itself. Napoleon III's superintendent of fine arts summed up for all time the attitude of officialdom toward such artists when he expressed his displeasure and disgust with their work, characterizing them as democrats who do not change their linen and hope to put themselves over on the world.

Impact of scientific dominance. As science became a dominant force, the intellectual prestige of the artist declined. The role of the prophet continued to be played by such writers as Thomas Carlyle and John Ruskin, both of whom were skeptical of the benefits of science and technology. No creative artist since Goethe, however, has gained the intellectual respect accorded to scientists. Modern art has produced no men of universal interests and abilities to match such Renaissance geniuses as Leon Battista Alberti or Leonardo, whose work and thought carried them into virtually every area of human activity.

The ideological gap between scientists and artists, between "the two cultures," as the English novelist C.P. Snow has called them, has widened dangerously over the past century. A number of artists, however, attempted to incorporate into their own work some of the methodology of science and to gain for themselves the intellectual status of the scientist. In a famous essay of 1880, *The Experimental Novel*, Émile Zola argued for a literature governed by science and claimed for the novelist the function of a biologist of society. The French Postimpressionist painter Georges Seurat, who set for himself the reconciling of art and science, believed that properly applied scientific theory could replace intuition as the basis of art. In the 20th century, apologists of Cubism, though not its leading practitioners, saw in it the artistic expression of the fragmented world of modern physics. Recently, electronic music, created mainly by university-based composers, has provided a new union of technology and art.

New areas of respectability. The actor-manager David Garrick had helped to make the theatrical profession respectable, but during the early 19th century, actors and actresses, however applauded or financially successful, continued to be regarded by many as slightly disreputable. By the end of the century, however, the situation had so changed that in 1895 Queen Victoria knighted Henry Irving, the first actor to be so honoured. Such an accolade became almost commonplace in following years, and in both Europe and America the actor is now accorded considerable respect.

The new arts of the 20th century—the cinema, radio, television, and recording—have all produced their mass idols. The fantastic adulation lavished on film stars or pop singers, however, has often been accompanied by an openly or covertly expressed intellectual contempt. Of all the artists involved in the creation of a film, only the director seems so far to have achieved intellectual respectability. Such directors as the Italian Federico Fellini, the Swede Ingmar Bergman, and the Frenchman Jean-Luc Godard became subjects of a serious aesthetic discussion comparable to that focussed on the most important of contemporary writers or visual artists.

Artists of the new arts

In music, the last 100 years have featured the virtuoso performer, the pianist, the violinist, the soprano, or the conductor. The separation of composer from performer began early in the 19th century, and instead of a Bach, Mozart, or Beethoven performing his own works, there is the virtuoso who interprets the compositions of others. The most publicized musician of the 20th century and the one most widely admired for his intellectual gifts was not a composer but the conductor Arturo Toscanini. The famous pianist Ignacy Paderewski, when he served briefly as premier of Poland in 1919, became the first professional artist to head a national government. One of the most striking phenomena of the rock music of the 1950s and 1960s was the closing again of the gap between composer-lyricist and performer.

Another important development in this century was the growing activity of black artists and women artists to break down the social and intellectual discrimination that historically has been directed against them. There has been, as a result, not only increased public knowledge of the contributions made by both groups but also a greater understanding of the artist as an individual who is directly involved in—and in much of his work reflects—the problems and stresses of his society.

THE ARTIST'S LIVELIHOOD

Throughout history, the economic status of the artist has varied as widely as his social and intellectual status, though changes in the one have not necessarily paralleled changes in the other. Since conditions favouring one art may have been unpropitious for another, artists in different fields often did not experience the same measure of economic security at any given time. Similarly, though artistic activity in general has followed economic development, there have been periods of great prosperity during which artists have starved.

Forms of patronage. Until recently, artists in most fields were directly dependent upon patronage—whether governmental, church, or private—for their livelihood. They produced works specifically commissioned for a particular need or occasion or designed to appeal to the tastes of a well-defined group or known individual. The notion of an artist working to please himself and then attempting to sell his product on the open market, either directly or through an intermediary, is a comparatively modern one. In such fields as architecture and serious music the artist even now remains almost completely dependent on direct commissions.

Quality and patronage

The system of patronage has had a profound effect on the work of art itself. The idea that institutional sponsorship always produces bad works of art is refuted by the facts. The buildings on the Acropolis of Athens, the Gothic cathedrals throughout Europe, the Sistine Chapel frescoes in Rome, and the sculptures of the Benin kingdom of West Africa, for example, were all commissioned by a governmental or religious body. Artists working to express a shared system of ideas and beliefs for a community of which they are an integral part are likely also to share a common and stable style. The rapid shifts of style that characterized the arts in modern times reflect, to a certain extent, the new economic status of the artist as well as the present instability of artistic traditions.

The open marketplace. Nevertheless, many artists of the past have chafed at the loss of individual freedom entailed by institutional support. This was particularly true during the 16th and 17th centuries, when the intellectual and social status of the artist was improving. When artists gradually freed themselves from the tyranny of patronage, they often discovered that they had submitted to the equally confining tyranny of the marketplace. In 17th-century Holland, for the first time in the history of art, painters began to produce works to be sold at auction or in art markets. The dependence on commissions did not cease, but direct patronage was no longer the only source of the artist's income. Most of Rembrandt's work, for example, was probably done to order, but many of his surviving works, including more than 60 self-portraits, probably were done without immediate commission.

In the 19th century, a number of artists attempted to capitalize directly on their own work, not by selling individual paintings to particular buyers but by exhibiting them for a fee to a mass audience. The American artist Rembrandt Peale earned $9,000 from the showing of his painting "The Court of Death," which was seen by 32,000 people during a 13-month tour in 1820–21. At virtually the same time, the French painter Théodore Géricault was exhibiting his famous, "The Raft of the Medusa" throughout England and Ireland. In 1855, Gustave Courbet constructed his own Pavillon du Réalisme at the Universal Exposition in Paris, charging the public an entrance fee to view 50 of his paintings.

In the other arts, as well, the artist emerges as an individual entrepreneur from the 16th century. Alexander Pope was perhaps the first poet in history to earn a living entirely from the public sale of his work. The new commercial orientation of the writer was summed up in the famous aphorism of Samuel Johnson, "No man but a blockhead ever wrote, except for money."

Institutional patronage has never, even in modern times, lost its importance for the artist. Eugène Delacroix, the supreme French Romantic painter, accepted a number of commissions for large murals from the French government. One of the most celebrated paintings of the 20th century, the "Guernica" by Pablo Picasso, was done for the Spanish Pavilion at the Paris Exposition of 1937. Both Marc Chagall and Henri Matisse created masterpieces of religious art on commission.

Artists in the modern period have had, nevertheless, to appeal to the general public in order to survive. For those who succeeded, the rewards were great, and many artists, both academic and avant-garde, became wealthy. On the other hand, innumerable others have been driven to despair by lack of public recognition and support.

Governmental and other subsidies. Various attempts have been made to provide some kind of governmental subsidy for artists. Among the most conspicuous of these were the Works Projects Administration (WPA) programs for painters, writers, actors, and other artists supported by the United States government during the depression of the 1930s. These projects were attacked bitterly by some politicians and eventually discontinued.

State-supported performing companies

Today, the principle of government support for the arts is widely accepted. Most European countries have state theatres and opera houses, and they support symphony orchestras, ballet companies, and other groups of artists. The British government grants sizable subsidies to the National Theatre and the Royal Shakespeare Company. In the United States, the federal government and some state governments have given financial aid to artists through such organizations as the National Endowment for the Arts and the New York State Council on the Arts. Universities and some private foundations have become important sources of economic support for the artist. Poets, painters, and composers in residence are now found in many universities, sometimes with teaching duties, sometimes free simply to create.

Conditions outside the West. The social forms of non-Western cultures have varied so widely, from place to place and from period to period, that it is impossible to make any valid generalizations about the economic position of the artist in such cultures. At one end of the scale, in simple nomadic or agricultural communities, the artist hardly has existed as a specialized person. Such crafts as pottery making or weaving generally have been diffused throughout the community, sometimes limited to one sex or another. In the sophisticated and highly developed imperial, monarchial, or tribal societies of China, India, or Africa, on the other hand, the trained artist has always been regarded as a valued servant of the ruling or priestly group and has been integrated completely into the economic structure of the state. With the vast social and political upheavals in the Orient and in Africa in recent decades, a period of rapid change in the economic and social status of the artist has set in.

INTERACTIONS AMONG ARTISTS AND WITH THEIR PUBLICS

The mingling of artists working in the same or different mediums always has provided a leavening element for

their creative energies and imagination. Similarly, collective or collaborative work, criticism and appreciation, and national and international awards, as well as collective action for the benefit of the artistic community, have contributed to artistic life and work.

The artist's working environment. The act of artistic creation is usually a solitary one, whether carried out in an isolated cell or in a crowded room. Most artists, however, have sought the stimulus of some special kind of environment, generally one in which they could have regular interaction with their peers. Only rarely has an artist who is cut off from others, by necessity or choice, been able to produce great work. It has been suggested that the American poet Emily Dickinson might have developed her poetic gifts even more significantly if she had not been so isolated from other writers. The battles of wit between Shakespeare and Ben Jonson at London's Mermaid Tavern may be apocryphal, but they describe a situation typical of the literary life.

A deliberately and formally organized artistic environment has proved only rarely to be conducive to creative work. Art colonies, either specifically limited to artists or existing within Utopian communities, have tended to fall apart rather quickly. One of the longest lasting of these, the MacDowell Colony at Peterborough, New Hampshire, has endured mainly because it offers a pleasant and quiet summer retreat for the artist. Brook Farm in Massachusetts, which Nathaniel Hawthorne joined briefly in 1841, and the Helicon Home Colony in New Jersey, founded by Upton Sinclair, were both short-lived.

The university has come to play an important role as a fostering environment for the artist. The many temporary or permanent positions it offers have provided not only a measure of economic security for artists but also an intellectually stimulating setting. There is some feeling, however, that permanent immersion in an academic environment may insulate the artist too much, eventually depriving him of his vital sources of inspiration.

Centres and schools of artistic activity

Historically, most artists seem to have thrived best in a city that is the centre of vigorous political, economic, and intellectual activity and in which they could interact with other artists when and where they chose. Florence in the 15th century, London in the Elizabethan era, Vienna in the late 18th and early 19th centuries, and Paris from the Revolution until World War II were such places. The Café Guerbois in Paris was the scene of innumerable heated discussions in the 1860s and the 1870s between many of the leading painters and writers of the period. Such interchanges undoubtedly provided much creative impetus for their art.

Sometimes, artists brought together by similar views have mutually inspired each other and worked in tandem to develop new styles. This was true probably of Giorgione and Titian in Venice, and certainly of Georges Braque and Picasso during the elaboration of Cubism in the early 20th century. There have also been looser groupings of artists with common interests, such as the Impressionist painters who exhibited together from 1874 to 1886. More recently, such groups as the Dadaists of immediately after World War I and the Surrealists of the 1920s for a time united various artists of sometimes diverse tendencies.

Artists have on occasion, particularly during the 19th century, combined out of a sense of shared poverty and neglect to express a common contempt for accepted values. They have formed a separate society of their own, a "bohemia" characterized by deliberately outrageous costume and behaviour. The ultimate expression of the bohemian disdain for middle class conventions was the glorification of suicide, and young French bohemians actually founded a Suicide Club in 1846 as a gesture of defiance.

Social integration outside the West

Such bohemian excesses generally have been unknown in non-Western cultures. In societies where the role of the artist has been clearly defined and established by long tradition, the notion of the alienated artist would be virtually incomprehensible. Like the European craftsman of the Middle Ages, the non-Western artist has functioned not in a special environment set apart from his society but as a respected member of the community. Here again, of course, as in other aspects of the profession of the arts, the very recent period has begun to see a breakdown in traditional values and relationships.

Multiple creators. Related to the question of the artistic environment is that of group creativity. A collective approach has proved far less viable in the arts than in such fields as scientific research. In the performing arts, film, and architecture, however, cooperative activity has been not only possible but often necessary, although usually there is one guiding vision—that of the director or the chief architect, for example. Many great buildings, such as St. Peter's Basilica in Rome, represent the effort of a number of different architects. Almost always, however, these architects have been in charge successively rather than at the same time.

Collaboration

More successful in most arts has been the collaboration of two individuals. The Parthenon was the product of the architects Ictinus and Callicrates. Collaboration was a common procedure among Elizabethan dramatists, a most notable example being the works of Francis Beaumont and John Fletcher. A collaboration between composer and librettist, such as that between Mozart and Lorenzo da Ponte, created many great operas.

Art collectives have been organized, with varying degrees of success, in many of the Socialist or Communist states, particularly in the immediate post-Revolutionary periods. In such countries as China, North Vietnam, and Cuba, deliberate emphasis has been placed on "the democratization of art," an attack on the notion of art as the province of the individual genius, and an attempt to create a collective and anonymous art. Numerous political events, such as the 1968 student strike in France, stimulated collective activity in the production of propaganda posters and the organization of guerrilla-theatre companies that gave informal but impassioned performances usually on social or political themes.

Guilds and unions. Artists have also banded together for very practical purposes. During the later Middle Ages and the Renaissance, European artists, like all other craftsmen, were organized into guilds that looked after their economic interests and regulated trade procedures. The bronze workers of the great African kingdom of Benin, which flourished in the 16th century, also had their guild.

Attempts to organize creative artists into modern trade unions generally have failed. During the depression in the United States, artists' and writers' unions enjoyed a brief period of growth, and during the 1960s black artists and women artists formed their own groups. Performing artists, in contrast to creative artists, have formed strong unions that have been able to exert considerable pressure to better their economic situation. American Actors' Equity Association and British Actors' Equity Association are probably the best known of these unions, and there are similar organizations in radio and television and in the film industry. On an international scale, there have been various attempts to bring together artists for political or economic activity. International PEN (*P*oets and playwrights, *E*ssayists and editors, and *N*ovelists) is an organization of writers that has taken vigorous public positions on matters of concern to professional men and women of letters. Founded in 1921 and actively supported in its early days by such prominent figures as H.G. Wells, Bernard Shaw, Anatole France, and Thomas Mann, it includes today more than 8,000 writers from 58 countries.

Impacts of criticism and appreciation. Among the necessary conditions of work for most artists are public criticism and appreciation. Although artists often have pretended to be scornful of critics, generally they have flourished best in an atmosphere in which their work was understood and encouraged. Almost every important artistic movement of modern times has had its critical spokesman and defender. Conspicuous among such champions have been the Frenchmen Zola, for the Impressionists, and Guillaume Apollinaire, for the Cubists, as well as the American critic John Martin, for the modern dance movement.

Many creative artists have been discouraged, however, by lack of appreciation of their work. The great Baroque architect Francesco Borromini committed suicide at least partly because of critical neglect and attack. A legend that the English poet John Keats died because of savage criticism of *Endymion* is false, but Keats was affected seriously by the attacks. Cézanne, in his later years, was deeply hurt by what he thought was complete public disregard of his achievement.

National and international honours

Like individuals of distinction in other fields, artists have been singled out for awards of various sorts. The most prestigious of these honours is the Nobel Prize for Literature, which has been awarded annually since 1901. The prize has sometimes been refused, notably because of political pressure, but the writers chosen usually have regarded it as the climax of their careers. There is no comparable award in the other arts.

In the United States, the Pulitzer Prizes annually honour outstanding work in drama, fiction, poetry, and musical composition, as well as in scholarly and journalistic writings. The National Book Awards also honour literary work in the different genres.

In other countries, artists are given prizes or other marks of merit for distinguished achievement. The Prix Goncourt is the best known of French literary awards. Many creative and performing artists appear each year on the honours list in Great Britain, as recipients of knighthoods or other distinctions. In the Soviet Union, the title of Honoured Artist is conferred as a badge of accomplishment, and in Japan, actors are often honoured with such titles as National Living Treasure, indicating the prestige they have acquired.

Artists are also eligible for election to honorary societies, some of which are limited to the arts. The greatest tribute that can be paid to an artist or intellectual in France is elevation to membership in the Académie Française, the famous company of "immortals." In the United States, the National Institute of Arts and Letters constitutes the chief honorary society of the arts.

THE ROLES OF THE AMATEUR

Amateurism has always been an important factor in the development of the arts. In the oldest sense of the term, the "amateur of art" is simply the lover of art. He need not necessarily be a practitioner of any of the arts himself, functioning instead to encourage and support the arts in every way possible. What distinguishes this kind of amateurism from passive art appreciation is its active commitment. From its ranks come the students, the connoisseurs, and the patrons of art.

There is also a more common sense in which the term amateurism applies to the arts. There are vast numbers of people who, without special talent or extensive training, write, paint, or play musical instruments primarily for their own pleasure or as a leisure occupation. Often such amateurs band together in musical groups or theatrical companies that perform for themselves and their friends. The line of demarcation between amateurs of this kind and professionals is usually clear, although their economic contributions in the form of royalties are often important for playwrights. Few Sunday painters, unlike Gauguin, ever abandon their businesses and families to risk everything on their talents.

Another kind of amateurism, rare today, though historically it has been of great significance, goes back to an idea that appeared first in antiquity and acquired force during the Renaissance, namely that proficiency in an art, particularly music or poetry, is an attribute of the gentleman. Art is valued as an accomplishment of the educated man, but not as a profession.

The idea of the gentleman artist

Unlike the amateur who practices an art simply for his own enjoyment and basically is not concerned with how well or how badly he does it, the gentleman amateur aims at excellence. His ideal is the quality of *sprezzatura*, or effortless grace, described by Baldassare Castiglione in his *Courtier* (1528) as one of the chief characteristics of the Renaissance courtier.

Chinese culture always has been unique in considering painting as the special avocation of the noble amateur. The man of letters, the statesman, and the upper class gentleman have traditionally been expected to be proficient in painting, and especially in calligraphy. In India, too, training in the arts has been considered a necessary part of the education of members of the highest caste.

The amateur in all of these senses has made invaluable contributions to the preservation and dissemination of art. The devotee of art has encouraged and supported artists, has collected and handed on works of art, and has founded museums, libraries, and schools for the arts. Above all, he has set a standard of taste in every generation, sometimes a false or artificial one but one that often has raised the level of artistic performance. If the skilled amateur at times has introduced a measure of snobbism into the appreciation of the arts, he has compensated for this by his defense of art against those who have attacked it as immoral, dangerous, or frivolous. Most of all, however, artists are indebted to the millions of amateur practitioners of art who, having experienced its value for themselves, can respond to it as practiced on the highest professional level.

BIBLIOGRAPHY. There is no one book that satisfactorily covers all of the issues discussed in this article. There are, however, a number of excellent works that deal in detail with one or another phase of the subject. CARL ROEBUCK (ed.), *The Muses at Work: Arts, Crafts, and Professions in Ancient Greece and Rome* (1969), contains a mass of fascinating information. RUDOLF and MARGOT WITTKOWER, *Born Under Saturn: The Character and Conduct of Artists: A Documented History from Antiquity to the French Revolution* (1963), is an invaluable work. NIKOLAUS PEVSNER, *Academies of Art: Past and Present* (1940), is not only the standard book on its subject but also contains many penetrating observations on the profession of the arts generally. QUENTIN BELL, *The Schools of Design* (1963), adds much interesting and useful detail to Pevsner's book and explores some new ground. CESAR GRANA, *Modernity and its Discontents: French Society and the French Man of Letters in the Nineteenth Century* (1967), is a brilliant study that was originally published in 1964 as *Bohemian Versus Bourgeois*. Much information about the profession of the arts may also be found in books on the history of the various arts, such as the volumes in the "Pelican History of Art" series; or JOHN REWALD'S magnificent two books on *The History of Impressionism*, rev. ed. (1961), and *Post-Impressionism* (1962). DONALD DREW EGBERT, *Social Radicalism and the Arts: Western Europe* (1970), provides a very full treatment of the effect of radical thought on the arts from the French Revolution to the present.

(Si.T.)

Arts, Social and Economic Aspects of the

It is necessary to begin a sociological analysis of the arts by identifying the various social frameworks within which artistic activities have been conducted and the influences that these frameworks have had on the style and content of the arts, the levels of creative attainment, the mode of living of the artists, and the uses to which their art has been put by society.

This mode of analysis is not concerned, as the histories of the various arts are, with describing how the particular arts have historically evolved and what they have meant to their users. Rather, it is aimed at discerning the basic alternative patterns of organizing artistic activities and the consequences, for society and for the arts, of adopting one or another of them.

Most of the necessary knowledge for recognizing these patterns is still lacking or is ambiguous in its implications. Indeed, there is no generally accepted theoretical basis for encompassing all the arts in relation to all sociological variables in all types of societies, from the simplest to the most complex. There is, furthermore, hardly any other field in the whole area between the humanities and the social sciences as inviting to partisan sensibilities as the relationship between art and society. Any general statements about relationships between art and society must therefore be treated cautiously, not as established knowledge but as tentative hypotheses.

This article is divided into the following sections:

I. General considerations
The field of art
The aesthetic function

Social uses of art
The cognitive character of art
Social dynamics of artistic creativity
II. Artists and artistic cultures
Social role of the artist
Artistic cultures
III. Economic support of the arts
Economic evaluation of the arts
Systems of financing artistic activities
The art market
Remuneration of artists and protection of their rights
Art collecting
Fraudulence in the arts
IV. Social control of art
Types of regulation
Conditions for social control
Implications of social control
V. The arts and religion
Social relationships
Aesthetic influences
VI. Technology, science, and the arts
Influence of technology on art
Other aspects of the relationship
VII. Aesthetic education
Basic conceptions
Supplementary approaches
VIII. Preservation and dissemination of art
The nature of art preservation
Systems of dissemination

I. General considerations

THE FIELD OF ART

There are two possible ways to conceive of art sociologically. In the nonhistorical view, art must be conceived as that which is defined by a society, or an artistically relevant part thereof, as art. This could be called the "labelling" view. The work itself may or may not claim to be art; it is recognized as such by those with an authority to do so (in modern societies, by artists and art critics) or by anyone interested. This definition implies that art did not exist in preliterate societies until it was recognized by the moderns, since what now appears to be art has been treated mostly as religious or utilitarian artifacts in such societies. In the absence of the label art, imposed by later cultures, these objects are indeed only utilitarian or religious artifacts. Art arises merely in our perception of them, but does not exist in the intrinsic qualities of the objects themselves.

Historical conception of art

The second, historical view of art is based on the assumption that art is what survives a series of "tests" given to objects that function as art. When such an object is initially presented to the view of people other than its creator, it could be viewed as representing a "claim to art." When it is accepted by large numbers of people in a society or by its established elites or by other artists and art critics, it could be said to have become a popularly, or authoritatively, or professionally validated work of art. But the ultimate test of its artistic quality is whether it can transcend the boundaries of time and space and be accepted by other peoples in other areas.

The historical conception of art implies that art represents a universe with *some* shared characteristics that are everywhere recognizable as artistic (whether or not the concept of "art" is consciously acknowledged). Experimental psychological studies provide some support for this view. It has been found, for example, that traditional Japanese potters agree to a high degree with U.S. art students on the relative merits of a series of works of art shown to them. The agreement is closer than that between American art and nonart students. Thus, practitioners of art appear to agree across cultural boundaries on the quality of artistic attainments.

The agreement on what constitutes a work of art has the following characteristics:

1. It is partial. Each society and period has its own particular standards, as well as generally accepted criteria, by which it judges works of art. The patrons and publics of art (and even art critics) probably insist on these special standards to a greater extent than do the artists themselves. Hence, local criteria should have greater weight in judging works of art when the artistic enterprise is dominated by nonartists. This is one reason for the great fluctuations over time in the economic evaluations of particular works of art, for persons who are not artists usually determine these judgments.

2. The artistic consensus is hierarchical. When artists are left alone, they can, in the long run, roughly agree on the ranking of individual works of art. This happens even in quite egalitarian societies, such as the Australian Aborigines, where most men participate in artistic activities but differences in the quality of achievements and in individual capacities are recognized. The notion of a hierarchy of artistic qualities is therefore not a superimposition of a social hierarchy on artistic experiences.

3. The consensus on art has expandable boundaries. The artists of a society can incorporate works they have been unfamiliar with into their notion of the "field" of art, with a discriminating sensitivity as to their merits. The entrance of new claims to art generates efforts at evaluation, clashes of particularistic values, and the rise of artistic schools and movements. Such claims are ultimately tested in terms of what is considered to be the total structure of the field of art. In turn this structure evolves, by a process of self-testing, through the evaluations it gives to works claiming the right to enter it.

THE AESTHETIC FUNCTION

The nature of the system of art seems to derive from sustained experience with the practical problems of making objects or acts that perform the functions of art and from a sense, which craftsmanship seems to generate, for what transcends mere craftsmanship. A mid-20th-century survey in the United States has shown that craftsmanship by itself is more highly regarded, in judging works of art, by nonexperts than by artists and art scholars.

Art and craftsmanship

Since craftsmanship is a purposeful activity, it appears that the art public is more apt to judge art by some presumed purpose, while artists judge it in terms of what transcends any presumed purpose. If it is ultimately the consensus of artists that determines what is included in the field of art, artistic value must lie in something that is not recognized, even by the artists themselves, as "the purpose" of art.

It follows that works of art cannot be understood by the manifest functions they have been specifically intended to perform. Where they function most purely as works of art, they perform latent functions—unintended and unrecognized. If this observation is valid, art could be regarded as the generalized system of the society for the performance of unintended and unrecognized (but nevertheless needed) psychological and cultural functions. If these functions are effectively performed by other systems of the society, art can remain implicit in them. To the extent that other systems become explicit about the functions they perform and rationalize them to exclude everything that does not clearly contribute to their purpose, art has to emerge as an autonomous system.

The autonomy of art from the social forces that would control it derives, as an immediate consequence, from the structural requirements of good design peculiar only to art; and, ultimately, from the origin of the aesthetic function in the intuitive organization of unintended and unrecognized functions of an interrelated psychological and cultural nature. To the extent that art performs an aesthetic function, it is not subject to intentional social control. A variety of consequences follow from this conception of the aesthetic function:

1. If its very essence arises from performing unrecognized functions, art must be a less self-conscious, a less "rationalized," and indeed a less professionalized activity than any other in the cultural sphere.

2. Since it must be ready to perform unrecognized functions as they unpredictably arise, the system of the arts cannot be a specialized one, adapted to a particular set of circumstances. It must remain generalized, to some extent maladapted to the existing state of society, and able to function in a wide range of areas of ambiguity.

3. The survival over time, and perhaps the aesthetic quality, of works of art depends on how wide a range of unintended and unrecognized functions they can effectively perform. It is because they have a wide aesthet-

ic range, in this sense, that great works of art function for us even when it is not known exactly what they have meant for their producers, as is the case with prehistoric art. The latent functioning of a work of art is not dependent on the grasp of its intended meanings.

4. It could be argued that an effective organization of any unrecognized function constitutes the aesthetic aspect of the social or psychological system in which it is embedded. Thus, scholars could analyze successful works of art as diagrams of effectively constructed but hidden psychological processes occurring in personalities individually or collectively, and could discern how these diagrams are used, by artists and art consumers, to deal with otherwise raw and chaotic psychological processes. The diagrams may be seen as providing rehearsals for effective organization of these psychological processes or as suggesting alternative models for them or as providing focal points around which they can crystallize.

If aesthetic value depends on consciously unrecognized functions, does an explication of these functions erase the aesthetic experience (or prevent it from arising for those who will, in the future, be conscious of the functions the work of art presumably performs)? Not necessarily—if the work of art, after one of its functions has been explicated, can still function effectively in other unrecognized and unintended ways. The interpretation of art could be viewed as a struggle against its inexhaustibility, but the functions that have been fully explicated would seem to become more cognitive than aesthetic.

SOCIAL USES OF ART

While, to be artistically effective, art must function as art—fulfilling a great variety of unrecognized and unintended functions—it does not generally operate as "pure" art. First, it may overlap with other cultural systems—religion, philosophy, science, a secular ideology—and perform, in part, the more clearly identifiable functions of these systems. It is then shaped, to some degree, by the superimposed functions it performs as a part of those systems.

Relation of art to other disciplines

In general, art tends to become increasingly differentiated from other cultural systems in the course of social evolution. Yet in some periods a closer integration of art and some other cultural system may be sought. Thus, the system of science has, in most cases indirectly, affected much of modern visual art. It cannot be taken for granted that the most complete differentiation of art from other cultural systems is most conducive to its authenticity. Nor does a self-conscious "integration" of art with another cultural system enhance art, as is shown by the failure of Socialist Realism, tightly bound to Communist ideology. It could be inferred from the notion of the aesthetic function that an indirect, unintended mutual interaction between art and some other cultural system would be most stimulating to the arts.

Second, art may be used by various social agencies or groups to perform functions these agencies are interested in. Such use of art by social agencies structures the content and style of art and influences its level of creative attainment and its total repertoire of functions. Again, it is not to be taken for granted that it is necessarily disastrous for art to be used for extra-artistic purposes. On the one hand, by using art to fulfill their purposes, social agencies may restrict art's capacity to function as art. This seems likely to happen to the extent to which social agencies successfully limit art to performing any set of consciously recognized and intended purposes.

On the other hand, by using art for their own purposes, social agencies may also stretch art's limits in directions that artists might not otherwise have been inclined to explore, enhance certain of its expressive potentialities at the cost of others, and increase its capacity to communicate with contemporaries while perhaps reducing its ability to communicate transhistorically and cross-culturally. By being forced to struggle against purposes imposed from the outside, artists may become more aware of what is both peculiar and essential to art.

Like other symbolic systems, art can function as a means of attaining the purposes of any system of society. Thus in the economy, art can be used as a means for attaining or symbolizing the possession of wealth; but also as a critique of particular ways of using it. In the political system, it can encourage or condemn a particular distribution or use of power; and, in the community, as a means of reinforcing or protesting against the existing order of sensibilities, expectations, social rankings, and social distances. In an ideological system, it can operate to strengthen the hold of established values by filling the imagination with forms or content suggestive of these values—or to question them by presenting forms and content that are irreconcilable with existing values. But even if art "objectively" functions to promote particular social ends, it is not necessarily consciously employed to promote these ends.

Extra-artistic functions

None of these extra-artistic uses of art seems by necessity aesthetically superior to any other. What seems important is whether artists accept the legitimacy of the extra-artistic expectations directed to their work and incorporate them, unselfconsciously, into their own notion of the artistic task. If they do so, they should be able to produce good art regardless of the type of extra-artistic expectations imposed on it.

The critical question, nevertheless, in evaluating the aesthetic quality of works of art is whether they can resist or transcend the social uses to which they have been put by their makers and whether they retain their own character and transmit their own message even when manipulated to serve the purposes of their sponsors. An art that has not been subjected to manipulation during its making could well be deprived of a powerful stimulus for acquiring a "resistable" character, an artistic toughness that ensures its aesthetic survival.

THE COGNITIVE CHARACTER OF ART

To what extent do the particular social uses of art shape its content and style? To what extent do content and style, therefore, correctly reflect or distort the "objective" realities of the society in which they have been produced? Does use of art by a social agency or group necessarily imply a distortion, in its interests, of the reality that art may be presumed to reflect? Or can some groups (as, in the Marxist views, the "progressive" ones, which identify themselves with the direction of history) use art in their own interests without thereby forcing it to distort reality?

The most general response to these queries is probably that art reflects either subjective affirmations or subjective denials, symbolic invalidations, of the existing reality. It can therefore be read only as a record of the history of subjective attitudes toward objectively existing reality. The ways in which art has been used can be assumed to influence the subjective responses it will express. The subjective responses that exist in the environment in which art is produced but that are not "useful" to the groups or individuals that provide resources for art creation, or for the artists themselves, are less likely to be reflected in art. Yet, insofar as art necessarily performs unintended and unrecognized functions as well, it may reflect even the subjective responses that it is not useful to anyone involved in the artistic process to reflect. These responses may contradict the conscious intentions of the artist. Art is never an objective record, and it is never fully controlled by those who use it—or it ceases to be art.

Inherent subjectivity of art

The most significant art may well express both the most striking characteristics of objective social reality and the sense of what is most missed in it—reflections of reality as well as utopian denials of it. The significance of such art may arise from its discovery of ways to articulate these mutually contradictory subjective responses to social reality, without suppressing one in favour of the other. Cognitive distortions in art arise not from an introduction of subjective attitudes but from a denial of the ineradicable contradiction between existing "objective" realities and their possible "subjective" negations. Art distorts the totality of human experience, in any social setting, when it is biassed in favour of either "objective" recording or "subjective" expression.

Art reflects, or compensates for the deficiencies of, ob-

jective reality not only in its content but also in style. Even completely nonrepresentational arts therefore have a cognitive character, and subjective orientations to social reality can be inferred from them.

Easily swayed modes of art

Some arts lend themselves more easily to deliberate manipulation in the interest of consciously distorting the ways in which they reflect reality—whether in style or in content. Fiction, the theatre, painting, and sculpture are more vulnerable to such manipulation than music or lyrical poetry. Representational styles lend themselves more easily to manipulation than highly symbolic or completely nonrepresentational ones. Indeed, one reason for moving toward the latter styles, in modern societies, is the desire of artists to escape manipulation. The mass arts are particularly susceptible to imposed distortions. It is deliberate manipulation, rather than merely the use of art by social agencies and groups, that would appear to give rise to distortions in the way works of art reflect reality in its interrelated social and emotional aspects.

SOCIAL DYNAMICS OF ARTISTIC CREATIVITY

If art performs unintended and unrecognized functions, the creation of artistic values should be affected by the degree to which art is needed, in particular social settings or by particular individuals, to perform these functions. That is, for artistic creativity to become possible, there must be many psychic needs that are neither met by existing social arrangements nor can be consciously identified and purposefully dealt with by means of social policy.

Conditions favourable to artistic creativity arise when rapid changes in either the organization of society or in the emotions of its members produce a sharply sensed disjunction between personal emotion and objective social structure. But while the need for art increases during such periods, the possibility of creating it also depends on the availability of resources for creating art. The supply of such resources tends to be diminished in the phase of most intense action of periods of radical change. This phase occurs during rapid economic accumulation, technological transformation, the high points of religious (or secular ideological) reformations, and struggles for imperial consolidation, national liberation, or change of political system. Artistic creativity is enhanced when an increased social need for art coincides with an ample supply of resources for artistic production—thus before and after, but not during, the most intense phase of any cycle of technological, political, ideological, or communal change in the society. Activistic social movements are unlikely to be artistically creative, but they increase the need for art—after they have succeeded or failed to transform society.

Creativity and social resources

Conscious social policy can affect artistic creativity by supplying, or failing to supply, social resources for artistic production commensurate with the existing social need for art. Not all such resources, however, can be controlled at will—the supply of cultural symbolism, for example, depends on its existence in any social system in a form appealing to the imagination of artists. It does not depend on the immediate policies of governments, churches, or parties, although these institutions can, to a high degree, determine the allocation of economic resources to artistic activities. But while a deficiency of such resources may prevent the creation of works of art for which the potential exists in a society, economic resources cannot generate artistic values if the social need for art is not sufficiently intense, as during periods of general cultural quiescence and social complacency, when existing conditions are taken for granted by most members of a society (as in 18th-century Italy); if the creative people are too widely scattered in a large population or too isolated from each other by ethnic, ideological, class, or disciplinary barriers to attain a necessary density of interaction; if an appropriate social organization for producing art, particularly important for the large-scale arts and the film, is not established; or if the general symbolic design of a civilization is not sufficiently developed or is already too "completed" to sustain efforts at producing great artistic designs.

In modern societies, perhaps especially in the large nations, two threats to artistic creativity have emerged in the possible bureaucratic overorganization of the artistic enterprise, leaving too little space for the unintended and unrecognized in art, and the early popularization by the mass media of new artistic movements before they have had time to mature their contributions. Once artistic movements become widely popular, they tend to drop what they have been doing, since the modern cult of originality discourages a continued exploration of what others have become familiar with. Hence, what a style is potentially capable of may never be developed.

II. Artists and artistic cultures

SOCIAL ROLE OF THE ARTIST

The two main variables in defining the social role of the artist are that role's degree of specialization and the extent to which it is conceived as involving manual labour or some sort of "spiritual expression."

Variables in defining the artist's role

By their nature, all the literary arts involve less visible manual labour than the traditional visual arts, and the sculptor or painter has been generally more easily and closely assimilated to the traditions of skilled manual craftsmanship. The poet has tended to be associated with the realm of religious ceremony and, later on, with record keeping within or outside of the various church organizations. Thus, some of the prestige of literacy in traditional civilizations has attached to the poet's role. Yet it was originally literacy in the service of a god as well as an illiterate aristocracy, and there is still some tendency to expect writers, more than other artists, both to entertain the classes that have become their sponsors in place of the aristocracy and to serve a moral purpose. It is they, of all the artists, who are expected to be "the moral conscience" of society. But besides this clerical tradition, there is also the more anarchic one of vagabond poets, who live "beyond good and evil."

In music, he who composes a work is frequently—at least since the German composer Beethoven—perceived as partaking of a more "spiritual" role, and he who performs it, particularly in groups of performers, has been frequently seen as more of a "craftsman." In the theatre, the actor is a physical labourer insofar as he uses his own body, but, in performing his role, his body assumes and acts out "spiritual substances" originally alien, and possibly greatly superior, to it. This is one reason for the ambiguities in the social treatment of the actor, who, like the singer and the dancer, has been the most exalted and the most despised of artists.

Within the limits shaped by the nature of the particular artistic medium, the social role of artists has been affected by developments in the organization of society and its value system. The main evolutionary trend has been toward increasing professionalization of artistic roles. In the preliterate societies of a more egalitarian character, all adult members of a society (or all members of one sex, and a few of the other) may be engaged in activities of producing objects or performances that, in addition to their consciously intended functions, have some kind of aesthetic aspect (superfluous from a purely utilitarian point of view). But some individuals are recognized by their peers as more competent carvers, potters, or dancers, and their products or performances, though not superior in a utilitarian way, are more highly esteemed; their authors receive a superior compensation in prestige and, less frequently, in material valuables. Yet even the best of tribal artists are spare-time specialists who devote most of their time to meeting obligations incumbent on all members of their society of their own sex and age.

Full-time specialization

It is only with the development of a hierarchic type of social organization that, for male artists, full-time specialization becomes possible, if their society is both interested in art and sufficiently prosperous to sustain a demand for full-time specialists. Women tend to remain semispecialists or "folk" artists, and, the more professionalized the artistic enterprise, the lesser their part in it. In the primitive states or hierarchically organized chiefdoms is formed the main pattern that has governed artistic enterprise in classical civilizations: the division

between "high" and "low" art, the specialization of male artists, formalization of their training, patronage by clients, and general subordination of art to the politically ruling class and the religious establishments.

This pattern has occasionally been modified in mature preindustrial civilizations. The basic hierarchic pattern, however, has been radically challenged only in consequence of the Industrial Revolution and the rise of a civilization associated with it, in which aspirations to universal participation appear. This transition is still going on. But "high" art has already become detached from its almost exclusive dependence on the ruling elite. It is now visible, through the mass media, to almost everyone and is intensely experienced by a recently developed loose coalition of artists and intellectuals. This coalition is the primary audience, even for art that has been explicitly produced for other social classes. An example is the Mexican murals of the 1920s that were intended to revolutionize the proletariat but are appreciated mainly by a bourgeoisified intelligentsia. With respect to art, the socially dominant class tends to lose confidence in its own taste and to become "culturally" subordinate to this coalition. From a symbol of privilege, art changes into a symbol of emotional aliveness; hence, the boundary between "high" and "low" art becomes blurred. But the artistic enterprise itself remains highly stratified in accordance with the prestige granted the type of activity the artist engages in and his personal achievement in it. Women once again return to greater participation in the fully specialized production of the "creative" arts. As for the performing arts, women have either always possessed or regained at an earlier time fully specialized roles in them. This is especially true of the dance, except in those cases where religious asceticism or some comparable influence has eliminated them from the performing arts as well.

"Cult of the genius"

Historically, the most significant single illustration of the ways in which the value system of a society affects artistic roles is the Western "cult of the genius." It developed under the influence of the great achievements and the striking personalities of artists like Michelangelo and Leonardo da Vinci in the Renaissance—a civilization that generally placed a very high value on achievement and expression of individual personality. Comparably great achievements in less "individualistic" cultures have not produced any similarly powerful notions of genius. The cult of the genius had the long-term effects of assimilating all the creative, and at a later date even the performing, arts to the "spiritual expression" of poetry; legitimating the demand for creative autonomy on the part of the men whose genius has been perceived as transcending the dimensions of established custom; and, on the basis of the above two changes, transforming those arts previously treated as crafts (and organized in guilds, etc.) into a peculiar kind of free profession.

All professions are corporate bodies of practitioners in a skilled activity who are devoted to an ethical purpose superior to the mere pursuit of wealth, power, or pleasure for themselves. The professions set their own standards, determine admissions to them, and judge their own members accused of transgressions against the profession's ethical code. Modern artists constitute a peculiar profession in that they do all this without necessarily being organized into an association that encompasses all of its members (as the medieval guilds did) and without necessarily being held accountable to public authorities for its activities. The ethos of the artistic profession, as it has emerged in the West since the Renaissance, is opposed to both general organization and any kind of public accountability of artists. Even with the evaporation of the cult of genius, in the 20th century, the conception of the artist's role as that of a free professional has remained as the ideal expectation. There are evident tendencies, from the latter part of the 19th century on, toward a partial collectivization of the role of the professional artist—whether in artistic movements in which several artists interact in producing individual innovations that have something in common (as did the French Impressionist painters of the 1870s) or in artistic workshops in which several artists cooperate in producing the same, or the same type of, works of art (as in the Bauhaus school of design in Germany, 1919–33). Yet it is a collectivization of a free profession, not a return to the guild structures, that *some* artists of industrial societies are willing voluntarily to accept.

Even as a liberal profession, the artistic enterprise remains professionalized to a lesser extent than other recognized professions. The artists are *morally* a profession, but the degree to which they *economically* depend on it as their main source of income varies tremendously. Most professional artists, in modern societies, do not derive the bulk of their income from the direct exercise of their profession, particularly where they have to sell their products under conditions of a free market.

Professional incomes

Estimates made at the beginning of the 1970s suggest that no more than one-half of 1 percent of the professional painters and sculptors in Paris and in West Germany receive a regular and sufficient income from the sale of their works. Some arts have retained, even in the self-perceptions of the artists, the character of vocations rather than professions: it is still awkward to describe someone as a professional poet.

It is conceivable that, in the more affluent and leisurely societies predicted for the future on the basis of technological progress, well-trained amateurs might again, as in the past in China, begin producing art of respectable quality. The nature of the aesthetic function sets limits on the extent to which a professionalization of the artistic enterprise can be sustained without aesthetic loss.

ARTISTIC CULTURES

Artistic cultures are the various basic situations in which art is produced. Each such culture involves a distinct type of social organization of artistic activities that is associated with a distinguishable attitude of artists toward their work. Artistic cultures arise from artists' relationships with other artists, their publics, their means of earning a living, and agencies affecting artistic activity; from their involvements with cultural systems not specifically artistic; and from artists' technologies, shared emotions, and ideologies. Changes in any of these variables modify artistic cultures and give rise to new variants of them.

Once the artistically undifferentiated unity of the tribal society has come to an end, several variants of "traditional" and "modern" types of artistic cultures can be distinguished.

Traditional artistic cultures. At least seven highly distinct types of artistic cultures can be identified in traditional (preindustrial) societies.

The folk culture. Artists are nonspecialized members of their community of residence and closely involved with all of its activities. Their art deals with typical experiences, concerns, or needs of ordinary members of that community. The real-life references are, however, linked with elements of highly utopian imagination (folktales, in which an attitude critical of the existing reality is frequently expressed), abstract stylization (geometric ornamentation of utensils), or symbolizing lyricism (folk songs). The artist is one of the people, but living more powerfully in the imagination than the others. Artists create by following or elaborating on traditional patterns. Much art is produced spontaneously for oneself or in friendship or for communal enjoyment, rather than for pay.

The artisan culture. The artist is a specialist—possibly a member of a specialized collectivity such as a guild or a workshop set up by a state or a church—who works on order and for pay (or on command) only. He develops the pride of good craftsmanship and habits of regularity and reliability in his work. He does whatever style or content is required by his client without asking about his intentions. Whatever individual needs he expresses in his work, his conscious ideology is that he is a skilled worker who respects high standards of craftsmanship and produces to earn a living. He belongs, not permanently to his community, but temporarily to his client, wherever the latter may be located, and perhaps to his clan or guild of fellow artisans who share his craft ethic. To the extent

Specialized craftsmen

that his client is his local community, however, he becomes a "civic" artist, whose highest purpose is, on command, to celebrate his community, as in the cities of the Mediterranean civilizations. (In modern society, this culture survives most nearly, perhaps, among orchestral musicians.)

The clerical culture. The artist is very much a craftsman, but his craftsmanship is subordinated to a highly valued tradition of a literate civilization, which he is committed to serve and to defend by his work. By his association with this tradition, he gains in prestige, but he also acquires a moral responsibility which the pure craftsman is not bound by. He belongs to a moral community, not necessarily localized, that provides him with criteria for judging which works of art are worth making. In this can be discerned the potential beginnings of an artistic consciousness and of a critical attitude of the artist toward society and its artistic ideologies. Most of the time, the clerical artist (for example, the medieval manuscript illuminator) is submissive to the discipline of the moral community of which he is a part, but, in the conception of the artist as a civilized servant of a higher moral purpose, there is, at least, a potentiality of criticism of the organization to which he belongs.

The ecstatic culture. This is a general term suggested for situations in which some type of artistic creation—though it is usually more than merely artistic—occurs in the midst of a mediumistic trance performed for religious or magical purposes, during an orgiastic happening, as a consequence of a mystical experience, or even as an element of prophecy. Art, performing or literary, arises from emotionally intense experiences, "divine madnesses," which are perceived to have high symbolic significance but over which the individual (or, sometimes, the group) that has them imposes no critical controls. Traditions that generate ecstatic art creation, such as the Dionysian ritual (performed by followers of the cult of the god Dionysus) in ancient Greece or the poetry of Hebrew prophecy, usually arise in loosely structured societies (as do the mediumistic priest-doctors known as shamans) or in times of social crisis (as mystics and prophets are likely to do). The shaman tends to be conservative with respect to his society, the mystic apathetic, the prophet radically innovative. Numerically, these types constitute a small part of the artistic cultures of the well-ordered literate civilizations and are least important in the visual arts dependent on organized patronage. But they have been sources of revitalizing impulses for established traditions. In the ecstatic culture, the artist is a nonspecialist who opens up higher, or more truthful, realms of existence and transcends the norms of everyday life. Theories of "artistic inspiration" and of the artist as martyr (or scapegoat) arise from this culture and continue to be generated in it.

Sources of revitalization

The courtly culture. The artist, who may be an aristocratic poet or a craftsman of ordinary social standing, is directly dependent upon a secular royal court or a household of the high nobility (or, by extension, plutocracy). To some extent, he is even socially involved with this household as a vassal, royal favourite, or court artist. He therefore not only produces for but also spiritually identifies with the high aristocracy for which he works, without actually being a member of it. The artist is a man ennobled by his art; his art must therefore be permeated with "noble" attitudes: heroic exaltation, fashionable late-medieval despair, or refined Rococo or Rājput sensuousness. While this art, by glorifying the establishment, reinforces its privileges, it may, at the same time, educate its—and other people's—emotions in novel ways.

The gentlemanly culture. When men of high social standing and independent economic means become active in the production of art, a tradition of gentlemanly art may arise (as is exemplified in much of the calligraphy and landscape painting in historic China). Art tends to be contemplative, subjective, aestheticizing—but not consciously "ennobling," since artists already possess sufficient "nobility" (as courtly artists do not). The artist is a man of independence who observes and contemplates without desire (an attitude that has been influential in modern aesthetic philosophies, frequently produced by such men).

The vagabond culture. All the traditional types of artistic cultures presuppose a firm social location of the artist in relation to a community, except where the artist is a vagabond adventurer who roams the countryside producing or performing works of art—comical tricks, storytelling, lyrical songs (for example, the medieval minstrels or the 15th-century French poet François Villon). In origin, he may be an uprooted folk artist, who belongs only to temporary communities. He is neither an apprentice journeyman nor a gentleman (who may also travel widely) but an impecunious amateur master unconcerned with material gain. (Masterful amateurism in the arts has two sources: the gentleman and the vagabond.) In contrast to the ecstatic artist, the vagabond does not assume that his works possess any higher symbolic significance. They are plays of fancy that amuse the fancier and whoever cares to join him. Individual vagabond artists have probably always existed, but vagabond cultures can arise only in the fringe areas of cities (for example, a transitional zone between industrial and residential quarters) or where nomadic tribes like the gypsies travel over a settled countryside. Travelling theatrical groups exemplify both conditions. Whatever the effect of vagabond artists on society, they, like the ecstatic artists, tend to have a liberating effect on the arts in which they engage.

Modern artistic cultures. Several types of artistic cultures, in some cases anticipated in earlier periods, have come to a full development in industrial societies. Some older types have been transformed, and new ones evolved. In the advanced industrial societies, artistic cultures tend to become fluid—composed of overlapping and fluctuating elements, components of a mixed orientation rather than rigidly separated alternatives. Nevertheless, several types can be distinguished.

Artistic cultures in industrial societies

The genius culture. The product of a conjunction of the artisan, clerical, and courtly cultures, the cult of genius emerged in the early 16th century during the High Renaissance and became more fully developed in the age of Romanticism at the beginning of the 19th century. The artist conceives himself as the "unacknowledged legislator of the world" (in the words of the English poet Percy Bysshe Shelley), an autonomous, godlike creator of new orders of reality obedient only to his perceptions and the categories of his mind. He is superior to the specializations of the sciences and the crafts: he creates the unifying symbols of a developing civilization. He is self-confident, a proclaimer of new values rather than a critic of the established ones. Historically, perhaps the greatest significance of the genius has been in the legitimation he has provided for the developing conception of the professional artist.

The professional culture. The most essential characteristic of the artistic professional is that he himself chooses what to produce or perform. But, unlike the gentleman artist, the professional produces in order to sell. His art is the activity of a specialist competent in the techniques of expressing subjective perceptions. In the visual and the literary arts, the professional usually works in his own studio, at times of his own choice. In the performing arts —which are less "professionalized"—he is dependent on facilities provided for him by entrepreneurs and impresarios. He may have an agent who represents his economic interests in any case. But the essential point, in the artistic profession, is not the ownership of the means of production; it is the artist's sense that he can, and indeed must, depend upon his personal aesthetic sense. This is the predominant image of the writer in modern Western societies.

A sociologically important distinction can be made between private and public professionals. The first have originated in the gentlemanly culture, the second in the salon–coffeehouse–journal circuit of the 18th and 19th centuries, in which the enlightened aristocracy and the increasingly popular press joined forces in supporting a socially critical collectivity of middle class artists, journalists, and hangers-on. In this public culture, it was most important to be up to date and to respond to current is-

sues by one's wit. As this culture began to disintegrate, the distinction between private and public professionals was partly replaced by that between bohemian and radical avant-gardists.

The applied-arts culture. The old artisan tradition has continued in the modern applied arts (for example, industrial design, advertising). The artist is, however, not usually organized in guilds but, rather, is employed in a white-collar type of job by large business firms or governmental organizations, sometimes by "artistic firms," which, in their management, are not very different from other business firms. But while the applied artist is employed as a craftsman, his attitudes have been affected by the culture of professional art. He may therefore be a craftsman with aspirations to professionalism and may tend to perceive himself as failing to realize his artistic aspirations, regardless of financial success and social recognition. Applied artists with the self-conception of professionals provide one of the main supports for movements of radical criticism of contemporary society and culture.

White-collar artists

A development toward the applied arts may be occurring at present in such an established profession as architecture: the increasing costs over which he has little control, the various restraints over his activity (zoning codes, labour union regulations, and the like), and group pressures to which he must respond are tending to transform the architect into a high-level organizational craftsman. As their opportunities for being artistic professionals decline, architects tend to conceive of themselves increasingly in the image of social engineers—a trend noticeable also in the other arts.

The mass-arts culture. The notion of the mass arts implies that the artist communicates with his public through the mediation of some type of mechanical or electronic machinery; that he does not know even the kinds of people of whom his public will eventually consist and may be unfamiliar with the characteristic experiences of most of his actual public. Furthermore, the responses that he will get from his public will probably also be, in most cases, mediated either in a mechanical manner (statistics of sales) or by professional intermediaries (newspaper critics). The film, television, mass-periodical fiction, and recorded music represent the clearest cases of mass art. Insofar as the mass media deal with any art, they tend to assimilate it to a greater or lesser extent to the mass arts. Because of the conditions under which he operates, the mass artist lacks a sense of general standards—other than the purely technical ones of craftsmanship—by which to orient his artistic activity. Nor, since he is dependent on large, impersonal audiences, can he trust his own convictions to guide him. His typical solution is to cultivate an "image" that has proved to be popular with a mass audience and that, once established, is "forced" on him by his public. He becomes a victim of his own image, in the sense that he is trapped within its confines. He is an uncertain specialist in the symbolic manipulation of diffuse audiences.

Cultivation of a popular image

Avant-garde cultures. Another characteristically modern artistic culture is definable by two aspects: a principled sense of alienation from significant aspects of existing reality and a conscious commitment to overcoming the deficiencies of existing reality by artistic means of a completely novel character. One of the distinguishable types of this artistic culture is the bohemian avant-garde, which tends to be alienated from all rational and utilitarian aspects of social organization and cultural tradition and aims to create a new kind of exaggeratedly irrational art and, perhaps even more important, an irrational style of life (typical of this group is the French poet and critic Charles Baudelaire). Its behaviour centres on the ambivalent and the self-consciously paradoxical: a tradition of pursuit of the new, the cultivation of a pleasureless hedonism, the development of systems for the liberation of spontaneity. Another type of this culture is the radical (politically committed) avant-garde, which regards itself as alienated from "oppressive" and "exploitative" political and economic institutions and tries to create a new kind of art intended to undermine faith in these institutions and to provide a basis for their abolition or reconstruction (an example is the German poet and playwright Bertolt Brecht). These two avant-garde cultures overlap, and, over time, one may change into the other. The alienated type is more influential among artists when faith in ideological utopias declines generally in their society. Perhaps because of its tendency to limit itself to consciously recognized and intended purposes, the aesthetic achievements of the radical avant-garde (the "prophets") have been less significant, so far, than those of the alienated avant-garde (the "mystics").

More recently, a third type has emerged: the anti-artistic avant-garde, which is alienated from the very notion of art and its practice as heretofore conceived, even (or perhaps especially) in the other versions of the avant-garde culture. Art itself is perceived as "oppressive" and "exploitative," and the obligation of the artist (and of the art critic) is seen as the promotion of the "end of art"—its diffusion into a "life" of childlike impulsivity. The very coherence of art is felt to be an imposition of an arbitrary system on the immediacy of "aesthetic experiences," which are to be pursued with a self-conscious repudiation of any deliberate control.

All the avant-garde cultures can be seen as diverse outgrowths of the culture of genius, with which the peculiarly modern developments in the arts began. But the bohemian and radical version of the avant-garde culture are also reactions, by artists and aesthetically sensitized intellectuals, to the development of the modern industrial society—a society highly rationalized and, in spite of its ideology celebrating the "common people," continuingly stratified. The anti-artistic avant-garde appears to be a "shamanistic" reaction against the professionalization of art, and, like the mediumistic trances of traditional cultures, its efforts may prove to be a socially innocuous (or conserving) ritual that is perhaps psychologically refreshing.

Total-command cultures. Hitler's Germany and Stalin's Soviet Union still provide the best examples of what happens when a modern artistic enterprise is subordinated to a political ideology that has the monopoly of organized power in a society. Artists are forced back into clerical roles—with the difference that the moral responsibility for upholding an ideology is imposed on them, without its necessarily corresponding with their own convictions; and, even when it does correspond with their ideological views, it conflicts with the conception of the professional role that they, as modern artists, regard as their primary orientation within the sphere of art. An artistic enterprise can be effectively politicized, without losing most of its aesthetic qualities, only if artists are content with assuming the role of "clerics" illustrating, with appropriate technical means, the manuscripts of a political ideology, yet do so without limiting their imagination to service in its mission.

Politicalization of art

The scientific–technological culture. Partly under the influence of imaginative developments in science and technology and partly out of disillusionment with the avant-garde cultures of art, an artistic culture in which the artist is assimilated to the role of a scientific researcher has arisen. The artist is a strategist of concepts, a deviser of technical systems for others to build, and a manager who invents ways for workers to relate technological or natural processes to each other. His difference from the scientist is that he seeks not new understandings but new perceptions; his difference from the technological experimenter is that the systems constructed are expected to have no utility. They are frequently evanescent or self-destructive. This type of artistic culture requires large financial outlays and tends to result mainly in entertaining decorative effects. It is neutral with respect to the particular institutions of a society but tends to glorify the technological and economic achievements that enable it to sustain such works of art. Thus it is the equivalent, in a technological civilization, to the monumental art by which absolute monarchs had been glorified in the agricultural empires, such as ancient Egypt.

Movement cultures. Artistic situations that arise in spontaneous movements of a general cultural, rather than

specifically artistic or primarily political, character (such as youth movements) may generate an immediacy and intimacy of the relationship between producers and consumers of art that is rare anywhere else in modern societies. The effect on the arts is to simplify aesthetic structures and integrate them more closely with the experiences of the many. The arts most frequently involved are poetry, music, and some form of theatre. The movements themselves are not highly productive of enduring artistic achievements. Their main significance is in opening up large audiences to a sensitive responsiveness to art or to particular kinds of art—a sensitivity anesthetized by the mass arts.

Amateur art making. With the increase in leisure time, amateur art making has become increasingly popular in advanced industrial societies, as had amateur musical performances in the European upper and middle classes in the 16th and 17th centuries. This phenomenon contributes to the blurring of the traditional distinctions between producers and consumers of art and to the involvement of larger numbers of people in a more active manner with the artistic enterprise, strengthening the social anchoring of art in the modern society.

III. Economic support of the arts

ECONOMIC EVALUATION OF THE ARTS

Factors in the allocation of wealth to the arts

The economic evaluation of the arts in general is indicated either by the absolute amount of economic resources put into their production, acquisition, distribution, and consumption or by the relative share of the total income of an individual, group, or society that these allocations represent.

While the absolute size of expenditures may be determined more by the level of wealth than by the degree of interest in art, it yet has a significant effect on the artistic enterprise, perhaps especially on art collection and prices of artworks. When the absolute amount of funds available for the arts is very large, it produces the phenomenon of "cultural imperialism": it enables the powerful, whatever the degree of their interest in art or their own creative attainments, to dominate and overwhelm the cultural activities of the financially less well endowed, unless the latter remain isolated or protect themselves by consciously designed cultural policies.

The percentage of the income devoted to art, on the other hand, bears a closer relationship to interest in art, and it could therefore be expected to be more closely associated with the level of artistic creativity—which does not necessarily require large economic resources to sustain it. Some of the finest works of primitive art, for example, have been produced in places where life is economically precarious, such as the swampy areas of New Guinea. The percentage of wealth devoted to the arts tends to be greatest in economically comfortable societies that have passed their peak of economic expansion—Italy after the middle of the 14th century and Spain in the late 16th and 17th centuries, to cite two instances.

The economic valuation of art is also affected by the degree of its association with activities that are regarded as possessing great importance in a particular society. Actual importance of an activity can be measured, in economic terms, by the size of economic allocations for its pursuit. The closer art is integrated with the activities (typically, the economic, political, religious, and military) getting the major share of available funds, the larger tends to be its own share. In the advanced industrial societies, a major financial basis of contemporary arts is their integration with the growing leisure industry (as in the Broadway theatre or the arts in the mass media) and with education (as in the widespread employment of artists in schools and universities). It is partly for this reason that modern states tend to allocate larger sums for the support of the performing than of the object-making arts and for the dissemination than for the production of art objects.

To the extent that art itself is a prestigious activity, individuals, groups, cities, or whole societies may compete for supporting its most important living practitioners or acquiring the most famous works. This motive has probably been present, to some degree, in all elite-oriented societies, but its significance tends to increase in civilizations that have many centres rather than a single "imperial" one—in Renaissance Italy, in Germany before its unification, and in the modern world in general. Competition of would-be patrons for the prestige of living artists increases not only the economic allocations to art but also the freedom of "elite" artists. Modern state support for the performing arts, in particular, rests to a high degree on competition for internationally recognized cultural reputations.

The value of art as an investment of wealth, either for ensuring its preservation or for increasing it, grows in periods of social and political instability and currency fluctuation. Small-scale artworks made either of precious materials or by prestigious old masters become the safest investment in times of trouble. This has been an important motive for the collection of art by the rich in the Renaissance as well as in the 20th century. It is characteristic of this system that a high economic valuation of art does not necessarily correlate with the income received by living artists, since the object of speculative interest is the reputation of a master, who is frequently dead, rather than the aesthetic merits of the work itself. It is a system in which artists, instead of being supported by the rich in the pursuit of aesthetic values, support the rich in their pursuit of further enrichment, since an artist may endure a lifetime of poverty to produce works that then become tokens of steadily increasing worth. As a consequence, the system encourages capital accumulation rather than artistic creativity. It has little effect on the performing arts or literature but is currently a major source of discontent among practitioners of the visual arts.

Relation of achievement to value

Increases in the social status of artists and in the prices their works fetch on the market have, in the past, tended to follow increases in the creative attainments of an artistic tradition. Thus, in England, an increase in artistic creativity in the earlier part of the 18th century (as exemplified by the paintings and engravings of William Hogarth) was not followed by a significant increase in the social and economic status of some artists until the 1780s. During the late 19th century, however, when contemporary English art was admired above that of any other period or society and highly rewarded, it produced only minor achievements. Clearly, economic allocations do not guarantee artistic efflorescences, and inferior art may be ascribed high economic value. Great artistic achievements, nevertheless, appear to increase the economic value of art, though perhaps not immediately. While the economic value of particular objects of art depends in part on their historical significance, uniqueness, and fashion, the economic valuation of art itself is, in the long run, not altogether unrelated to the presence in a society of large numbers of aesthetically meritorious works.

SYSTEMS OF FINANCING ARTISTIC ACTIVITIES

The economic support of artistic activities can be provided by the artist himself, who derives his income from sources other than the rewards he receives for producing art, or it can be furnished by nonartists supplying the economic means for the artist to survive while he is making works of art.

Self-financing of artistic activities has always been, and still is, considerable. The gentlemanly type of artistic culture is wholly supported by artists of independent means. The writing of lyrical poetry has very generally been a leisure-time activity of persons deriving the bulk of their income from other sources. In modern times, self-financing is present when the artist survives from an activity that does not require him to produce works of art, or when his income from producing works of art is significantly lower than what he could earn by being employed in a legitimate alternative occupation accessible to him. The dance is heavily self-financed in the second sense.

Sponsorship of artistic activities may be said to occur when the artist draws upon the economic support of individuals or organizations that do not necessarily expect to get anything for themselves in return for it. They support the artist because they are committed to art for its

own sake or to a particular artist. In this specific sense, artistic activities may be sponsored by government agencies supporting art as a "social service" for the people, by private foundations, by individual sponsors, by relatives who support the artist while he is economically unsuccessful, or by friends and acquaintances whom an artist "sponges on." Sponsorship may thus be, in some cases, unintentional.

Financing outside of the market system

Self-financing and sponsorship are economically "abnormal" systems of financing the production of goods and services that potentially possess an economic value to people other than their producers. These modes of financing are found principally in the "cultural" activities, above all religion and art (but also in "ideological" politics), and they occur because artists are not treated as if they were engaged only in the production of economic values. They produce economic values, but there is also a "religious" aspect in their activity that requires to be supported for its own sake. If the arts were to be supported entirely by a market economy, this aspect might well disappear: art would then become purely a commodity.

The arrangements of economic support that have traditionally carried the main burden of sustaining artistic activities are the three types of normal economic systems into which the arts may be integrated like any other productive activity: the exchange, command, and market economies.

The exchange economy. An artist in the exchange economy produces for a customer familiar to him who, in turn, directly or through a series of similar exchanges, provides him with a desired good or service of equivalent value. The transaction between the producer and consumer is, ideally, a ritualized exchange of gifts, and, in addition to providing the goods that each needs, it reinforces their social relationship. This type of economy prevails in the egalitarian type of preliterate societies and in the folk type of artistic cultures. In more advanced societies, this becomes a minor part of the economic support structure of the arts; it tends to be limited to the private circle of artists and their friends. The economic rewards to artists within this system are small, fairly continuous, and not greatly differentiated in magnitude.

The command economy. The artist in the command economy depends on a consumer—an individual or organization—of superior status who has extracted a great deal of wealth from the producers in the economy and can distribute it, perhaps within the boundaries established by tradition, as he sees fit. This economic system is dominant in a range of societies extending from the Kwakiutl Indians of the northwest coast of North America through the classical absolutist monarchies to modern totalitarian states. Usually, however, it does not completely displace elements of other economic systems. In this system, the artist depends for his support on relatively few patrons, who are far more powerful and socially resourceful than he is. He may be well rewarded by them, but he must produce works that will be regarded as possessing high value—usually both aesthetic and economic —by his patron. He is paid well if he produces economically valuable aesthetic goods and services. (A significant exception to this principle has been built into the situation of women in the performing arts: in male-dominated societies, they tended to be well paid if they performed sexual as well as artistic services.) Before the modern age, all artists who produced expensive works of art were supported by economies of this type. Even today, architecture—to the extent that it is an art—and the large-scale "technological" arts are supported in this manner.

Patron systems

The market economy. In the market economy, many artists compete for the favour of numerous customers not greatly differing in their purchasing power and not personally known to artists. Since artists are not familiar with their customers, their products become either acts of self-expression or of appeal to an image, an abstract stereotype, of the "average customer." In reality, market economies in the arts only approach this condition to varying degrees: some publics are known to their artists, and there are always elements of the command economy limiting the operations of market economies. There are always economic or political elites or powerful business and political organizations, at whose command, to varying degrees, artists must be in order to achieve "success" even in market economies. And in the modern system of art trade, the mass media have assumed the character of a command element in a market system. To be able to compete for numerous potential buyers, artists must first ingratiate themselves into the favour of a relatively few taste makers in positions of power within the system of mass communications. The more centralized this system is, the more of a command character it acquires.

In a pure case of market economy, the artist's freedom is great, but his rewards are highly unpredictable and uneven; his uncertainty and sense of alienation from his public is a natural consequence. In market economies, the arts are particularly dependent on support through the economically "abnormal" systems of self-financing and sponsorship. These systems typically provide the economic basis for radically new departures in the arts in market economies. They also subsidize the established arts when their costs are too high to be supported by the market (symphony orchestras, for example).

Competition for the "best"

When there are several basically independent command economies in adjoining areas, all interested in art, they are likely to compete with each other for possessing the "best" living artists (as well as the works of the most prestigious dead masters). This competition is apt to favour artistic creativity in a way that pure market economies, concerned with the production of the most usable rather than the most perfect, seem to be less capable of. First of all, pure market economies are hard put to concentrate sufficiently large economic resources to finance expensive artistic achievements (except where they can be immediately "consumed" by large numbers of people, such as in film making). Beyond that, the expectation, typical of command economies, that the artist will produce the most perfect object he is capable of producing may actually increase his motivation (and consequently his capacity) for doing so. Market economies do not hold forth any such expectation. No longer economically supported, expectations of the artist's commitment to excellence in a market economy come to depend mainly on the professional integrity of the artist.

It seems probable that a given allocation of funds to artistic activities can be more directly translated into high artistic achievements in command economies than in market economies—provided, however, that the command economy, in addition to insisting that an artist produce the most perfect, also permits him to do what he is best at. It is in the latter respect that the command economies of modern totalitarian states have fallen short: they have prescribed for their artists a manner of working in which the latter could not do their best work. Mass-media "command economies" may also work to that effect.

Market economies are potentially more capable than command economies of integrating art into the private life of ordinary members of society, and not only into the life-style of the elite. But market economies seem less capable, by themselves, of generating high artistic achievements (as well as producing artistic public environments). There is, therefore, likely to be an increasing demand for art in market economies but a reduction in the quality of what is being produced. There may be other, quite powerful stimuli to artistic creativity in societies with market-type economies, including the creative entrepreneurs who succeed in joining the market mechanism to an artistic purpose. But the impersonal workings of the economic system of the market itself appear to interfere with the accumulation of the "creative capital" while encouraging its dissemination.

THE ART MARKET

The art market is a complex system, with roots in preliterate societies and historical civilizations, by means of which artistic activities are organized as profit-making enterprises. Expectably, this system is most fully developed in capitalist economies, but, since economic enter-

prises generally have to take profitability into account, even in state-managed economies, the art market has some degree of existence wherever art is treated as an economic good that is offered, for a price, on a market.

The operations of the art market are determined both by the general dynamics of market systems and by the peculiar character of art as a commodity.

Sellers' options

People buy art for various reasons. But the organizations that specialize in selling art (or also in producing it for sale, as theatrical companies) are built in most cases on one of two assumptions about the motivations that propel people into spending their money on art. There is an elite market that sells reputation and a popular market that sells entertainment.

Selling entertainment. The popular market provides art for customers who seek primarily entertainment. It must, therefore, adjust to the prevailing conceptions of what is entertainment or shape such conceptions to correspond with what it is capable of providing. It conducts market research into audience preferences, but it also manipulates these preferences by spreading the impression, through its publicity agents and the mass media of communication, that entertainment is what the popular art market dispenses: not only its finished productions but the whole behaviour of the people involved with them is "entertaining." The artist must "sell" not only his art but also his behaviour outside of the art system. The implicit aim is absorbing everything that appears to the mass audience to be entertainment into the ambience of the popular art market—by transmuting the entertaining aspects of life into salable works of art, by associating entertaining people who are not artists with the popular art market, and so on.

Conversely, whatever cannot be regarded as entertaining tends to be forced out of the works of art or artistic performances handled by this system. It puts a premium on "slickness," the glittering or sensational packaging, and discounts both high seriousness and unique sensibility, unless they can be seen as "entertaining." Its influence spreads even beyond the arts it deals with, affecting the character of art that people highly exposed, from their early years, to this system produce and consume even when they operate outside of it.

Selling reputation. If the popular market sells entertainment (an important means of overcoming sensory deprivation), the elite market sells reputation (an important resource for legitimating high social status or aspiration). It is, therefore, involved only with customers concerned with reputation. Such customers may be individuals or organizations. While the "masses" are mainly involved with the popular market, organizations, including states, tend to be more closely linked with the elite market. This market attunes itself to the prevailing hierarchies of artistic reputation, but it also reinforces these hierarchies by spreading the news of the high prices (or prizes) fetched by the works of art handled by it, by encouraging competition among buyers (or supporters) of unique art objects, and by subsidizing the work of scholars and critics that is likely to draw the attention of purchasing elite groups to their goods, increasing their cultural reputation.

Transforming cultural significance into economic value

The elite market has a monopoly of the "measuring rods" used to transform cultural significance into economic value. For the buyers, these scales of measurement indicate the degree of their "cultivation." If impresarios in the popular market have the role of tastemakers (as well as organizers of a complex set of productive activities in the performing arts), dealers in the elite market operate as cultivation makers, cultivation being economically definable as the rate at which cultural reputations are paid for. Art sellers for the elite do not create the reputations, but they prosper by increasing the price of acquiring cultivation.

On the living artist, the elite market has the effect of encouraging him to produce a reputation that, highly paid for, identifies the purchaser as a person of cultivation. The artist's works must be the opposite of what the popular market deals with—they must *not* be entertaining but rather culturally "significant" (in accordance with current conceptions of cultural significance). In the 20th century, cultural significance is regarded as virtually equivalent to being innovative.

Interaction of popular and elite markets. Markets that are oriented to elites but deal in mechanically reproduced multiples, such as the publishing of "serious" (as contrasted to "popular") books, represent a special case in between elite and popular markets. The objects they deal in are not inherently "entertaining," and yet the cultural significance an individual object is seen to possess does not lend itself, when the object is multiplied many times over, to being transmuted into inflated prices on the market for reputations. Publishing costs have become so high that any published book, unless subsidized, must enter a popular or semipopular market. The transformation of book publishing into an industry oriented to a fully popular market has been encouraged by the paperback revolution and the profitability of motion-picture rights for both publishers and writers. A peculiarity of the book market is that it has developed semicaptive publics, in the form of book clubs, which the other arts have not been as successful in organizing.

The motion-picture industry is in the popular market. So are all the "spectacular" arts, such as the opera and the ballet, insofar as they are economically dependent on a market—even though their markets have been, by tradition, smaller and more selective than those of the cinema. The smaller and more select a market of the popular type is, the more important seems to be the influence of professional critics on it. And the more influential the critics, the more closely does a popular market approximate an elite market.

Relation of selling to quality

As the history of the opera and the ballet indicate, it is not to be assumed that a market concerned more with selling entertainment than with selling reputation does not permit the production of high artistic values. Even though a public is presumed to be spending its money for entertainment, entertainment is not necessarily what the artists think they are selling—one reason for the alienation of the artist from his public. The great impresario in the popular market is the one (such as Sergey Diaghilev) who arranges for artists to sell their best art to a public convinced it is getting glorious entertainment. Pioneering in new forms of entertainment may both change notions of entertainment and produce objects that will enter the elite market. It is not primarily by the objects produced but, rather, by the manner of their economic operation that one may distinguish between the popular and the elite market in the arts.

Popular and elite markets are not wholly separate. Prices of paintings by old masters in the elite market tend to be increased by their pleasing appearance and are usually reduced when they depict inelegant scenes of suffering. The elite market operates most purely, in terms of reputations, when art museums are involved. Conversely, any "respectable" popular market uses (and manufactures) reputations to increase the economic value of its product. Only anonymous entertainment—by "nameless" artists—can constitute a purely popular market.

The popular market, by treating art as entertainment, imposes on consumers' contact with art the expectation of fleeting experiences and of the absence of any cultural significance transcending them. The elite market imposes the expectation of enduring values and of the presence of the highest kind of significance. In both cases, the market exercises an influence over the definitions of art held by the art consumer (and even by the artist, to the extent he is governed by his market situation rather than other elements of the artistic role). The effectiveness of the influence depends on the degree to which market considerations alone govern artistic experiences and activities.

In both the popular and the elite markets, contrary to the usual dynamics of the marketplace, increases in the number of artists and in the amount of their production, relative to the size of the demand for their work, do not necessarily reduce the prices of works of art. In modern societies, these trends produce a focussing of attention on a few (who receive very high rewards) and the neglect of the rest of art producers, many of whom may be only slightly inferior (or not inferior at all) to the "stars."

Publicity in the mass media, which both the popular and the elite markets manipulate and benefit from, reinforces this tendency by promoting celebrity cults. The objects of these cults may be intrinsically meritorious or not, but the cult, in either case, destroys proportionality between merit and reward. Mass media tend to promote in the arts a phenomenon similar to the medieval cult of sacred relics. Current profitability of investments in art objects rests, perhaps mainly, on a cult of this sort.

Artists' unions

Trade unions of artists—a recent phenomenon—are important mainly in the popular art market. They establish lower limits for the incomes of working artists but also increase the cost of full-scale artistic productions, particularly in the performing arts. In effect, they seek to eliminate the need for the performing artists to finance their art by accepting an income less than commensurate with what they could get from an alternative occupation. But the increases in costs of artistic performances render the performing arts more dependent on sponsorship systems. Neither the European theatre nor symphony orchestras anywhere could survive in a perfect market system. While increasing the economic security of employed artists, artistic trade unions tend also to promote the unemployment or the underemployment of the performing artists and the growth of a "serious amateur" system (for example, the off-off-Broadway theatre and other experimental fringe theatres), which, in addition to developing a style alternative to that of the professional theatre, may also nurture fresh talent and feed it into the professional system. Artists working for the elite market have been less concerned with establishing trade unions. They either survive as celebrities—or fail to survive.

Marginal phenomena. In addition to the two major types of art markets, one can distinguish two currently marginal phenomena: traditional (that is, noneconomic) elements in the art market, such as ideological commitments to particular types of art or group commitments to particular artists, and the various "little" markets, at present increasingly numerous in Western societies, in which art is sold not exclusively, or not mainly, for profit and which may indeed be partly financed by proprietors or participants from income not derived from the sale of art or through their own unpaid services (little magazines, artists' cooperatives, the U.S. "underground" cinema at its beginnings). Little markets may be individual or cooperative, expert or amateurish, and built on the most varying motives. While their cultural role may be important, particularly in nurturing difficult artistic innovations, the little markets collectively cover only a minor part of the art market. If they become economically successful, they tend to be incorporated into one of the two major systems of art trade.

Incorporation of traditional elements into art markets

The "traditional" elements are usually incorporated into one of the three types of art markets and limit the intrinsic logic of their operation. In the elite markets, nation states impose restrictions on the export of historically significant works of art; all respectable book publishers print a certain number of worthy books they expect to be altogether unprofitable. In the popular markets, traditional criteria establish limits of "proper" and "improper" entertainment (performances of avant-garde musical works, for example, are consistently less profitable, both in Europe and the United States, than those featuring more traditional ones). On the little markets, which tend to lack economic staying power, commitments to an identifiable tradition exert a stabilizing effect.

REMUNERATION OF ARTISTS AND PROTECTION OF THEIR RIGHTS

Methods of remunerating artists depend on whether they are involuntarily bound (serf artists) or, with their own consent, are attached (palace artists) to a consumer or organization that uses their works, or are free to sell them on the market.

Bound artists are provided with subsistence by their lords; attached artists receive regular salaries (which may be high in the case of palace favourites and academicians) and additional bonuses for particular successful works of art. Freedom to sell on a market may be unlimited, as in modern societies, or it may be conditional on membership in a socially recognized group of producers (the medieval guild system). To the extent that it approaches a perfect market (numerous artists of approximately equal artistic reputation competing for numerous customers of approximately equal purchasing power), the economic transaction takes the form of a sale; in command economics it is more usually a commission. In guild systems, the association of producers may itself transact the sale or at least regulate the conditions of sale, by its members, of their products. In a dealer system, which was known in classical antiquity but (in the visual arts) has developed most fully since the 17th century (in the Netherlands) and the 19th (in France), a group of commercial intermediaries—dealers and artist's agents—has stepped in between individual artists and individual purchasers. Their task is to introduce the one to the other and conduct the economic transaction in accordance with the practices of good business. (Some art dealers and their literary equivalents—writers' agents—have operated also as patrons and even sponsors of young artists.)

Methods of payment

In market systems, methods of remunerating artists depend on the traditional cultural reputation of the art they practice, the degree to which the artists' rights of authorship are legally recognized in an economically consequential manner, and the technological requirements for transacting the sale of an artwork.

When the traditional cultural reputation of an art is high, the artists who engage in it, while not necessarily highly rewarded in economic terms, tend to acquire a moral right to a socially recognized authorship of their works. Their names are affixed to the works they have produced; these are no longer products of anonymous craftsmen or works signed by the supervisors of art workshops, as were lacquerworks from the imperial workshops of China.

The conception of individual authorship began developing in Greece around 700 BC and in China more than 1,000 years later. It was known in the medieval West and India but remained undeveloped in the Byzantine civilization. This conception is directly dependent on the development of a professional or gentlemanly or genius type of artistic culture. Indirectly, it is supported by a strong stress on individualistic values in a cultural tradition, but, as in eastern Asia, the notion of authorship could arise even in the absence of any strong commitment to individualism.

The notion of a legally protected intellectual property of an art creator in his work, which he retains even after he has sold his work to a user, presupposes a conception of authorship but extends beyond it. Historically, it is a much later development. It began in the West in the late Middle Ages and has been encouraged by technological inventions facilitating the reproduction of works of art (before book printing, possession of a manuscript implied the right to reproduce it); the growth of a competitive market in such reproductions, which needed to be regulated (the most prominent motive in the early phases of the copyright law was the desire to protect the economic interests of the book publishers, rather than the intellectual rights of the authors); and the image of the artist as a man of genius who may sell, like other producers, the works of his labour but retains a right to the spiritual substance, uniquely his own, that he has invested in them. What the copyright and unfair-competition laws protect are tangible contributions to a product that remain identifiably individual. The artist has a legal right only to what he has worked out; the law protects the labour of elaboration, not the idea or intuition. Yet, once the notion of a legally protected intellectual ownership has taken root, it tends to become self-perpetuating, even in the absence of the conditions that have favoured its development.

Implication of intellectual ownership

The retention of a moral right in the artistic product provides the rationale for an author's sharing in the income that the buyer derives from the use of his work and from the appreciation of its value over time. The logic of this system has been worked out most fully in the publishing and the mass-communications industries (radio,

the film, and television). In the visual arts, efforts are still being made to extend the economic implications of the recognition of the author's moral right to his work by developing a model contract for the sale of a work of art that will not only guarantee to the artist control over the reproduction of his work (which he may claim even now, by registering his works for copyright protection or stipulating at the time of the sale that he retains this right) but also provide for his share in any profits from the resale of his work.

The contemporary copyright system guarantees to the writers covered by it that their incomes will be proportional to the current economic value of their works (the total income from their sale). The visual artists' income, on the other hand, is unrelated to the current economic value of their works, once these have been sold (except indirectly, as the prices of the works they will sell in the future are affected by appreciation or depreciation of the works sold earlier). Thus, there is a disproportionality in the effects of the law on the relationship between the artist and his product in literature and the visual arts.

In the performing arts, unless the performance is recorded, the artist cannot be guaranteed any direct economic benefits from being recognized as the author of his performance. The performance "vanishes" after it is completed. It is only what is being performed (the text, choreography, musical composition), and not the performance itself, that can be protected by custom or the law. There is thus a difference between the legal rights of performing and "producing" artists in their art, and this difference may well affect their overall social status and self-esteem.

Legal rights in the mass arts

The development of the mass-arts industry has made it possible for performing artists to enjoy the benefits of legal protection of their continued "intellectual possession" of their performances, if they have been made to be recorded and can be reproduced. Recordings of live performances have remained unprotected and frequently "pirated" or reproduced without permission. In 1971, however, a convention was signed by 53 countries prohibiting international piracy of musical recordings, and individual countries are developing legislation to control its domestic variants. If only in the mass-arts industry, the situation of performing artists has been made similar to that of such "producing" artists as writers and composers, who receive royalties on their past work. The performers in a film continue to receive a set rate of payment from each showing of that film and residuals from reruns on television. While in this way they have a sort of continuing property right in their performances, as specified in their contract, performers have no moral right of authorship that would permit them to control the manner in which their work is presented to the public (as writers of books do). This right is usually vested in the film director, and even then, only if it is specifically assured to him in his contract.

Whether the legal protection of artistic property has in fact encouraged the production of high-quality art appears debatable: the system does not distinguish between high- and low-quality products. The operation of the legal-protection systems in the arts must be judged by the degree to which it helps artists both to earn an income and to derive self-respect from their work. It must, however, also be judged by inquiring whether its provisions result in larger benefits to the artists collectively or to the sellers of their works collectively.

ART COLLECTING

Art collecting and the building of architectural ensembles, such as churches and palaces, which can be regarded as immobile "art collections," have probably provided the most important source of income for the major visual artists, particularly in command economies. Once artworks enter a collection, however, they tend to stay there for long periods of time, potentially reducing the space and the demand for the works of later artists. There is thus a certain ambiguity in the attitude of artists toward art collections and museums, particularly in societies that possess well-stocked ones, such as Italy or, by now, the United States. Museums that tend to "enshrine" artists are in this respect worse, from the living artist's point of view, than private collections, which get reshuffled every once in a while—a process encouraged, at all times, by defeats in war, social upheavals, economic troubles, and, in modern societies, by high inheritance taxes. In contemporary societies, collecting by nonartistic organizations—governmental agencies, business firms, universities—seems likely to become increasingly important. This would mean a greater incorporation of art into real-life (as contrasted with museum) environments. As an encouragement for living artists, collecting seems to be most effective when any particular collection is intended to be temporary: that is, when it is established, exhibited, and then dispersed to make room for a different one.

Relations between museums and living artists

Purposeful collecting of works of art is most eagerly pursued in periods of great affluence and a reduction of the creative drive—in the Hellenistic age (323–30 BC), in ancient Rome, and in 18th-century Europe. The relationship between art collecting and political power has been more ambiguous. On the one hand, the powerful have liked to surround themselves with the grandeur emanated by an "overpowering" collection of art. On the other hand, states have compensated for their loss of power by cultivating an image of their cultural grandeur and refinement; art collections serve this purpose, too. But art collecting is also an appropriate response to great outbursts of creativity that have occurred in the past. And it establishes reservoirs from which later artistic efflorescences will, in part, be fed: art collections, if open to the public, provide a place for young artists to study their craft and for people in general to develop an interest in the arts.

Collecting of the art of the past may be encouraged by a lack, or loss, of faith in the creative accomplishments of living artists; by the proved prestige of the works of the past (important when the newly rich collector has neither taste nor understanding to judge by himself); and, particularly in contemporary society, by the profitability of economic investments in famous works of art. It is estimated that art prices in general multiplied some ten times from the early 1950s to the beginning of the 1970s.

The recent rise in art prices

Private art collecting on a smaller scale has been developing in the middle classes since the beginning of the modern age. Currently it shades off into "temporary collections" of multiples from the popular market, such as musical records or posters used to decorate student rooms. Attitudes derived from this practice are spreading even into sophisticated responses to art (*e.g.*, conceptions of the social function of museums).

FRAUDULENCE IN THE ARTS

Fraudulence is misrepresentation of authorship, usually for financial gain but sometimes for more complex reasons of a psychological or ideological nature. The sale of a copy of a unique work of art claiming to be the original and the sale of an original work claiming to be a previously unknown product of a famous master or of a highly valued anonymous tradition are typical instances of fraudulence. From an aesthetic point of view, a good falsification may be preferable to a mediocre original, and it would seem to "corrupt" the immediate aesthetic experience of amateur art consumers less than authentic but poor art presumably does. But a falsification distorts the evidence used for constructing an understanding of what "has really happened" in the past and for formulating sophisticated criteria of aesthetic judgment.

As an economic phenomenon, fraudulence represents primarily an assault on the elite market. On the one hand, it may reduce the funds available for buying authentic works of art; on the other, it tends to discredit the system of buying reputations. Yet it appears, in the modern world, not to have had much of either effect. Perhaps the most enduring sociological significance of fraudulence in the arts is that it has encouraged economic investment in the fraudulence-control system—art scholarship, gallery and museum expertise, and so forth. Beyond its intellectual reasons for existence, art scholarship has justified itself economically as art policing.

IV. Social control of art

Little art, except perhaps in the "intimate" genres such as lyrical poetry, can have been created, until the latter part of the 19th century, without some degree of social control—that is, influence exerted by nonartists—over its creation. The influence has varied greatly in degree, in the groups and agencies exerting control, and in the means used.

TYPES OF REGULATION

Six basic types of social control of art can be distinguished: (1) suppressive censorship by agencies in total control of channels of possible expression (the "medieval system"); (2) product specification by a relatively few customers, each of whom is not in total control of channels of artistic expression (the "Renaissance system"); (3) administrative restriction of the artist's access to his potential audience without a complete withdrawal of his opportunity to communicate and without destruction of his works ("enlightened censorship"); (4) organizational incorporation of artists into either nonartistic institutions, such as churches or business firms, artistic groups under the authority of nonartistic agencies, such as the academy under the French king Louis XIV (reigned 1643–1715), or state artists' unions (the "organizational system"); (5) expression of preferences of taste by large numbers of individual art consumers not differing greatly in the power to command and to purchase (the "democratic system"); and (6) criticism by experts specializing in sustained analysis and evaluation of individual works of art (the "intellectual system").

Censorship compared with criticism. Of the six modes of control, censorship is the most severe and least sensitive to the aesthetic merits of works of art. Art is on principle judged in terms extrinsic to artistic values and subordinated to explicitly utilitarian considerations of political, religious, or "moral" character. Criticism, at the other extreme, at least uses the means of imagination, rather than of power, to control products of the imagination and mediates between artists and groups of users of art (however small) whose standards particular critics articulate. Thus, criticism helps to relate artists to the subcultures in which their work is most responsively consumed and to reveal the character of these subcultures to their own members. At best, criticism expands the artist's experience in areas relevant to his understanding of how art operates in the minds of people and increases the consciousness of the community to which the critic "speaks." At the same time, the artist can avoid being influenced by critical interpretations (at the risk of not learning what they reveal) by the simple decision not to read the reviews. Scientific research into empirically demonstrable effects of particular kinds of art on particular types of personalities, under specified conditions, can be treated as a newly evolved element of artistic criticism.

Benefits derived from criticism

Censorship of the arts has been typically justified by "clerical" or "civic" conceptions of art—the view that art has a moral obligation to defend a cultural tradition higher than art itself or a civic obligation to celebrate the community that supports the artist. The first view is usually espoused by various ideological elites, but the second can be quite popular and widely supported, even in a democracy. The clerical conception of art also generally implies the assumption that art has a great power to corrupt or to save. It is, indeed, this tendency to overestimate the power of art that leads to demands for censorship. The civic attitude toward art presupposes merely a tendency to suppress anything that offends local self-esteem.

Whatever the conception of art, censorship becomes "necessary" only when an established clerical or civic view of art is powerfully challenged by other artistic cultures—the ecstatic, the vagabond, the genius, the professional, or the avant-garde. It is perhaps from such challenges to civic and clerical assumptions by ecstatic (Dionysian) or emerging professional types of artists that the first explicit philosophical defense of artistic censorship, by the philosopher Plato, emerged in classical Athens. A conflict between two clerical cultures, the Catholic and the Protestant, together with the invention of the printing press (which made literature more "dangerous"), led to the great development of formal, organized religious and political censorship in 16th-century Europe. Censorship declined, especially in the Anglo-Saxon countries, in the 18th century, with the establishment of religious tolerance and in conjunction with the displacement of clerical by professional conceptions of the writer.

Targets of censorship. Censorship has been mainly directed against the ideological (stated or implied views) or depictive (represented scenes) content of the arts, thus primarily against the arts in which content is more important—literature, the theatre, and the visual arts (including the film and, to a lesser degree, photography). But whole types of art have been outlawed. The ancient Spartans expunged music and dance, as well as poetry, on the grounds that they might promote effeminacy and license in a population that had to be hardened for heroic militarism. Early Christianity suppressed the theatre and fictional literature of the Greco-Roman civilization. Muslims, Calvinists, and in some periods the Byzantine iconoclasts outlawed religious visual art.

Suppression of content

While, in general, secular states have been concerned only with the content of art, ideological organizations have been sensitive, especially in 20th-century totalitarian systems, to the attitudes and values suggested also in style. It was primarily by stylistic characteristics that "degenerate" art was defined in Nazi Germany. Totalitarian movements have not only prohibited some styles but have also prescribed others for their artists to work in (for example, Socialist Realism in the Soviet Union between the 1930s and the 1950s).

The 17th-century absolutist state also perceived the value implications of artistic styles, but (like the medieval Catholic Church in its approach to the visual arts) it relied more on the techniques of product specification—patronage, by a royal court, and promotion, by an official art academy—rather than on prohibition, enforced by police power, against working in particular styles and having works in these styles exposed, in some manner, to artists' customers. As a mode of control, product specification is more congenial to artistic creativity than is censorship; and 17th-century France sustained creative attainments of a high order, even in the visual arts in which the court specified its demands most insistently. But it achieved still more in dramatic literature, over which the court had a less direct influence.

In the 20th century, the large economic organizations that have come to dominate the mass arts have acquired a capacity to exercise a private kind of censorship by depriving artists of their means of work. In the "image" industries, the reason for such blacklisting, as in the U.S. film industry after World War II, has tended to be the political image of the artist rather than either the style or the content of the work he was proposing or had done in the past. In the democratic societies, private watchdog and pressure groups are likely to be more effective in imposing their demands for censorship on the mass-arts industries or on the business firms whose advertising sustains them financially than in influencing governmental agencies. The newer mass media are, because of their dependence on public licensing (television stations), more vulnerable to governmental pressures than the traditional arts.

Self-censorship

Self-censorship by artistic enterprises (for example, the movie rating system of the U.S. film industry) has developed largely in response to the influences private pressure groups have brought to bear on the culture industries, by threatening their mass sales. A different variety of self-censorship is practiced by art museums when they ban types of art with contents that are not congenial to the economic or political interests of the business leaders who usually control their boards of trustees, or that seem capable of causing libel suits, or that appear to conflict too sharply with the traditional conceptions of what art museums should exhibit.

In neither of these two situations does "censorship" completely prevent public exposition of the works censored or terminate the public career of the artist con-

cerned. These cases thus do not represent true, suppressive censorship but, rather, a system of limiting the public's opportunities for viewing certain types of art. Prohibition of the sale of certain types of art to juveniles also represents restrictive censorship. In respect to the arts, the Western democracies have by now virtually abandoned suppressive censorship in favour of the restrictive.

In the U.S.S.R., there have been some trends after the death of Joseph Stalin (1953) toward a transformation of the system of suppressive censorship into a de facto system of administrative restriction: allowing for privileged exhibitions of modernistic art in scientific institutes, showings of avant-garde foreign films to select circles (partly specialists and partly political elite), publication in small editions or in recondite journals. But a somewhat relaxed suppressive censorship is still in operation. China has a completely suppressive system.

In contrast to the censorship practiced by the ideological organizations, style is usually of no concern to the economic interest and private pressure groups endeavouring to subject art, or what is presented as art, to censorship either by pressuring the mass-arts industries and large artistic enterprises or by demanding action by government agencies (in the United States, most frequently on grounds of "obscenity").

CONDITIONS FOR SOCIAL CONTROL

Factors in the susceptibility to restraints

In general, the arts seem to be most susceptible to social control when artists are dependent on a relatively few important consumers or on a great mass audience of a fairly homogeneous social character, and when their works are either very expensive "uniques" or highly profitable "multiples." The susceptibility of artists to social control thus depends, to some extent, on the kinds of art they choose to produce. This susceptibility tends to decline when there is a differentiated, heterogeneous public with political institutions permitting a free expression of individual choices and an inexpensive access to a wide range of works of art. If access to art is provided at public expense, however, artists again become susceptible to social control—this time by the intermediaries staffing the cultural organizations, such as museums and state publishing houses, that determine which artists and which works will be exposed to the public.

Art is least susceptible to social control when it is sponsored and financially supported by other artists, who acquire their income in some manner other than through the sale of their works (for example, through teaching). The institutionalization of the professional conception of art is perhaps, in the long run, the most reliable defense, from within the artistic enterprise, against deliberate manipulation of the arts by social agencies. But a purely professional tradition of art is so inoffensive as to appear dehumanized, and artists themselves may seek to overthrow it, opening themselves up to deliberate manipulation by those (frequently political movements) they align with to achieve this goal.

IMPLICATIONS OF SOCIAL CONTROL

The issue of censorship and other kinds of deliberate manipulation of art has ultimately to be dealt with in terms of moral and political assumptions about human nature and the proper character of the artistic enterprise, not primarily in terms of the effects of these policies on artistic creativity. The effects of social restraints, or their absence, on artistic creativity seem, in any case, somewhat ambiguous. In general, no direct relationship seems to exist between the degree of personal freedom enjoyed by the artist and the aesthetic quality of his work. The tolerance and cultivation of the art patrons in Italy during the 17th and 18th centuries did not prevent a decline in creativity. On the other hand, a general cultural repressiveness does not always preclude an artistic efflorescence: the period of the most intense activity of the Spanish Inquisition in the 16th and 17th centuries coincided with El Siglo de Oro (The Golden Age), one of Spain's most important periods of artistic creativity.

Freedom, however, may have become a necessary—though not sufficient—condition for artistic creativity in industrial societies. Artists, particularly those who have been affected by artistic developments in the West since the Renaissance, have come to expect creative freedom and cannot help but feel illegitimately constricted in its absence. Their personalities are no longer sufficiently congruent with an authoritarian structure of external controls, as the personalities of medieval artists may have been, to be able to produce aesthetically significant work when subjected to such controls. By abolishing artistic freedom, both the Soviet Union and Nazi Germany destroyed flourishing artistic movements and proved unable to generate aesthetically valid substitutes for them. But a relaxation of earlier rigid controls may be more conducive to artistic creativity than the maximum of freedom, because under restrictive controls an unspent tension accumulates, which is then available to be released in an explosion of creative activity when pressures are relaxed. Under complete external freedom, such energy may never accumulate; unless the artists possess an extraordinary degree of self-discipline, their energies tend to be immediately used up through a variety of outlets.

Effect of easing controls

What seems to be specifically fatal to artistic creativity is not social control of the arts but an imposition, whether by outsiders or indeed by the artists themselves, of a completely intentional conception of the artistic enterprise: that is, its successful limitation to the expression of any set of recognized and intended functions.

If there is to be any social control over the arts, which many artists would dispute, its least objectionable form would seem to be one limited to a combination of expressions of preference by large numbers of interested art users with analysis and reasoned evaluation by art critics and scholars. The two modes of control in conjunction balance each other's biases and increase the range of options available to artists. These controls can be dismissed from consideration only if it is assumed that art is totally irrelevant even to the people most interested in it (other than the artists themselves). Art would then have to be treated as either a private affair of the artists or as pure entertainment without any cultural significance.

V. The arts and religion

In both preliterate societies and historic civilizations, the arts have frequently, but not always, had a close relationship with religion. In the past, this relationship has tended to be less direct when art objects were being produced by women, such as the pottery of the Pueblo Indians of the southwestern United States. Women, who have rarely been the religious specialists of their societies, have mostly produced an art of decorated utilitarian artifacts or of personal intimate expression, in either case not characterized by high cultural symbolism. Men have been more frequently disposed, or constrained, to justify their aesthetic interests by relating them to a metaphysical purpose. In effect, this means that men have had more reasons, in most preliterate and historical societies, for making art than women did.

SOCIAL RELATIONSHIPS

Interaction of art and religion. An explanation for the frequently close relationship between art and religion may be found in the areas in which they are similar or overlap. In both art and religion, there is much concern with the basic needs of the imagination and with a valid perception of subtle qualities of experience, and there is no binding requirement that they provide references to demonstrable fact for their insights (as there is in science). The similarity means that the arts and religion might, in part, operate as functional substitutes for each other: the more successfully one system functions, the less need for the other, and the less successfully, the more need. This may be one reason for the modern growth of interest in the arts, especially among the intelligentsia and the alienated of the middle class.

Religious and artistic similarities

Religion, however, tends to be more dependent than the arts on those aspects of culture that are communal, in-

volved with abstract ideas, and concerned with standards of conduct. It must provide guidelines for action that whole communities (as well as individual persons) could live with. While the arts can perform such "religious" functions, they can also survive quite well without performing them: inherently, the arts have a less direct, less intentional, less "responsible" relationship to social action than does religion. In their specifically aesthetic essence, the arts are more private, more individualizing, less binding than religion. There is more play than obligation in the arts. The opposite is true of religion. Popular interest in the arts tends to decline, and interest in religion to increase, during life-threatening historical crises, such as the Black Death of 1348–50 in Europe or destructive wars. Interest in art frequently increases after such periods. Religion is more of a crisis-management phenomenon, art that of postcrisis integration.

The impact of art on its consumers is likely to be magnified by its association with religion or even, in the absence of any formal association with organized churches, by its perception in terms appropriate to religious experiences. If close alliance between the arts and religion means that they are "used" by religion, the arts become collective liturgies, providing sensuous elements to increase the hold of a religious doctrine over the more private aspects of human experience. If, on the other hand, the arts "interact" with religion, if artists can have an influence on the development of religious orientations (instead of merely being "guided" by them), religion might acquire some of the characteristics of an art—become less systematized, more private, less dogmatic in its claims, more individualizing in the experiences it permits. The medieval Catholic Church, by and large, "used" the arts, while the Asian religions—other than Confucianism—tended to "interact" with the arts. In classical Greece, the arts became so emancipated from religion that they can be said to have "used" it, with more enduring benefits for the arts than for religion.

Separation of religion from art. A sharp separation of religion from the arts, such as tended to occur during the Protestant Reformation, promotes the rise of a rationalized, "disenchanted" religious world view and an anarchically romantic tradition of the arts. Like the Reformers, religions generally have been willing to accept some kinds of music and of literature and have varied mainly in their attitudes toward the visual and the "bodily" arts—the dance and the theatre. Attitudes toward the religious significance of the materially existent underlie this variation. But religions have also varied in the degree to which they made use of, or interacted with, the arts. "Aesthetically deprived" religions lose one sort of appeal to potential adherents and, perhaps especially in times of widespread discontent and cultural crisis, do not compete well for popular support with religions that possess more powerful aesthetic (or mythological) resources. Soviet explanations of the survival of religion in the U.S.S.R. stress increasingly its aesthetic aspect —a problem for Marxist ideologists, who are beginning to perceive a lack of this element in their own system of faith.

Effect of separating art from religion

A religion or a secular ideology without the arts loses one type of symbolic resource for its perpetuation and a possible source of modification of its doctrine. In addition, the arts represent one possible way of testing human significance of various aspects of the message of a religion or secular ideology. Do they survive, when artistically treated, without the support of enforcement by the established authorities of a church (or party)? It has been suggested that some ideological and religious concepts failed to be retained in the popular consciousness because they never received an effective artistic elaboration.

For the arts, a complete separation from religion or secular ideology means, on the one hand, a release from the obligation to serve a tradition regarded as superior to art itself and from its various efforts at social control —including censorship—over artistic expression. On the other hand, the separation of art from religion or ideology means: (1) the loss of a symbolic resource that either can be used directly in art creation or that stimulates artistic creativity indirectly by "disturbing" the imagination of artists; (2) the loss of one type of opportunity to create art that is both appreciatively used by large numbers of people and regarded as significant by them; and (3) the loss of a type of patronage that has historically been more continuously interested in the arts (other than literature), in times good and bad, than any other and that has tended to exhibit more concern even for the aesthetic merits of art than did the two most important types of patrons—the high bourgeoisie and the secular state—that followed the virtual demise of church patronage.

A preoccupation with patronage, however, has not always governed the behaviour of artists. Even though painters could expect to lose their most important kind of patronage if the iconoclastic Protestant Reformation succeeded, a surprising number of them supported the Reformation. Historically, literature has been least dependent on religious patronage; music (because of its almost ubiquitous linkage with ritual), perhaps most.

Attitudes toward patronage

AESTHETIC INFLUENCES

Influence of kinds of religion. A strong influence of doctrinal religion over the artistic enterprise disposes it toward a "clerical" conception of its task: upholding of a cultural tradition regarded as higher than art itself. It also tends to eliminate women from roles, except auxiliary, in the artistic enterprise—as in organized religion. But the *mystical* streak of religious experience has frequently influenced the artistic enterprise in a generally liberating direction, producing an ecstatic type of artistic culture. Any kind of direct religious influence on the arts tends to give them a more consciously symbolic character, but the doctrinal religions tend to impose on them an authoritarian rigidity (for example, the Romanesque style current in Europe from the 11th to the 12th centuries), whereas the mystical religions produce tendencies toward a more fluid style that could be interpreted as more egalitarian (Buddhist wall paintings at Ajantā in Mahārāshtra State, India). Mystical religions may also open the arts more equally to both sexes, as it did for such women mystic writers of the late Middle Ages as the English Julian of Norwich and the Italian St. Catherine of Siena.

Feeling-oriented religions like Buddhism tend to produce preferences, in the personalities influenced by such religions, for sensuous styles of art; belief-oriented religions like Calvinism favour austerity in art styles.

Confucianism has also tended to favour austerity, but the Chinese Confucian gentlemen painters drew most of their artistic inspiration from Taoism and variants of Buddhism. Perceptual tendencies may persist even after the religious tradition that has shaped them is no longer consciously adhered to.

Changes in the religious system have, in strongly religious periods, provided impulses for subsequent artistic changes. Periods of increased artistic creativity have frequently followed those of intense religious struggle—conversions of nations, great heretical movements, successes or failures of popular reformations, even iconoclasm. But this is not an effect peculiar to religious changes: other types of intense social action have also preceded artistic efflorescences. A more specifically religious influence on artistic creativity is suggested in the historic tendency for artistic efflorescences to occur after periods of great religious creativity. This suggests that the arts benefit from a partial secularization of intensely religious traditions or that religious efflorescences shape symbolic or emotional resources that are most usable for artistic expression when they have aged ("mellowed") somewhat but have not been completely discarded from the living experience of a people. Expirations of previously potent religious traditions, as of Buddhism in India, have coincided with declines in artistic creativity.

The modern secular equivalents of religion have generally been quite inferior to it in their effects on the arts. This may be partly because most of the "secular religions" of 19th-century origin have been rationalistic, pur-

Inferior influence of secular "religions"

posive ideologies with little mythological content to disturb and stimulate the artistic imagination. In fact, artists have been most affected by the drama of events (or "theatre") brought into being by secular religions, rather than by their "mythology."

A great many modern artists have been willing to let themselves be influenced by the secular ideologies or by the more ancient religious traditions of their own civilization or by religions of cultures alien and exotic to them. The latter two influences have generally been aesthetically more auspicious than the former, but most artists do not experience even these influences with sufficient intensity to be "disturbed" by them and tend to exploit, almost at random, their more superficial characteristics.

Art's shaping role in religion. It would be difficult to demonstrate that developments in the sphere of art have had any independent effects on religion. Throughout much of history, art has not been sufficiently independent from religion to be perceived as a "cause" of religious developments. More likely, it may have given focus, a tangible concreteness, a dramatic shape, a memorable melody, to abstract religious notions or shapeless feeling states and provided means for celebrating and transmitting them—rather than initiating the experiences and the notions themselves.

It seems likely that the cultural attitudes of the growing number of alienated intelligentsia of the West are greatly affected by its artistic culture. The conception of the "revolution" as espoused by the New Left has, in many ways, been a "surrealistic" one ("all power to the imagination," etc.). Fascism, too, has been considered by some observers to be an aesthetic phenomenon almost as much as a political one. It is, in any case, through the shaping of fantasy dispositions—congenial modes of perception—that art can influence the development of general value orientations, which in turn dispose people to favour particular religious or political ideologies, when choice between alternatives is possible.

The impact of fantasy

Yet the shaping of fantasy dispositions is not a foolproof method of manipulation, particularly in modern societies: it may provoke a subjective rejection of overly insistent attempts to influence—an anarchic response to the rigid authoritarianism of the "classical" styles or, conversely, a demand for new dogmas and rituals in response to more ambiguity, suggested by the arts, than an individual can live with in his own life.

VI. Technology, science, and the arts

INFLUENCE OF TECHNOLOGY ON ART

To varying degrees, the arts depend on technological evolution for the very techniques used to create works of art —music considerably, literature least, and architecture most of all. The cinema has been made possible only by recent technological developments.

The arts also depend on technology for the dissemination of the creative product. Book printing has made possible the development of a mass reading public, which in turn has facilitated the rise of new literary genres, such as the novel. Modern techniques rendered objects of visual art mechanically reproducible, hence perceived and treated as less "unique" than they had been in the past. The film is a peculiar art in that the technique of production and the technique of dissemination immediately imply each other, both having been produced by the same technological development.

Beyond these direct technological influences on the arts, there is an indirect one, mediated through the effects that particular technologies have on the imagination of artists or of larger populations.

Effect on content. Generally, the arts do not mirror, in any direct manner, the technological developments occurring in the society in which they have been produced. In their contents, the visual arts are more likely to touch upon the consequences of technological developments that have become intimately familiar (for example, the artifacts represented in still-life painting) but are not necessarily its objectively most important result; or perhaps upon the more vividly visible agents of technological development (such as draft animals, machines, or electronic processes) rather than on the whole system of technology or the basic principles underlying it.

In other words, the arts focus on those aspects of technology that "grasp the senses." Thus, they frequently dwell on technological innovations in a very early stage of their development, then let them drop as subjects for artistic concern after they have become fully developed, attained a dominant role in the economy, and are taken for granted emotionally.

But it is also what technological change has eliminated or does not permit to exist that can grasp the artists' imagination; they will then be concerned with what they miss in a particular technological system and will depict, perhaps, the opposite of what they see as existing in it. It has been found that in preliterate societies where the house shape is circular, straight lines are preferred in art style and that, conversely, where the house type is rectangular, art styles tend toward the curved line. This finding suggests either that art supplies what is most lacking in the technological environment or that one art must provide what another art does not.

Art as a celebration of the absent

Technological developments may both suggest new contents for art and encourage the elimination of old ones. Thus, the invention of photography has virtually eliminated the need for realistic portraiture.

Effect on style. In the development of styles of art, technological factors appear to be more important than they are in the choice of subject matter. Certain styles have been made possible only by particular technological developments (electronic music). In other cases, technological developments (new types of paints) have converged with new scientific theories (in optics) and perhaps psychological changes (the decline of middle class democratic militance) to produce a particular style, such as Impressionism in French painting.

In still other cases, a style may be regarded as a symbolic projection of psychological attitudes inherently linked with a technological process. It has been argued that tendencies toward more geometric styles of visual art have emerged in connection with both of the major technological transformations of society—the agricultural and the industrial revolutions, while periods that have preceded these transformations have favoured more "realistic" styles (*e.g.*, Paleolithic cave paintings and the figurative arts of urban pre-industrial civilizations). A possible explanation of the linkage of the technological revolutions with more abstract styles of art may be the alienation from nature or a sense of mastery over it suggested by the geometric styles. Through them man imposes "his own" type of order on nature, where no such geometricism can be found. This attitude is also implied by the great technological transformations through which man has objectively increased his control over the forces of nature.

This process could have started in the imagination of artists inventing the notion of mastery over nature as a dream; in the attitudes of the people or of the religious, economic, or political leaders, reflected in both technological and artistic changes; or in the process of technological transformation, which, by its success, led to a widespread sense of mastery, projected into artistic style.

Industrial mastery and style

In the industrial transition the apparent sequence has been change in popular or elite attitudes, encouraged by the Protestant Reformation; speeding up of technological changes; and the rise of geometric styles of visual art. Music has "lagged behind" the technological transition even more than have the visual arts, but literature may, in certain respects, have anticipated it.

Thus, it could be argued that the poetry of courtly love and of religious affective mysticism of the later Middle Ages suggested an attitude of mastery over nature (over sexual impulses, limits of time and space). Even if this interpretation were accepted, however, it could not be concluded that literature is necessarily a more sensitive indicator of underlying psychological changes than the other arts. It may be in periods in which the other arts are heavily dependent upon organized patronage and a guild organization, and literature is comparatively free of such encumbrances, that it registers psychological changes

more sensitively and at an earlier date. Technology, in this perspective, appears not as the "ultimate" determinant of artistic expression but as an aspect of basic psychological and cultural change—an aspect that became crucial at a certain point.

The interest of artists in science appears to have been greatest on the threshold of the Industrial Revolution, in the 17th century, and toward its end—when technological developments, particularly in the field of electronics, have altered the material basis of advanced societies in the second half of the 20th century. In between, artists, especially writers, tended to be repelled by technology and its effects on people and the environment (and by the science they associated with these effects).

In the latter part of the 20th century, the attitude of repulsion still tends to hold with respect to the machine technology, but a distinction is made between it and the electronic (and cybernetic) technology, which is viewed more optimistically. Electronic technology, by the miracles of its circuitry that exhibits an almost human responsiveness, is expected to overcome the chasm between the "mechanical" and the "spiritual" that the Industrial Revolution had deepened.

Changes in the character of science during the 20th century, its recognition of the principles of relativity and indeterminacy and of the importance of the researcher's subjectivity, have contributed much to the revived appeal of science to artists and to their high expectations with respect to the technology based on this kind of science. In the 20th century, psychoanalysis and several perspectives in the social sciences have also been influential in the arts.

The response of the arts to science

In the long run, it is likely that the artists' response to the sciences and technologies of advanced industrial societies will follow the usual logic of the workings of the artistic imagination. That is, scientific technology should influence the arts through the artists' choices between affirming it and modelling their own work after it or repudiating it and concentrating on what it cannot recognize or suppresses or, if left to itself, would tend to destroy. The power, mathematical clarity, and systematic nature of the sciences and technologies may produce an artistic response in the form of sensuous, amorphous styles, carrying suggestions of impotence, of mysticism, of unique experiences.

There is no reason to anticipate that, in the long run, artists will be overwhelmed by science and assimilated into its mode of operation. Indeed, the more science and technology develop, the more they may be subjectively taken for granted, the more imagination may be captured by what science and technology cannot encompass, and the more an art that has not been overwhelmed will be needed. Artists tend to be overwhelmed by a great increase in the technical possibilities of expression only when they have no reason of their own—a mode of perception, a value commitment—to express anything in particular.

Effect on creativity. An essentially descriptive science may influence artistic creativity favourably when it fuses with an older tradition of an artistic craft, as anatomical and optical research vitalized painting in the Italian Renaissance. But a highly abstract science, removed in its formulations from directly perceivable realities, seems to be of most benefit to artists when they permit their imaginations to be stimulated by the general atmosphere of scientific discovery, of the revision of concepts about the nature of reality, or of new perceptions, rather than when artists have set out to apply scientific principles consciously to their work. Even an incorrect interpretation of scientific models by artists can be artistically productive (and perhaps scientifically suggestive).

OTHER ASPECTS OF THE RELATIONSHIP

Effect of art on the social sciences

To what extent the arts can influence the sciences is still uncertain. If "visual thinking" precedes conceptual thought, or if there is a basic background element common to both visual and conceptual thought, changes in perceptions that art either reflects or stimulates may generate responses in conceptual theorizing and, in this manner, affect developments in science as well as in philosophy. But it is on social and psychological theory and on the perceived shape of history that developments in the arts may be expected to have the strongest effect.

The association of artists with technology, especially in the visual arts and architecture, has been far closer and more durable than it has been with science, and the requirements and achievements of artists have often led to technological discoveries. Literary artists, on the other hand, have, particularly in the periods of great scientific discovery, such as the 17th century, been influenced more by science than by technology.

It might seem that the new technologies of the mass media of communication, by permitting more pervasive dissemination of particular works of art throughout society, are likely to increase the impact of art on people's personalities and modes of existence. But the mass media may also have the opposite effect: by making of art a commonplace occurrence that is taken for granted but does not generate an enduring emotional response, by assimilating it more completely to "entertainment," the mass media may decrease the effects of works of art on people and, indeed, on the artists themselves and on their aesthetic experiences.

At the same time, the media may be increasing the psychological impact of nonartistic events, particularly those that resist being assimilated to entertainment, such as wars in a foreign country, which were easier to take for granted before the electronic media became fully developed. This has led to tendencies to substitute, especially in the visual arts, the structures of events or of social systems for aesthetic structures (for example, an art exhibition consisting of photographs of slum buildings and landlords' names).

While it is debatable whether the mass media have enhanced the impact of art, they have certainly increased the social status of the performing artists (and tend to transform all artists whom they capture into performers).

VII. Aesthetic education

The broad and narrow views of aesthetic education

In the broad sense, aesthetic education refers to everything that art is, or may be, used for in the education of nonartists. In the narrower sense, aesthetic education is the developing of a sensitivity to aesthetic qualities and works of art and of an understanding of the criteria by which some works of art are regarded, by artists and art critics, as possessing more highly valued aesthetic qualities than others. Aesthetic education could hardly avoid, but it does not need to be limited to, what is implied in the term's narrower meaning. Approaches to aesthetic education vary mainly in how they conceive of its "broader" responsibilities.

As soon as art is created and exposed to others, it always educates, whether effectively or not, in some way. A study could well be made of the implicit philosophies of aesthetic education, inherent in the manner in which art is used and particularly in the manner in which the growing individual is exposed to it, of the preliterate societies in which the role of an art educator—as distinguished from the practicing artists—does not exist.

BASIC CONCEPTIONS

The formal development of aesthetic education, like the formalization of all education, has occurred in the classical civilizations. In the Western tradition, four basic conceptions of aesthetic education for the nonartist have been sociologically most important, although in practice they frequently overlap.

Didactic. The didactic theory regards art as a means for shaping a particular type of personality (the view of the Greek philosopher Plato). Modern notions of proletarian, black, or feminine "consciousness raising" by means of art, insofar as they presuppose that art is to be used for developing an attitude specified in advance, represent variants of the didactic approach to aesthetic education. But even when the goal is developing a particular type of aesthetic taste (as, for example, a taste for the "classical" styles or for "modern" art), a philosophy of education is didactic, albeit in a more subtle way. The didactic theory implies that if art does not do the task as-

signed to it, it has either no place in education or a subordinate one; and it may lead to justifying censorship of the arts, "to protect the innocent from corruption." This view of aesthetic education is most congruent with the "clerical" conception of art.

Therapeutic. The therapeutic theory views art as the supplier, for individuals or groups, of experiences that everyday life in society fails to provide but that are assumed to be necessary for a "whole" and "healthy" existence. In a highly rationalized society, art supplies or reveals the irrational, supplementing or confronting reason and duty with spontaneity and sensuousness. Or, in a more sophisticated conception, art reconnects reason with sensuousness, which had first to be separated from each other, for man's self-awareness to advance itself.

Since this philosophy of art as therapy developed in an early phase of the Industrial Revolution, it usually conceives the deficiencies of society in terms of what was becoming suppressed then—needs for emotional expression —or what was promised ideologically but has not been adequately delivered in reality even in the most modernized societies, such as equality and participation. But what appear to be increasingly missed in the advanced industrial societies are credible designs for the symbolic coherence of life; a sense for experiences and objects that resist being used up and are immune to planned obsolescence. The therapeutic function of aesthetic education may be changing accordingly.

Changes in the therapeutic function since the Industrial Revolution

The objective of a "therapeutic" aesthetic education is to use art to bring out the missing elements and either to promote their integration with the existing state of affairs or to overthrow the latter. The task may be conceived as "sociotherapeutic" or "psychotherapeutic"—overcoming the deficiencies of a whole civilization or those of an individual personality. In contrast to the didactic approach, only the initial problem and the general direction of the effort, but not the final solution—a specified type of personality or society—are presupposed. Artworks can be judged, in a general way, by the degree to which they fulfill the therapeutic task set to art by the character of the society or the stage of the evolution of civilization contemporaneous with it. But no artwork can be in principle excluded from aesthetic education, since any work may be "therapeutic" for someone.

Developmental. The developmental theory aims at making it possible for anyone to choose from the realm of art whatever he needs to develop his unique potentialities. The evolution of society has not given rise to any particular type of need that art should be meeting in a given social context. The task of aesthetic education is, perhaps, to increase the range of exposure to art so that every individual will choose on the basis of a more complete knowledge, more intelligently, how to develop his own self aesthetically.

This is a liberating educational philosophy, particularly in a tradition dominated by didactic approaches. But it has three unintended effects: (1) Since it puts the emphasis on what a person gets from a work of art for himself, it destroys the reason for trying to understand the work of art in itself and the ways it has functioned for other people, including its author. (2) While this approach permits individual experiences and descriptions of such experiences, it eliminates the basis for a more generalized critical evaluation of works of art. (3) By its emphasis solely on the individual's aesthetic experience, it eliminates the need for art. What have the works of art got that "aesthetic experiences" on a crowded street or in one's dreamworld do not give?

The developmental approach applied to children

The developmental approach is most applicable to the aesthetic education of children, where it is provided with a sense of direction by the psychologists' notion of the stages of intellectual and perceptual development of the personality, each offering distinctive possibilities and limits to aesthetic education. In the aesthetic education of adults, however, by implying that neither art nor the society nor history are particularly important as compared with a person's "experiences," this approach defines itself as a luxury object, likely to appeal to a social group that is prosperous but impotent to shape its destiny; and it needs to be balanced by various other kinds of aesthetic education.

Culture-critical education. The purpose of another form of aesthetic education is criticism of culture, in the broad, anthropological sense, through analysis of concrete works of art and of their functions in the life histories of individuals and in the historical existence of societies (and particular groups within them). This culture-critical approach differs from the therapeutic in that it is based less on the personal feelings and cast of mind of the educator and anchored more in the empirical study of culture history and of people's relationships to the symbolic expression of their experiences. It considers art in the context of all human experience, not just in the narrow social setting of an individual perceiver of art (as the developmental theory necessarily does). It is not concerned with recapitulating human experiences in the technical detail of historical monographs. Instead, it is involved with the recognition of patterns in which experiences are "fitted together"—externally influenced, intuitively structured, and their qualities and meanings interpreted to affect later experiences.

Aesthetic education as culture criticism is concerned with producing an ability to make aesthetic judgments that are founded on the knowledge of the ways in which whole civilizations, as well as everything involved with them, have been working. An analytical understanding of parts and of particular relationships is encouraged to grow into a capacity to perceive and evaluate the overall connectedness of a civilization.

Aesthetic education as culture analysis is a method that can be adequately used only with adults; but it can be placed at the very centre of the education of adults, including college students. Its purpose is not to produce art critics but, rather, persons more competent of judging civilizations, including their own, on the basis of an increasing understanding of what they have done, or failed to do, to human sensibilities.

SUPPLEMENTARY APPROACHES

The culture-critical approach to aesthetic education is predominantly analytical. It needs to be supplemented, for most people, by practice in the making of works of art. The writing and performing of dramatic works permit an exploration of the potentialities of social interaction and its limits, which a person has either had no experience with in his own life and yet senses as potentially significant or which he has experienced in an "incomplete" manner and whose full logic and his own design for interpreting it he wishes to work out. Other kinds of literature seem to have a similar role in aesthetic education. Music and dance clarify the dynamic rhythms in terms of which personal experiences and the character of civilizations are perceived, while the visual arts crystallize images of the emotional qualities these experiences and civilizations are sensed to possess. The film provides a potential basis for integrating the functions that literature, music, and the visual arts have in aesthetic education, but the separate experience of these arts provides a clearer understanding of such functions.

Architecture, with its commitment to relating dispositions of the imagination to the practical exigencies of life —and therefore to social policy—in a viable overall design of aesthetic merit, deserves, perhaps, a central place in aesthetic education that has not been recognized in any large-scale educational system. It may have a particularly important role in the education of those who refuse to recognize the linkages of their imaginations with the requirements of actuality.

The importance of aesthetic education beyond purely artistic considerations

It would appear to be arbitrary and self-defeating to limit aesthetic education entirely to considering the work of art in its isolation. It also must give some attention to the ways in which societies organize, or fail to organize, themselves to build aesthetically adequate whole environments (physical, social, and even "spiritual"). It must demonstrate how individuals can demand and, by their own political actions and spending patterns, support the building of such environments. A complete program of aesthetic education should include a consideration of the

costs of building and tearing down (or maintaining) aesthetically inadequate environments as compared with the costs of building environments capable of adequately performing a great many cultural functions for the people inhabiting them. And it must teach the techniques of social action for insisting effectively that the kind of aesthetic environment needed is provided.

The importance of immediate contact with practicing artists in the aesthetic education of nonartists is not primarily in the teaching of the technical skills of making art, which the practicing artists are not necessarily superior to "art educators" in transmitting. What good practicing artists can provide for aesthetic education is the demonstration of how "aesthetic structures" (meaningful connections between disparate elements of experience) arise from skilled labour under the constant judgment of a personal sensibility. Good artists also reveal in workaday practice how the nature of a subject, a design, a genre evolves in the process of elaborating its implications and simultaneously develops intrinsic requirements of its own, a set of "norms," a "logic," that the artist cannot disregard without diminishing his creative attainment. In this way, practicing artists transmit a sobering sense for what is required of creators by the inner logic of their works.

The role of the art critic in aesthetic education is to clarify the criteria by which he judges the aesthetic fitness of works of art and distinguishes between artistic successes and failures. The critic should be distinguished from the art interpreter, who explicates the meanings that a work of art has for him. The aesthetic philosopher's most productive role in aesthetic education would be to compare, in some systematic manner, the criteria of the critics, the meanings of the interpreters, and the intentions and the practices of the artists of various societies. In practice, the aesthetic philosophers mainly study each other. Partly for this reason, a role in aesthetic education has been opening up for the sociologists and psychologists of art, who are concerned with how art actually functions in the life of societies and personalities.

Since different individuals are likely to benefit most from different aspects of aesthetic education, it would seem to be a mistake to have a single model of aesthetic education for everyone.

VIII. Preservation and dissemination of art

THE NATURE OF ART PRESERVATION

Art preservation in preliterate societies

Even in preliterate societies, not all art is created for the occasion of its use and then abandoned. First of all, the basic design of the work of art survives in the collective memory of the tribe, or of its more artistically inclined members, and can be reproduced from this recollection, frequently with creative modifications. This principle of preservation operates most clearly in oral literature, but it gives continuity also to the other arts, both performing and objectifying. And it is not limited to preliterate societies.

Art objects may also be preserved either for their utility or for their religious significance. Cave paintings have been continuously "refreshed"—that is, restored by overpainting—over long periods of time by the Australian Aborigines to retain the benefit of their magical effectiveness.

Even before the rise of literate civilizations, art collecting had become a symbolic exhibition of wealth and power. Forms of writing, which made easier the collection and preservation of literature, probably also were developed, in the classical civilizations, for their utility in economic management, the pursuit of religiously significant activities such as astronomy, and the more effective exercise of political power. But once traditions of collecting works of art and of literacy have evolved, they have tended to acquire a degree of autonomy from the purposes they may originally have been associated with. Systems of musical notation have evolved later than literacy and without any significant economic or political motives to necessitate their development; it was encouraged, however, by a religious need to stabilize the liturgical uses of music. Systems of dance notation have been produced for purely artistic purposes. Systems of photographic, sound, and motion recording have been provided by the progress of modern technology. The development of these systems has been propelled more by curiosity than by the anticipation of the great profits they eventually proved capable of producing.

Collections of manuscripts existed in the ancient world, and a system of state and school libraries was established in Rome. But a network of public libraries has evolved only since the middle of the 19th century. The notion of a collection of books has not been as closely associated with the aristocracy as that of a collection of works of visual art: the chief connection of libraries has been with scholarship and its practical applications (in preaching, in administration, and, in modern times, in self-advancement through education).

Art collecting in the industrial age

The nature of art collecting has also changed in the age of industrialization and the democratic revolutions that started in the second half of the 18th century. Art collections, which previously had been possessions of the monarchy, privileged classes, and the church and (except for the visible part of church art) were rarely opened to the public, became, in most cases, public museums.

The traditional association of high art and art collecting with class privilege has led some of the revolutionaries and avant-gardists of the 19th and 20th centuries to conclude that the establishment of a democratic culture requires the destruction of the monuments to an aristocratic culture collected in the museums—and of the museums themselves as repositories of that type of culture.

Not only art museums but also symphony orchestras and even the theatre (in contrast to Elizabethan times, 1568–1603) have, in Western societies, little attraction for the working and lower middle classes. A study made in France in the 1960s found that 1 percent of museum attenders were agricultural labourers, 4 percent industrial workers, 5 percent artisans and tradesmen, the rest white-collar workers and higher social classes. In eastern Europe, where there is more of a tradition of high culture being in alliance or, indeed, in secret emotional conspiracy with the people (an attitude of the folk culture that has been retained), working class attendance at performances or exhibitions of high culture is higher. But even though it is encouraged by government, party, and trade-union agencies (and indirectly by the monotony of much of the officially sponsored "popular" culture), interest in high culture is still stratified by class.

This lack of interest by manual labourers in art poses a contradiction. Works of art, by whomever they have been sponsored or collected, have always been produced by craftsmen who surpassed mere craftsmanship. They represent a conjunction of craftsmanship and sensibility of the men and women who have been in the vital centre of their own times, working and imagining. If there is a monument to the immortality of manual labour, it is a museum of art.

Museums of modern art

Museums devoted to contemporary art have developed only in the 20th century. Instead of preserving what has survived repeated tests of critical judgment over the ages, museums of contemporary art delve into the flux of ongoing artistic developments, at best endeavouring to sort them out into intelligible patterns and to enlighten the public's consciousness of its own times. Frequently, however, such museums have become powerful trend setters for the fashionable, producing an ephemeral new "movement" each year and becoming, in effect, adjuncts of the mass media of communication rather than seekers for the surviving values of art. In the 1960s, there developed demands that contemporary-art museums be conceived of as "houses of controversy"—where anything arousing concern of contemporary artists could be exhibited, whether it is art in the traditional sense or not. There are tendencies toward differentiation between two types of art museums: one engaged in controversy and the other a repository of art-historical collections whose holdings are critically evaluated for their aesthetic merit. There also seems to be an increasing need for museums in which all the arts of a period or a historic group could be shown together, to reveal the overall cultural atmosphere of the

period or group and the interconnections of its arts, placed within their sociological context.

To preserve artworks implies the desire not to deprive future generations of the kinds of human experiences that the circumstances of the past and present have generated and the future may no longer be able to produce. Possession of what one is not able to produce stretches the mind in ways that, in the absence of these possessions, could not even be imagined. By not preserving works of art, a society would lose much of its awareness of the limits of its own imagination and would tend to treat its present modes of existence and of aspiration as though they were absolute.

Unselective preservation of everything that has claimed to be art would, however, very quickly overload the depositories of culture. A decision to preserve therefore implies a criterion of selection to determine what is worth preserving. Such criteria may be the aesthetic merit of the work of art or its historical significance for a nation, a church, a political movement, or a family lineage. Preserving samples of the "popular culture"—presumably lacking in aesthetic merit but representative of the character of a period—is also justifiable. Since judgments of aesthetic merit are revised and possibly improved over long historical periods, "accidental" preservation of works that have not been judged by their contemporaries to possess aesthetic merit can also become important.

The greater part of the art produced by the preliterate societies and historic civilizations has been lost due to neglect and historical misfortune. Only 12 percent of the tragedies said to have been written by Aeschylus, Sophocles, and Euripides in Greece during the 5th century BC have survived. Of the period that the Chinese traditionally regard as representing the historical peak of their painting—the T'ang (AD 618–907)—hardly any paintings have been preserved. Many works of art have been intentionally destroyed by organizations or individuals (for example, medieval peasants burning classical marble statues for lime to fertilize their fields). Among the major social organizations, the churches have probably done more for the preservation of works of art than any other, but they have also destroyed art in religious wars and campaigns for the eradication of heresy. States have both protected and destroyed. Business organizations, on the other hand, rank among the major destroyers of architectural monuments, particularly in contemporary societies. Only by governmental regulation can the popular market, whether capitalist or Socialist, be restrained from destroying unique works of art to build parking lots.

Destruction of art

SYSTEMS OF DISSEMINATION

At least seven types of systems through which artistic products are disseminated to art consumers may be identified. These systems differ in the manner in which contact with art is established and the attitude with which it is approached.

1. Ritualistic systems. Art is incorporated into the conduct of special, repetitive occasions and presented as an integral part of them (for example, church services and political rituals in which particular kinds of music are performed). Highly specific types of art receive wide exposure on an occasional basis. An air of the extraordinary, of festivity, is attached to these types of art. The modern opera has inherited some of the trappings of the ritualistic system. Certain avant-garde occasions may approach it. So do art festivals, by breaking with the routine of permanent museum collections or environments, which gives art the character of being taken for granted.

2. Environmental systems. Art is incorporated into the organization of the everyday, stable environment visible to everyone in the normal course of his life. Architectural works, landscaping, public monuments, and private libraries are cases in point. Transmitted in this way, art influences the perceptual expectations of most everyone, but it tends to be taken for granted, to become "invisible." In modern societies, economic organizations and governmental agencies are primarily responsible for the dissemination of art (bad as well as good) through this system. Governments could, indeed, do more for aesthetic education by controlling the environmental system than through their hold over aesthetic education in the schools.

3. The utilitarian system. Art is incorporated into the small, usually portable objects of everyday use, from eating utensils to automobile designs. In its effects, this system is comparable to the environmental system. But it is industry, rather than the government, that is in a strategic position to affect art's distribution through the utilitarian system.

Art in everyday utilitarian objects

4. The art-trade system. Art objects are produced for and sold on the market specialized for goods of their nature. Depending on their cost, they may appeal to a "mass" or a "class" public but, in any case to those possessed of a purchasing power they can devote to objects of no material utility and who have a tradition of buying such objects. In practice, the system of trade in the objects of fine art tends most frequently to become oriented to economically privileged groups. The market for music records and tapes is, however, creating a tradition of buying art objects even among the underprivileged. The book trade, particularly after the development of the paperback in 19th-century Germany, is oriented to a broader section of the population than is the art trade; it tends to divide into an elite-oriented and a popular component, but the two overlap.

5. The mass-entertainment system. Artworks are constantly exposed, at low cost, to large audiences, but not as a stable and integral part of their everyday environment (as in the environmental and utilitarian systems) nor with an attitude of the extraordinary attaching to it (as in the ritualistic system). The ubiquity of exposure, the fleetingness of the occasions of exposure, and the general atmosphere of fun and games surrounding the operations of this system have the effect of obliterating the unique significance of any particular object of art. Everything processed through the mass-entertainment system tends to become like everything else.

6. The educational system. This is the concern not only of schools but also of modern museums and libraries. Works of art are collected and disseminated with an educational intent—in schools potentially to everyone in the right age groups, in museums and libraries to those who choose to make use of the education offered. Museums, however, have not been fully assimilated to the educational system. In the United States, they are still not regarded, in their claims for public support (or for an automatic tax exemption by virtue of their status), as quite equivalent to schools. There are also notable tendencies in the latter half of the 20th century to incorporate museums into the mass-entertainment system. On the other hand, a politically controlled system of "mass entertainment," as in totalitarian states, is in fact heavily "educational," in a didactic manner. The street theatre of the "counterculture" is also intended to be educational.

The educational system typically lacks the power inherent in environmental or utilitarian systems and the intensity characteristic of ritualistic systems. Possibly for these reasons, the art disseminated through the educational system tends to acquire connotations of what in the upper middle class of Western societies used to be perceived as "femininity"—a refined irrelevance to the world of affairs. The possession of "real life" power by the large art museums constitutes one of their educational advantages over art teaching in the schools, where the art teacher is still frequently treated as an inferior type of educator.

7. Movement systems. There are indications that a new system of dissemination of art may be emerging in the youth movement and perhaps other popular movements of a basically "cultural" (rather than traditional "political" or "economic") character that have become a feature of advanced industrial societies. Any particular movement system is temporary, but, while it lasts, it generates an intense audience response to certain kinds of art, and the audience may even be drawn into the collective production of works of art. The dividing line between producer and consumer of artworks becomes attenuated. Art can be consumed without waiting for a set

occasion, but its consumption always has an aura of rebellious celebration attaching to it. If movement systems are successful, however, they tend to be assimilated into mass-entertainment systems—revitalizing them, but losing their own vitality.

While all known societies produce (and therefore presumably need) some kind of art, not all their individual members use art or exhibit any kind of response to it, even when it is offered to them at no cost and in an environment that is not intimidating. Therefore, all societies have some system, or a combination of systems, for the dissemination of works of art, but not all of their members are involved with such systems. While art-dissemination systems might aim at a universal exposure to works of art, they cannot be blamed for not generating a universal response to art.

The sources of the responsiveness to art lie partly outside of the whole artistic enterprise, in the structures of individual personalities and in their experiences in confronting social systems and historical processes. In this sense, an individual's contact with art starts not with society's systems of art dissemination but with the character of his existence.

BIBLIOGRAPHY

Anthologies: On the visual arts in preliterate societies, see CAROL F. JOPLING (ed.), *Art and Aesthetics in Primitive Societies: A Critical Anthology* (1971), an important reader. On the arts in historic societies, MILTON C. ALBRECHT, JAMES H. BARNETT, and MASON GRIFF (eds.), *The Sociology of Art and Literature* (1970), has very extensive coverage and bibliographies. The best overview of the sociology of literature is LEO LOWENTHAL, "Literature and Sociology," in JAMES THORPE (ed.), *Relations of Literary Study: Essays on Interdisciplinary Contributions*, pp. 89–110 (1967).

Art and society: ROBERT ESCARPIT, *Sociology of Literature* (1970); I.C. JARVIE, *Movies and Society* (British title, *Towards a Sociology of the Cinema*, 1970); and ALPHONS SILBERMANN, *Wovon lebt die Musik?* (1957; Eng. trans., *The Sociology of Music*, 1963), provide overall analyses of the social organization of particular arts. CESAR GRANA, *Bohemian versus Bourgeois: French Society and the French Man of Letters in the Nineteenth Century* (1964); FRANCIS HASKELL, *Patrons and Painters: A Study in the Relations Between Italian Art and Society in the Age of the Baroque* (1963); BARRINGTON KAYE, *The Development of the Architectural Profession in Britain: A Sociological Study* (1960); HARRISON C. and CYNTHIA A. WHITE, *Canvases and Careers: Institutional Change in the French Painting World* (1965), are studies of changes in the social role of particular types of artists. An account of a social setting in the performing arts (the Broadway theatre) is SAMUEL W. LITTLE and ARTHUR CANTOR, *The Playmakers* (1971).

Basic works on social elements in the style and content of the arts: WALTER ABELL, *The Collective Dream in Art: A Psycho-Historical Theory of Culture Based on Relations Between the Arts, Psychology and the Social Sciences* (1957); PIERRE FRANCASTEL, *La Réalité figurative: éléments structurels de sociologie de l'art* (1965); ARNOLD HAUSER, *Sozialgeschichte der Kunst und Literatur*, 2 vol. (1953; Eng. trans. of vol. 1, *The Social History of Art*, 2 vol., 1957); VYTAUTAS KAVOLIS, *Artistic Expression: A Sociological Analysis* (1968); ALAN LOMAX, *Folk Song Style and Culture* (1968); LEO LOWENTHAL, *Literature and the Image of Man: Sociological Studies of the European Drama and Novel, 1600–1900* (1957); RENATO POGGIOLI, *Teoria dell'arte d'avanguardia* (1962; Eng. trans., *The Theory of the Avant-Garde*, 1968); PITIRIM A. SOROKIN, *Social and Cultural Dynamics*, vol. 1, *Fluctuation of Forms of Art* (1937); and MAX WEBER, *The Rational and Social Foundations of Music* ed. by D. MARTINDALE, J. RIEDEL, and G. NEUWIRTH (Eng. trans. 1958). Important studies of particular historic cases are LUCIEN GOLDMANN, *Le Dieu caché: étude sur la vision tragique dans les Pensées de Pascal et dans le théâtre de Racine* (1956; Eng. trans., *The Hidden God: A Study of Tragic Vision in the Pensées of Pascal and the Tragedies of Racine*, 1964); and PHILIP E. SLATER, *The Glory of Hera: Greek Mythology and the Greek Family* (1968).

Social conditions of artistic creativity have been studied in VYTAUTAS KAVOLIS, *History on Art's Side: Social Dynamics in Artistic Efflorescences* (1972); and A.L. KROEBER, *Configurations of Culture Growth* (1944). Influences of art on society are theorized in HUGH DALZIEL DUNCAN, *Communication and Social Order* (1962); and RADHAKAMAL MUKERJEE, *The Social Function of Art* (1948). VICTOR W. TURNER, *The Ritual Process: Structure and Anti-Structure* (1969), provides a framework for analyzing the social effects of the performing arts. For a survey of some empirical findings, see JOSEPH T. KLAPPER, *The Effects of Mass Communication* (1960).

Art and economics: There is no basic overall inquiry into the role of economics in the development of art. Good historical studies of art collecting are FRANCIS HENRY TAYLOR, *The Taste of Angels: A History of Art Collecting from Rameses to Napoleon* (1948); and NEILS VON HOLST, *Künstler, Sammler, Publikum* (1960; Eng. trans., *Creators, Collectors, Connoisseurs: The Anatomy of Artistic Taste from Antiquity to the Present Day*, 1967). For trends in art prices, see GERALD REITLINGER, *The Economics of Taste*, 3 vol. (1961–70); and GERALDINE KEENE, *Money and Art: A Study Based on the Times-Sotheby Index* (1971). On economic support of contemporary arts in the United States, see WILLIAM J. BAUMOL and WILLIAM G. BOWEN, *Performing Arts: The Economic Dilemma* (1966); and WILLIAM JACKSON LORD, *How Authors Make a Living: An Analysis of Free Lance Writers' Incomes, 1953–1957* (1962). In nine west European nations, see FREDERICK DORIAN, *Commitment to Culture: Art Patronage in Europe, Its Significance for America* (1964). On the development and current status of legal rights of artists, see BRUCE W. BUGBEE, *Genesis of American Patent and Copyright Law* (1967); BORIS I. GOROKHOFF, *Publishing in the U.S.S.R.* (1959); GEORGE HAVEN PUTNAM, *Books and Their Makers During the Middle Ages*, vol. 2 (1897); HOWARD WALLS, *The Copyright Handbook of Fine and Applied Arts* (1963).

Art and Politics: There is no reliable survey of relationships between art and politics. Good examples of monographic studies of particular historic cases include: HUGH DALZIEL DUNCAN, *Culture and Democracy: The Struggle for Form in Society and Architecture in Chicago and the Middle West During the Life and Times of Louis H. Sullivan* (1965); DONALD DREW EGBERT, *Social Radicalism and the Arts, Western Europe: A Cultural History from the French Revolution to 1968* (1970); JAMES A. LEITH, *The Idea of Art As Propaganda in France, 1750–1799* (1965); GEORGE LACHMANN MOSSE (ed.), *Nazi Culture: Intellectual, Cultural, and Social Life in the Third Reich* (Eng. trans. 1966); JAMES L. PEACOCK, *Rites of Modernization: Symbolic and Social Aspects of Indonesian Proletarian Drama* (1968); HAROLD SWAYZE, *Political Control of Literature in the USSR, 1946–1959* (1962).

Art and religion: Relationships between religion and the various arts are comprehensively surveyed in GERARDUS VAN DER LEEUW, *Vom Heiligen in der Kunst* (1957; Eng. trans., *Sacred and Profane Beauty: The Holy in Art*, 1963). For sociological studies of religion and the visual arts, see the works of Abell, Francastel, Kavolis, and Sorokin listed above. Of numerous historical studies, G.G. COULTON, *Art and the Reformation*, 2nd ed. (1953); and JEAN GIMPEL, *Les Batisseurs de cathédrales* (1958; Eng. trans., *The Cathedral Builders*, 1961), are concerned with the effects of religion in organizing the arts, while HELEN GARDNER, *Religion and Literature* (1971); and ANDRE MALRAUX, *La Métamorphose des dieux* (1957; Eng. trans., *The Metamorphosis of the Gods* 1964), with the "spirit" of religion in art.

Art, technology, and science: A historical survey is provided in CYRIL STANLEY SMITH, "Art, Technology, and Science: Notes on their Historical Interaction," *Technology and Culture*, 11:493–549 (1970). On the impact of contemporary science and technology on the visual arts, JACK BURNHAM, *Beyond Modern Sculpture: The Effects of Science and Technology on the Sculpture of This Century* (1968), is more imaginative, but less balanced, than C.H. WADDINGTON, *Behind Appearance: A Study of the Relations Between Painting and the Natural Sciences in This Century* (1969). On the influence of science on poetry, see DOUGLAS BUSH, *Science and English Poetry: A Historical Sketch, 1590–1950* (1950); and MARJORIE NICOLSON, *Science and Imagination* (1956). A mythology substituting electronic technology for the spirit of religion has been elaborated in MARSHALL MCLUHAN, *Understanding Media: The Extensions of Man* (1964).

Art and education: The most influential recent prescriptions for aesthetic education have been JOHN DEWEY, *Art As Experience* (1934); and HERBERT EDWARD READ, *Education Through Art* (1943). For psychological studies of perception and imagination in art, see RUDOLF ARNHEIM, *Visual Thinking* (1969); and ANTON EHRENZWEIG, *The Hidden Order of Art: A Study in the Psychology of Artistic Imagination* (1967). The utopian and the analytical dimensions of aesthetic philosophy may be illustrated by HERBERT MARCUSE, *Eros and Civilization: A Philosophical Inquiry into Freud* (1962); and RALPH A. SMITH (ed.), *Aesthetic Concepts and Education* (1970). See also *The Journal of Aesthetic Education* (quarterly).

(V.Ka.)

Arts, Style in the

Definitions of "style"

Like much of the vocabulary of aesthetics, the word "style" resists straightforward definition. The word may point to little more than a mode or form of artistic production; or it can designate traits regarded simply as aids in the task of dating, grouping, and attributing works of art; it can imply skill, grace, or some other sort of excellence; it can mean a manner sanctioned by a standard; it refers to a mode, form, manner, tone, theme, subject, or quality—or a combination of such—that is felt to be characteristic enough to evoke a person, a group, a class, a nation, a place, a period, or a civilization; often also the reference is to features that are said to express an outlook, a doctrine, or a program. As a rule, even in the most carefully controlled context, several meanings will be present; and the tidy meanings will tend to bloom, or decay, into the untidy. Thus, although "sonata style," strictly constructed, should point only to a mode of musical production, in fact it will usually suggest a Classical outlook, the European 18th century, and perhaps the compositions of Haydn. Although to a field archaeologist "Late Helladic III" may designate merely a device for classifying pots of the ancient Greek city of Mycenae, in many imaginations the phrase is apt to inspire a vision of an entire culture, or perhaps of the legendary King Agamemnon bleeding in his bath after being murdered by his wife, as related in Aeschylus' tragedy.

The resulting confusion in thinking and talking about art is often deplored. One art historian likens style to a rainbow, a phenomenon of perception governed by the coincidence of certain physical conditions, which vanishes in the attempt to approach it. Another scholar takes the view that an adequate theory of style awaits a deeper knowledge of the principles of form construction and expression and a unified social theory comprising the practical means of life as well as emotional behaviour.

Such comments have not, however, had much effect; the majority of artists, critics, historians, and ordinary appreciators have continued to employ, loosely but confidently, the familiar term. And this persistence is not indefensible. Rainbows, after all, do exist; the chaser who fails to catch one demonstrates nothing except a mistake about their mode of existence. Also, in talking about art one can always cite the principle, first enunciated by Aristotle, that every study has its own degree of certainty and that a well-educated man will not ask for an unsuitable degree.

With that principle as a point of departure, this article will undertake to expand and combine the several meanings of "style" into a discussion of the inner nature, the varieties, and the dynamics of the phenomenon as it manifests itself in all the arts. (By "dynamics" will be meant the patterns of movement and change that are the concern of style-conscious historians and biographers.) Although the expression of personal opinions will be unavoidable, an effort will be made to present the current state of each question as that state was manifest in the 1970s and to stick closely to accepted usage; the aim will be mostly to discover what the commonly used terms actually mean and only occasionally to suggest what they ought to mean. The whole discussion will be on the level of a general theoretical introduction; material from the history of art and the history of aesthetics will appear only in the form of examples intended to provide clarification and some historical perspective. For full treatment see also RELATED ENTRIES under ARTS, STYLE IN THE, in the *Ready Reference and Index*.

This article is divided into the following sections:

I. The nature of style
  Invention and discovery
    Historical background
    Contemporary thought
  Principal aspects of style
    Value aspects
    Creative aspects
    Formal aspects
    Metaphorical aspects
    Polar aspects
    Measurable aspects
II. The varieties of style
  Single-culture varieties
    Personal styles
    School styles
    Social styles
    Ethnic styles
    Regional and national styles
    Ecological styles
    Religious styles
    Period styles
  Cross-cultural varieties
    Outlook styles
    Contextual styles
    Procedural styles
    Professional styles
III. The dynamics of style
  Historical origins
    Politico-economic factors
    Cultural factors
    Technical factors
    Artistic factors
  Diffusion
    Correspondences in real space
    Correspondences in the arts
  Change and duration
    Cyclical theories
    Dialectical theories
    Sequential theories
    Satiation theories

I. The nature of style

General suppositions

It is easy to suppose that present notions about the nature of style are as old as the human ability to perceive differences, and a sampling of reasonably ancient cultural activity can seem to confirm the supposition. Much of the history of Chinese painting, for example, seems unthinkable without something like a modern critical apparatus for sorting out the dynastic periods, the local schools, and the copiers of venerated masters. It is almost impossible to imagine that the ancient Greeks were not thinking of style when they noted the differences between Doric and Ionic orders in architecture. The way in which ancient Athenians contrasted the "rational" sound of the stringed cithara and the "irrational" reed-pipe wail of the aulos suggests the distinction between classical and romantic styles that was made in the 19th century, and the parodies of the tragic dramatist Euripides by the Athenian comic playwright Aristophanes imply modern conceptions of a personal style. Much of what is now called stylistics seems to have existed already in the long succession of ancient Greek and Latin treatises on rhetoric. In sum, examples from both East and West can seem to support the assumption that there has always been something that may be called style.

INVENTION AND DISCOVERY

That "seem," however, needs very heavy emphasis, for some unexpected facts lie below the surface of the sort of sampling that has just been cited—facts that are open to more than one interpretation, but certainly not wide open.

Historical background. The admirer of Chinese painting who consults the old texts on the subject will find illuminating accounts of brushwork and much wisdom about the creative process but practically no discussion of style in the full modern sense of the word. In Indian aesthetics since ancient times, the doctrine of *rasa* (Sanskrit: "essence") has been used in reference to the flavour or sentiment of works of art and to the modes of affective response to them, but not to what is properly called style. In the West, ancient writers on art exhibit a similar and finally rather enigmatic failure ever to focus squarely on the subject. Linguistic evidence, while inconclusive, suggests that in the Greco-Roman world there was no word that meant quite what is now generally meant by "style."

Vitruvius, the Roman authority on architecture, writing sometime during the 1st century BC about the Doric, Ionic, Corinthian, and Tuscan orders, avoided even the Latin word *ordo* ("arrangement") and contented himself with *opus* ("work") and *genus* ("kind"). The Greek traveller Pausanias, writing about the visual arts in the 2nd century AD, used *kataskeuē* ("device" or "method of fit-

ting out") and *ergasia* ("work"). Among writers on literature and rhetoric a parallel tendency is apparent. Speaking of the style of an author, Aristotle was likely to refer simply to *lexis* ("speech" or "word"), and he was also likely to be talking merely about lucidity. The unidentified Greek critic known as Demetrius, writing probably in the Hellenistic era, used *charaktēr* ("quality" or "mark"), but the noun carried overtones from the verb meaning merely to scratch or engrave. The anonymous Latin text called the *Ad Herennium*, dating from around 85 BC, used *figura* ("figure"). A generation later Cicero used an arsenal of terms that included, with greatly varying degrees of precision, *figura*, *color*, *habitus* ("condition," "character"), *dictio* ("diction"), *elocutio* ("elocution," in a very general sense), and *genus*.

"Style" and *stilus*

Only in Late Latin does *stilus*, the word for the sharp-pointed instrument for writing, usually on wax, begin to mean also a manner of writing, as "pen" now does in such expressions as "a fluent pen" and "an acid pen"; and even here modern readers must be alert, for the derivation of English "style" from *stilus* does not prove that *stilus* always meant "style." The Latin term was reserved entirely for discussions of writing and speaking and usually for treatises on rhetoric; moreover, it seems to have implied little more than style in the sense of a skill or grace, and of a manner sanctioned by a standard. Apparently an author or orator in the closing years of the Roman Empire, in the 5th century AD, could have a periodic, loose, effective, ineffective, elevated, elegant, plain, high, middle, or low *stilus* but only very exceptionally, if ever, an idiosyncratic *stilus* that expressed a personality. And, again apparently, an architect, painter, sculptor, or musician could not have a *stilus* at all.

No important change in the usage thus established can be detected during the European Middle Ages. Words that suggest a kind, category, or mode of artistic production continued to be used in contexts in which a modern critic or historian might think in terms of a characteristic or expressive manner, or style. In architecture and the other visual arts, *opus* continued to be favoured; the French Gothic style of building was called Opus Francigenum, and English-style embroidery was know on the Continent simply as Opus Anglicanum. In music, the stylistic change apparent at the beginning of the 14th century, especially in France, was referred to as a new art: Ars Nova. The Latin *stilus*—and eventually its derivatives in other languages—was used only for talking about writing and speaking, and normally in the old rhetoricians' sense of a nonpersonal style sanctioned by a standard. When Dante (*Purgatorio*, xxiv) refers to the *dolce stil nuovo* ("sweet new style") that appeared in Italian poetry at the end of the 13th century, he stresses the importance not of the authors' personalities but of making the manner suit the matter and the occasion. The Host in Chaucer's *Canterbury Tales*, of the late 14th century, has this sense in mind when he addresses the Clerk of Oxford, in the prologue to the latter's tale:

> Your termes, your colours, and your figures,
> Kepe hem in stoor till so be ye endyte
> Heigh style, as whan that men to kinges wryte.
> Speketh so pleyn at this tyme, I yow preye,
> That we may understonde what ye seye.

In Renaissance Italy a shift in attitudes is apparent. Giorgio Vasari, for instance, in his widely influential *Lives of the Most Eminent Italian Painters, Sculptors and Architects* . . . (first edition in 1550, enlarged edition in 1568), built up a fairly consistent terminology on the basis of the word *maniera* ("manner"); his *maniera tedesca* ("German manner") refers to Gothic architecture, *buona maniera greca antica* ("good antique Greek manner") to ancient classical architecture, *maniera vecchia* ("old manner") to Byzantine or Byzantine-influenced painting, and *maniera moderna* to Renaissance architecture and painting. The *stilus* of the rhetoricians, *stile* in Italian, was still, however, reserved for literature. Not until around 1600 did musicians use such expressions as *stile moderno* and *stile rappresentativo*, and *stile* in criticism of the visual arts came still later.

In Britain, the equivalent extension of usage does not occur until the 18th century; the *Oxford English Dictionary* gives 1706 for the earliest reference to "style" in painting and 1728 for the earliest application of the term to music. Concerning the earliest English application to architecture there is some disagreement involving connotations, but a case has been made for a passage in Henry Fielding's novel *Tom Jones*, written in 1749: "The Gothic style of building could produce nothing nobler than Mr. Allworthy's house."

Earliest English usage of "style"

After that there was an era of the refinement of labels. A sharpened distinction between the art of ancient Greece and that of Rome began to be made in the 1760s. The division of British medieval architecture into Norman, Early English, Decorated, and Perpendicular styles dates from 1817. The term "roman" (Romanesque), referring to the architecture of western Europe before the Gothic, appeared in French criticism around 1820. Around 1850, "Rococo," after a career in slang, became a serious term for the fanciful style of the 18th century. The idea of the Renaissance as a cultural period, and not just an artistic movement, became fully-fledged around 1860. Definitions of the post-Renaissance styles of Mannerism and the Baroque were elaborated in the late 19th century. By the early 20th century, journalists were applying their own coinages to the styles discernible in modern painting and poetry.

Contemporary thought. The interpretation of the centuries-old mass of mostly semantic data on style is difficult and, for many people, exasperating. How is it possible to reconcile seemingly adequate perceptions of style throughout history with the long lack of adequate terms for it? How can the apparently adequate term, *stilus* or a derivative, be reconciled with inadequate perception of what it now connotes? Assuming that Chaucer's pilgrims reached their destination, what did the Host, familiar as he was with the figures and the high style of the rhetoricians, think when he was confronted by the Frenchness of Canterbury Cathedral?

Retrospective interpretations

Some historians have taken the easy course of assuming that when their ancestors used such words as "kind," or "work," or "speech" in certain contexts they somehow actually meant not what such words normally meant but what is now meant by "style." When *stilus* or a derivative was used, according to these historians, it somehow actually meant not what the rhetoricians meant but what is now meant in references to the composer Igor Stravinsky's or the novelist Ernest Hemingway's personal style. More rigorous minds have decided that style, viewed in the perspective of linguistic and general cultural history, is a will-o'-the-wisp. The majority of scholars, however, to judge from published essays, are reluctant to tamper with the evidence or to indulge themselves in a comfortable skepticism. They might therefore agree with the position that will be adopted in this article, which is that style in the arts is to a considerable extent a discovery, and to a large extent an invention, of a surprisingly late date.

Like many another cultural invention-discovery, style had a basis in human behaviour, but the distance between an ordinary ability to recognize things and what is meant by style in the arts is about as great as the distance between ordinary human memory and what is meant by history. Again like many another invention-discovery, style had a long phase during which some important levers and gears were already in place and more or less functioning. But the main job of assembling and powering the apparatus has been done since the end of the Middle Ages. Contributing to this development were Renaissance ideas of the importance of the individual personality, as opposed to medieval collectivism; 18th-century ideas of order and taxonomy in the natural sciences; and 19th-century ideas of biology, history, and, of course, aesthetics, all of which provided analogues for stylistic perceptions. The inventing and discovering are still going on, with much help from the experiments of artists and from such disciplines as archaeology, anthropology, psychology, sociology, and linguistics.

To talk of the nature of style in these terms is to raise some difficult philosophical questions, ranging from those

Philosophical questions on the nature of style

posed by ancient thinkers down to those of recent logicians. For although in practical affairs it may be willingly granted that every invention is to some extent also a discovery and that every discovery involves some invention, the fact is that the word "invention" implies one sort of being and the word "discovery" quite another sort. Moreover, by a curious reversal of what happens in many inquiries, the philosophical questions that are thus raised for art critics, historians, and appreciators tend to be less pressing in regard to style as a concept or a collective noun than to style as it is actually experienced. People speculate calmly, if at all, about the nature of their comprehensive stylistic assumptions and become heated about the alleged reality or unreality of the style of the 19th-century French sculptor Rodin or of the *style galant* in 18th-century music or of the style of the Renaissance.

Can anything useful be said in this realm? Certainly any attempt to deal thoroughly with the issues would lead far beyond the scope of this article and deep into problems for which professional philosophers have not yet found accepted solutions. But it seems legitimate to confront antistyle "realists" with the suggestion that a style is no less real, or no more unreal, than a work of art, which is also a kind of invention-discovery, and to add that this degree of certainty is all that a well-tutored man should expect. Works of art can be, among other things, physical objects, imaginary objects, enduring possibilities, realized possibilities, sensible phenomena, insensible phenomena, sheer processes, and even, according to respectable opinion, transcendental entities. They can also have—and this is important to the argument—what has been called an emergent mode of existence and might be called a "do it yourself" mode; a picture, for instance, emerges from the blobs of pigment on a canvas when the viewer steps back and perhaps squints, and a symphony emerges from blobs of sound in the same way. All these modes of existence, and the emergent in particular, can be found among styles. Rodin's style emerges when a sufficient number of his works are contemplated from a certain psychic distance, and it is also a bronze entity in one of his statues of Balzac. The *style galant* emerges from compositions by Bach's sons, and it is also in a single Mozart serenade, which itself exists on paper, in performances, and, above all, as an enduring possibility. The Renaissance style, according to the scholar Arnold Hauser, "is at once more and less than what has actually been expressed in the works of the Renaissance masters. It is something like a musical theme of which only variations are known." In short, it is probably most usefully thought of as having an emergent mode of existence.

All this might be summarized by remarking that the process of invention and discovery that produced the general notion of style over a period of centuries is constantly being recapitulated by individuals, sometimes over a period very brief indeed, for particular styles. From this, one might hastily conclude that practically anyone can invent-discover what will pass unchallenged as a style; and anyone acquainted with modern art scholarship and art publicity must grant that there is a measure of truth in the conclusion. But in the long run, of course, there are certain limits to what can pass as style, just as there are certain limits to what can pass as a work of art; eventually opinion accumulates to the effect that the invention-discovery in question does not work well enough, or is simply not important enough, to qualify for the standard label. The annals of archaeology in Latin America and the eastern Mediterranean are strewn with styles that broke down, often after extensive repairs by their inventors. And, in fact, the world's major accepted styles turn out on examination to have more rigour and clarity in their nature than might be supposed. They have a number of recurring features that can be extracted and combined so as to constitute a working model of style, a sort of metastyle, that can be used for dealing critically with new labels.

PRINCIPAL ASPECTS OF STYLE

These recurring features can be grouped and considered under the headings of value aspects, poietic (or creative) aspects, morphological (or formal) aspects, metaphorical aspects, polar aspects, and measurable aspects. The word "aspect" is preferable to "feature" or "element" or some other possibility, for it must be kept in mind that a style is an invention-discovery with several modes of existence. Moreover, in each instance, what is being talked about is likely to be affected by the different viewpoints of producers, consumers, historians, critics, and other observers.

Value aspects. The first group on the list is logically defective, since all the other aspects have value aspects, but it is important enough to merit some separate preliminary treatment. That styles are regarded as desirable is evident from common usage. Merely to say that an artist, a work, or a period has style is to judge him or it favourably; merely to use "style" in preference to such words as "manner" or "fashion" is to imply value; even to say that a thing is in a poor style is often to suggest that it does not have enough style. And that styles are actually worthwhile is difficult to doubt. They provide the art appreciator with the pleasure of recognizing somebody or something, a pleasure that is certainly among the basic ones of human existence: witness the popularity of handbooks that tell how to distinguish Tang from Sung styles in Chinese art or Louis Quinze from Louis Seize in French art. Styles provide the artist with the pleasure of being recognized, and they do so without the self-display of a signature. They are like codes; in the language of information theory, they help obtain an invariant output from a variable input, and nearly all art history, which is a comparatively recent phenomenon, and much criticism makes use of them. They also have many less evident sorts of value. They function as the signs and, to some extent, as the agents of integration in individual artists, groups of artists, and sometimes whole cultures; they are brakes on alienation. They are appetizing and preservative, like spices; they have saved from oblivion a number of great minds whose ideas have lost their fascination. To cite only a few examples from British critical and historical literature, it is hard to imagine anyone still reading much of Dr. Samuel Johnson, Edward Gibbon, Thomas Babington Macaulay, Thomas Carlyle, or John Ruskin for content alone.

Evidence from common usage

These positive aspects are accompanied, of course, by some negative ones, and the latter have been worrying critics increasingly during recent decades. Since styles stress similarities and work partly as codes, they tend to blur differences and to simplify excessively; they encourage many viewers to see, for example, merely "a Picasso" or "Cubism" where one ought to see the rich uniqueness of a painting. The pleasure of being recognized, and the money that may come with it, can encourage some artists to develop a mere trademark. Successful period styles of the past may foster, as they did among 19th-century European architects, an absurd amount of eclecticism and fancy-dress historicism. Successful current styles can generate an equally absurd amount of imitation among artists. The difficulty of defining specific styles and style in general creates serious problems in art history, which will be discussed in the last section of this article.

Negative aspects

Do these negative aspects outweigh the positive? The majority opinion is clearly that they do not, for there is no prospect of a return to the supposed innocence of the centuries before the massive, intricate, dangerous, useful invention-discovery got up steam.

Creative aspects. One of the complaints, however, is very legitimate: it is that people who talk about style do so too often from the viewpoint of an appreciator. The same complaint can be made about art discussions in general; a good deal is heard, for instance, about disinterested contemplation, which is fair enough from an appreciator's viewpoint but close to wild calumny from an artist's viewpoint. The hardworking men who built the Parthenon were certainly not disinterested in any usual way, nor were John Milton in writing *Paradise Lost*, Michelangelo in painting the frescoes for the Sistine Chapel, Richard Wagner in composing the opera *Tristan und Isolde*, and Leo Tolstoy in creating *War and Peace*, and what they were doing can scarcely be called contempla-

tion. It was rather what the Greeks called *poiēsis* ("creation," or simply "making").

Style, then, has what can be called poietic aspects; the adjective brushes jargon but lacks the Romantic connotations of "creative" and the matter-of-factness of such an alternative as "productive." The existence of these aspects can be posited etymologically, with the risk inherent in arguing from dead metaphors; *stilus*, as has been noted, originally meant a writing instrument; and such near equivalents for "style" as "manner" (Latin *manuarius*, "of the hand") and "fashion" (Latin *facere*, "to make") also have clearly poietic pedigrees. Moreover, in some contexts "style" still means little more than a mode of artistic production.

But here an attempt to sharpen common usage seems called for, since a style in its poietic aspects is not the whole of an act of making. It is only the part of the act that represents a deviation from a norm and that, as such, is apparent enough to offer the pleasure of recognition. The norm may be provided by a more inclusive style, by a tradition, by material conditions, or by some other frame of reference within which artists work. It may also be assumed, more or less arbitrarily, by observers. Thus the personal style of the 17th-century Flemish painter Peter Paul Rubens is not, poetically speaking, the whole of his way of applying paint to canvas but merely that part of his way that constitutes a recognizable deviation from a norm—the Baroque style—in which he worked. The Baroque, to continue the illustration, is not the whole of the Baroque painters' ways of applying paint to canvas but merely the part of their ways that constitutes a recognizable deviation from another norm, the general style that prevailed in European painting from roughly the middle of the 15th century to the end of the 19th, though in many histories of art it is instead assumed to be the High Renaissance style as exemplified by Raphael.

The role of norms

One can conclude that style is dependent on originality and the will to exercise it. But much depends also on the norm. In the first place, many norms are imposed, sometimes by accepted authority, sometimes by social pressure, and most often by unawareness of an alternative; the average Western composer, for instance, between Bach in the 18th century and Schoenberg in the 20th, seems to have regarded the tonal-style norm—the organization of tones and chords in a composition in relation to a keynote—as something like a law of nature. In the second place, certain norms contain fewer variables than others and therefore offer fewer opportunities for deviation; a Byzantine mosaicist had no possibility of developing a marked personal style. And, finally, there are the supernorms constituted by each art, by artistic materials, by languages, and by much else; at this level the number of variables may be decisive. Traditional sculpture offers fewer variables than painting and hence has yielded a much smaller number of styles. Granite offers fewer variables than bronze and hence has what might be called a lower yield in terms of style. Specialists in stylistics—the branch of linguistics that studies the variables in a language and their manipulation—have noted that one of the secrets of the 20th-century Welsh poet Dylan Thomas's strongly personal style was his discovery of unsuspected variables in English: thus where the norm had seemed to insist on nouns of temporal, or linear, measurement he could write "All the sun long," "A grief ago," and "farmyards away." In "Spelt from Sibyl's Leaves," the 19th-century English priest Gerard Manley Hopkins showed a comparable talent for finding possibilities of deviation:

> Earnest, earthless, equal, attuneable, vaulty, voluminous ... stupendous
> Evening strains to be time's vast, womb-of-all, home-of-all, hearse-of-all night.

Formal aspects. Much of the above might have been put under the heading of the morphological—or "formal," if certain connotations are ignored—aspects of style, and much that is traditionally morphological is equally poietic. The form of a work of art can be regarded as the record left behind by the making process; this idea, implicit in ancient rhetoricians' descriptions of prose and poetry (*e.g.*, as laboured), has been prominent in modern criticism since the appearance in the 1950s of such process-emphasizing accomplishments as Action painting, in which the brush strokes and textures may be regarded as a record of the creation of the work; aleatoric music, in which the notes or sounds are selected by chance; and Brutalist architecture, the sort that leaves the concrete raw and plank marked.

The making process, however, involves the whole work, whereas form may be regarded as excluding content and including only shape, volume, space, structure, pattern, organization, texture, rhythm, imagery, emphasis, balance, and the like. This separation is often denounced by careful critics, and a successful work of art, when contemplated in a properly focussed and expanded state of awareness, does indeed present itself with form and content organically tangled. But there is little likelihood that art appreciation suffers when things are untangled for discussion, since most persons are quite capable of distinguishing between the critical analysis of a work and actual aesthetic experience. So it seems safe to accept common usage concerning "form" and then to agree with commentators who have been assuming for centuries that style has purely morphological aspects—without joining them in assuming that it has practically no other aspects.

Separating form and content

Immediately an apparently drastic shift in emphasis occurs, from one of variables and deviations to one of repetitions and conformity. In fact, style can be reasonably, if incompletely, defined as constant formal elements and their combinations, with content excluded except in certain circumstances. But the shift in emphasis is obviously just a result of the play of aspects. A style is always generated by the manipulation of available variables in such a way as to yield a recognizable deviation from a given or an assumed norm. If, however, the deviation is to be recognized as something more than an accident or a solitary impulse, it must be repeated, either identically or in a recognizable variation. It must become understandable, which a unique event—an accident—cannot be.

Thus a style is always, when perceived from what can be called the poietic stance, rather surprising; the Baroque norm and the English-language norm do not lead to any expectation of Rubens' fleshly swirl and Thomas' "A grief ago." But thus also a style is always, when perceived from the morphological stance, rather familiar; what was unexpected in terms of the act of making becomes expected in terms of form. Again the illustration can be continued above the level of personal style: the energized masses found in Baroque painting of the 17th century are constant in their deviation from the balance of High Renaissance paintings of the 16th century. In sum, an artist, or a group of artists, is obliged by the nature of style to move freely into constraint, and heretically into orthodoxy.

Metaphorical aspects. Hence certain styles are commonly said to be "characteristic," or "expressive." An aesthetic deviation, repeated sufficiently, may become converted into a form and yield recognition pleasure. The result may fairly be described as a manifestation of personality. The flame shapes in El Greco's paintings may be thought of as a handwriting, Bach's driving rhythms as a gait, Proust's long sentences as a voice; and usually no harm is done to understanding.

Such metaphors, however, along with the words "characteristic" and "expressive," can become whimsical or misleading in many situations. Moreover, the stylistic aspects in question are best perceived in depth and in the aggregate by noticing that nearly every style is itself a metaphor, functionally speaking. It implies, as Aristotle said a good metaphor does, an intuitive perception of the similarity in dissimilars. It works, as an ordinary simile does, by seizing striking likenesses and neglecting differences; to see that the 18th-century English poet William Cowper wrote in a Miltonic style resembles, as process, seeing that Robert Burns's sweetheart was "like a red red rose." There is a substitution of a part for a whole, as in synecdoche; in many eyes a pointed arch and a flying buttress are enough to evoke the Gothic style, and for many ears a single chord can summon up Beethoven.

Similarity in dissimilars

Most importantly, in nearly every stylistic context, and not just in those involving personal and "characteristic" styles, there are two sections, like the two sections in the comparing process of metaphor; these can be thought of as the "window" and the "view." In the simplest situations the view through the window of the style is of an individual artist, or of a well-defined group; the 19th-century French poet Stéphane Mallarmé may be sensed through his repetition of the word *azur*, and the contemporaneous group of Impressionist painters that surrounded Claude Monet are recognizable through their deviant, flickering brushwork. The situation, however, is seldom quite that simple. Through Mallarmé's style as well as through Impressionism the view may be of a doctrine or a program; in other situations it may be of a class, a nation, a place, a period. There are styles that bear a functional resemblance to myths, if the latter are thought of as communal metaphors. The view that looms through the symmetrical frontality of the rigidly posed figures in ancient Egyptian sculpture is of an entire culture, a culture that is dramatically different from the one that looms through the asymmetrical twist, the contrapposto, of figures in Italian Renaissance statues.

These remarks have to be qualified, for if common usage is accepted there are styles that seem to have no metaphorical aspects, that are practically opaque. Possible examples are procedural styles in general: those described by the ancient rhetoricians, those associated with the fixed forms of music and poetry, those that are just simplifications, often geometrical, of natural forms.

Polar aspects. Common usage also provides evidence for the existence of certain structural or self-defining tendencies in style that can be grouped under a single heading as polar aspects. These have long been noticed; Hellenistic and Roman literary critics have a lot to say about the flowery, redundant oratorical prose known as Asiatic, which is presented as the diametrical opposite of the plain, economical Attic. The invention or discovery of such polarities did not get seriously under way, however, until the 18th century in western Europe; and most of it has been accomplished, mainly by German thinkers and largely in the visual arts, since the late 1800s.

Among the great number of such contrasting pairs that could be mentioned are haptic–optic (*i.e.*, oriented to the sense of touch as opposed to sight orientation), idealistic–naturalistic, multifarious–unitary, closed–open, linear–painterly, and many more. It will be noticed that the trend is toward all-inclusive world styles, or at least toward constantly recurring stylistic features, and that the emphasis is strongly on morphological aspects. This emphasis, although open to the usual objections to "formalism," has compensated handsomely, in terms of instrumental value, for the sometimes naïve scientism and general overconfidence implicit in the labelling. Indeed, it is not an exaggeration to say that the best visual-art criticism and history published since World War I could not have been written without the help of polar analysis of form.

Applicability of polarization

But pairs like the rather abstruse ones mentioned are not the whole story. Any style, including the familiar established ones, may be polarized; the 19th-century French painter Delacroix's Romantic style may be paired with his older contemporary Ingres' Neoclassical style, the Gothic with the Renaissance, the French with the English, the Christian with the Islāmic, the Eastern with the Western. Each member of each pair is defined in terms of what the other member is not; each is at once the deviant from the other and the norm for the other. Often in such pairs the forms are not seriously analyzed; mere diametrical opposition, as in traditional political parties, is felt to be enough. This peculiarity may be especially evident in connection with a modern style, which is always polar to begin with and which may stay that way until in its turn it begins to cease to be modern. Only then may historians have a good chance to cut through the partisan propaganda, get at the constant forms, and decide if the style is internally consistent enough to merit a label of its own and a place in the parade that began some 40,000 years ago.

Measurable aspects. The historian who undertakes such a task is not likely to approach the constant forms with a yardstick, for in general the measurable aspects of style are not very highly considered by artists and art critics. But such aspects do exist.

The differences between Doric and Ionic orders in classical architecture are not only in the abacuses and volutes that decorate the columns and their capitals but also in proportions and in the number of flutes on a column. The 20th-century French architect Le Corbusier performed his subtle manipulation of architectural variables with the help of a system of proportion, which he called the *modulor*, based on the human figure. Sculpture styles were influenced for centuries by the canon (now lost) of Polyclitus, a Greek sculptor of the 5th century BC who believed that "the beautiful comes about little by little, through many numbers." In a portrait in the Mannerist style of mid-16th-century Italy, much of what makes it Mannerist may be a matter of how long the body is. Some of Hogarth's pictures can be analyzed in terms of what he called "the line of beauty," obtainable by winding a "precise serpentine line around the figure of a cone." Painters of the 20th century have revived the interest Renaissance artists had in the proportion (about 8:13) known as the golden section. An important part of the stylistic difference between a movie director of around 1930 and one of around 1970 may be discovered by simply noting the smaller number of camera shots and sequences used by the latter. Styles in poetry can be specified, often with surprising results, by counts of images, rhymes, run-on lines, and metrical variations; and such methods are accurate enough to help date Shakespeare's plays. Prose styles are, for certain modern linguists, a matter of the statistical averaging of the use of certain words, performed with the help of a computer. That musical styles have measurable aspects has been clear, of course, since at least the time of Pythagoras, who discovered in the 6th century BC the relationship of musical intervals to the lengths of strings; and the fact has become freshly clear under the impact of 20th-century science and technology. Computers have become standard equipment for many composers; synthesizers have become generators of styles translatable into mathematics; Beethoven's poietic deviations from a norm have turned out to be quantifiable in somewhat the same way as the unforeseeables studied by information theorists.

Qualifications on quantification

Some qualifications are in order. The analysis of a style is not the same as the experience of a style: the whole of a work of art is certainly not the sum of its measurable parts. While it may be true that under certain conditions quantity turns into quality, it does not follow that every quality can be quantified; and certain conditions—Beethoven's genius, for instance—remain to plague the quantifier. But all this does not alter the fact that styles do have certain measurable aspects. Nor does it excuse the neglect of these aspects by some art appreciators, critics, and historians. Perhaps the remedy both for the shortcomings of the quantity-minded and for the attitude of the quality-minded will eventually be found in interdisciplinary work on aesthetic problems.

II. The varieties of style

Since it is part of the nature of style to provide recognition pleasure, and since this pleasure is usually accompanied by an irrepressible impulse to name, one can suppose that unlabelled styles are rare. Would that they were not, an archaeologist may say; would that there were an opportunity to classify works of art scientifically and to substitute numbers or New Latin labels for such misnomers as "Gothic" (which is unrelated to the Goths) and "Cubist" (which has little to do with cubes). Actually, however, the downright mistaken or merely derisive labels are neither very numerous nor very misleading; to anybody who knows enough to be interested, "Gothic" is likely to mean something like "Medieval West European III," and "Cubist" something like "genus, partly abstract; species, Picasso–Braque 1907–14." And if the majority of style names are not scientifically descriptive, they do as a rule offer adequate clues not

only to the thing being talked about but also to the class, or classes, to which the thing belongs. In other words, the familiar nomenclature, although accumulated apparently haphazardly through the centuries, has taxonomic—more precisely, typological—implications. To speak of Rembrandt's style is not only to refer to certain poietic deviations from a norm and to certain recurring formal elements; it is also to imply the existence of a personal variety of style plus a more general sort that includes the personal variety. To speak of a realistic style, or of one of its several polar opposites, is to imply a different variety and a correspondingly different general sort.

When such implications are grouped, they yield some 13 varieties of style in the arts. (The "some" is inserted here to allow for reasonable differences of opinion as to where the dividing lines should be drawn.) These 13 varieties are clearly of two general sorts, which emerge from two types of "view" beyond the metaphorical "windows" that styles create in a sufficiently knowledgeable imagination. The first general sort, to which Rembrandt's style belongs, affords views that focus on single cultures; the second sort, to which a realistic style may belong, affords views that cut across cultures. It will be noticed that a given style may move from one category to another; more will be said later about this mobility. But most styles are reasonably stable, and for the moment it is convenient to assume that the others have certain recognizable home categories. Even a slightly unstable classification can stiffen discussion.

Two types of "view"

SINGLE-CULTURE VARIETIES

Single-culture styles are usually inhabited, so to speak. They usually evoke, more or less in the foreground of the contemplating imagination, either a person, a school, a social class, an ethnic division, a regional community, a nation, an ecological division, a religious community, or the generations that constitute a period. Rembrandt's style usually evokes the bulb-nosed, sad-eyed person known through dozens of remarkably self-searching self-portraits; the Venetian style usually evokes the 16th-century masters Giorgione, Titian, Tintoretto, and Veronese; each style in African sculpture usually evokes a tribe; and so on down the list. Single-culture styles are therefore frequently said to be "characteristic," or "expressive," of a particular people; and in this context these familiar terms of interpretative art criticism may seem to triumph over the mild objection raised earlier, in the discussion of the general metaphorical aspects of style. In fact, these adjectives, and also the noun "people," raise some awkward problems even here.

Personal styles. No variety of style seems, at first thought, quite as vividly, specifically, indubitably inhabited as the personal variety. Quoting the celebrated and seldom-read *Discours sur le style* (1753), by the Comte de Buffon, and neglecting his qualifying remarks, many appreciators assume confidently that "the style is the man himself." In a somewhat modified form, the same assumption can be found as far back as the 1st century in the Stoic moralizing of the Roman philosopher Seneca. In a somewhat pseudoscientific form it has produced some disturbingly glib psychoanalysis of works of art. In its plebian form it has led to such suppositions as that the flame shapes in El Greco's paintings are evidence of astigmatism, the long sentences in Proust's novels evidence of asthma, the rhetoric of Liszt's piano pieces evidence of Gypsy blood, and the right angles of Mies van der Rohe's architecture evidence of a subtle totalitarianism. The notion may be said to have reached one of the peaks of its career in 1935 in the earnest excogitation of the literary scholar Caroline Spurgeon; after counting and sorting Shakespeare's images, she concluded that the poet was

Style and the man

> a compactly well-built man, probably on the slight side, extraordinarily well coordinated, lithe and nimble of body, quick and accurate of eye . . . probably fair-skinned and of a fresh color, which in youth came and went easily . . . very sensitive to dirt and evil smells . . . gentle, kindly, honest, brave and true.

She also saw, through the obviously wide-open window of the personal style, a man who at 35 had "probably experienced heartburn as a result of acidity."

It is easy to call this sort of interpretation wrong and not easy to explain exactly why it is wrong. After all, everyone indulges in it to a degree. Miss Spurgeon, of course, went a bit too far; she was neglectful of complexity and of the possible differences between a dramatist and his personages. But her counting and sorting of images was a valuable and influential piece of research; it demonstrated, in an irrefutable way, the existence of some of the deviations and constants that make up Shakespeare's personal style. If the person who is certainly recognizable in this personal style is not quite the Stratford man himself, who is he?

When the question is asked about a large enough number of personal styles, a tentative answer may emerge. The style, it appears, is not the man himself but the artist himself—mostly, at least. Naturally, the artist is to a considerable extent the man; he has much of the latter's native capacities, acquired skills, secret drives, and painful defects. But the artist is a professional role, a programmatic personage, a cultural configuration, a persona; in sum, he is a remarkably, often deliberately, synthetic personality. Further, he is conditioned by much besides the man himself and notably by the work of other artists. To put all this another way, every personal style is in part sheer performance, and every artist as such is in a sense (not a pejorative one) a performer. Mies the man went on living in his relatively old-fashioned Chicago apartment, while Mies the artist was "performing" with gleaming right angles in the tall apartment buildings he designed for Chicago's fashionable Lake Shore Drive; Mozart the man disliked flute music, while Mozart the artist "performed" by composing flute music; Petrarch himself philandered, while Petrarch the poet "performed" as a faithful worshipper from afar of the idealized Laura.

The person and the performer

Hence, in the opinion of many modern critics, the once-popular problem of personal stylistic sincerity is meaningless; to pose it is to mistake art for life. Hence also the distinction that is often made between creative personal styles and performing personal (or group) styles should not be too categorical. Although a dramatic text, a musical score, or a notated ballet may seem to constitute a norm that offers a very small number of variables, in practice a competent actor, musician, dancer, conductor, or director usually manages to produce enough deviations to have an easily recognizable personal style. No opera-record collector is likely to confuse an interpretation by the intensely dramatic soprano Maria Callas with one of the same role by the serenely lyrical soprano Renata Tebaldi; and ballet literature suggests that the ethereal 19th-century ballerina Marie Taglioni was as different from her rival the sensuous Fanny Elssler as the 19th-century Italian operas of Vincenzo Bellini were from those of his contemporary Gaetano Donizetti. Moreover, an interpreter has about the same choice as a creator among the more inclusive styles that can always be recognized simultaneously with a personal one; he can be modern, traditional, Classical, Romantic, Baroque, or whatever. And finally, if he tries to be as faithful as possible to his text, the result will approach another personal "performing" style, that of the artist who composed the work. To succeed in producing *Phèdre* exactly as it was conceived would be to play the player Jean Racine.

These remarks should not be interpreted as a complete denial of the presence of "the man himself" behind a personal style—into the polar opposite of Miss Spurgeon's error. Since the artist is partly conditioned by the man, so, of course, is the style; and stylistic evidence can often be made to match biographical information in an enlightening way. Friends of Mies noticed in his manners and dress a certain fastidiousness and a love of good material that reminded them of his architecture. The reported simplicity, honesty, and modesty of Haydn are an agreeable match for qualities in his musical style. Most readers probably sense an authentic personality, a real voice, behind the sprung rhythm and breathless rush of the verse of Hopkins. Also, some quite successful methods of sty-

listic analysis apparently depend on a strict correspondence between the manner and the man himself; a fascinating example is the technique for attributing paintings, developed by the Italian art critic Giovanni Morelli in the 19th century, which assumes that the touch of a particular master can best be detected in unimportant details, such as the ears in a portrait, that were presumably painted without taking much thought.

None of the counter-evidence, however, seriously shakes the argument that the person in a personal style is mostly the artist as such, and some of it does not stand up very well under scrutiny. Fastidiousness, simplicity, and vehemence do not become meaningful in this context until they are given energy and focus by the artist; and the telltale details used for attributing paintings may have about as much aesthetic interest as fingerprints.

School styles. When artists are considered as a stylistic group, or school, all the problems raised by personal styles reappear in new guises in the company of other problems; and one of the more nettling of the latter is the precise meaning of "school." For there is no denying that this part of the apparatus of criticism has got badly out of hand since its invention and discovery in the 18th century (largely by the pioneer Italian archaeologist Luigi Lanzi). Art critics, historians, and especially painting-museum curators have acquired the habit of using the term as an elegant variation for what usually turns out to be merely a country of residence; thus J.M.W. Turner is said to be a painter of the British school and Thomas Eakins of the American school. Sometimes the geographical designation is narrowed, with a commensurate gain in stylistic information; thus Mantegna is said to belong to the North Italian school and Perugino to the Umbrian school. The gain, however, may be in confusion; the division between the so-called Northern and Southern Sung schools of Chinese painting, for example, has been called the most misleading and arbitrary in art history. Sometimes geography is abandoned for a general stylistic designation; thus the 18th-century French painter Jean-Baptiste Chardin is said to belong to the Realist school. Sometimes the stylistic designation is narrowed drastically, and then a gallery visitor may be confronted by a brass plaque attributing a 15th-century Florentine painting to the school of Fra Angelico, for instance; this can mean that documents point to the studio or to a follower of Fra Angelico; or that the picture is an ancient, and therefore respectable, copy; or that it looks rather like a genuine Fra Angelico without his customary quality—the unstated premise being that all pictures by Fra Angelico are first class.

The meaning of "school"

The situation is regrettable not only because one word is being forced to do things other words can do better but also because "school" has its own work to do. Musicologists can profitably use the term to talk about the centres of musical activity that existed in the 12th century at such places as Paris, Compostela, Padua, and Winchester; or about the late-16th-century amateur "academy" known as the Florentine Camerata, in which the monodic style that led to opera was fostered; or about the group of composers of atonal music that surrounded Arnold Schoenberg in 20th-century Vienna. Painting historians must refer to the Tours school of Carolingian miniaturists, the Shen Chou school of 15th-century Chinese ink artists, the Barbizon school of 19th-century French landscape painters. Literary historians must consider such schools of poetry as those of the 16th-century French Pléiade, the English Lake poets of around 1800, the Tokyo (then Edo) haiku masters of the 17th and 18th centuries, the American Imagists of around 1914. Even architectural historians, who tend to think in large units, have to take account of such phenomena as the Burlington group of Palladianists in 18th-century London, the Glasgow School in the 19th century, the slightly later Chicago School, and so on. Each of these examples, and of the hundreds of others that could be cited, involves a well-defined and usually not large geographical area, a relatively short time span, and a relatively small number of artists working in a describable shared style; here are the essential requirements for using the term "school" profitably. Of course, a curator, in announcing that Turner is of the British school, may really intend to commit his museum to the proposition that a nationally shared painting style has been perceptible in Great Britain down through the centuries. Such is not usually the intention, however, and when it is, it should be made explicit.

The confusion is worth dwelling on because in many works of art a school style, defined as the shared style of a relatively small number of artists, is as striking as a personal style. It is a "window" that affords a "view" inhabited by a synthetic personality, a programmatic personage, almost (but not quite) vivid enough to justify thinking in terms of something like an overartist, as some German philosophers have. Moreover, in the interaction between a personal style and a school style, there is a model, manageable for study, of the complex relations that emerge when several styles are found together in a given work. The personal style of John Donne, the 17th-century English poet of the Metaphysical school, is recognizable in the dense, macabre imagery, the sometimes violently wrenched metre, and the self-dramatizing switch of the following lines from his "Nocturnall upon St. Lucies Day, being the shortest day":

Interaction of artist and school

> The world's whole sap is sunke:
> The generalle balme th' hydroptique earth hath drunke,
> Whither, as to the beds-feet, life is shrunke,
> Dead and enterr'd; yet all these seeme to laugh,
> Compar'd with mee who am their epitaph.

At the same time, the style of the Metaphysical school is apparent in the rather conversational tone, the compact syntax, and the use of extravagant poetic conceits. Deviations and a norm solicit attention, and the solicitations will multiply if the recognition process is continued into the maze of styles—the Elizabethan, the Jacobean, the English, the religious, the aristocratic, the Mannerist, the formal, the haptic, etc.—which a sufficiently subtle and patient critic may discover in Donne's poetry. The common reader, in Dr. Johnson's sense, can be excusably irritated at a certain point by the game of hide-and-seek between the different and the same, the self and the other. But, in fact, such simultaneous recognitions are normal in the actual experience of art; and they are no more mysterious than recognizing, with confidence and usually without being able to say exactly why, that a given face is of an individual, a family, a region, a nation, and a race—is at once itself and not itself.

Social styles. On a scale of recognizability, many critics would probably put social styles—those associated with a particular class or section of society—directly after personal and school styles, at least when much of the art of the centuries before the Industrial Revolution is being considered. Even an untrained appreciator can sense courtly styles in the intricate fixed forms of troubadour verse, the stiff etiquette of a Louis XIV portrait, and the languors of medieval Japan's *Tale of Genji*, by Murasaki Shikibu; bourgeois styles in the solid forms of 17th-century Dutch still-lifes, the uncomplicated rhythm and harmony of a Protestant hymn, and the matter-of-fact, 18th-century English prose of Daniel Defoe; and peasant or proletarian styles—often more accurately described as traditions—in songs, carvings, and embroidery. In the 20th century such clear-cut social styles have become less and less noticeable, partly because class distinctions have become much less evident in the technologically more advanced nations. Also, artists have ceased to know for whom they work, the art market having replaced, except on special occasions that are most frequent for architects, the old system of direct commissioning by patrons. Nevertheless, a sophisticated, or merely mischievous, critic can point to contemporary social styles; some evoke the established families, others the new millionaires. A number are related to young people, who since the 1960s have exhibited many of the economic and cultural characteristics of a separate class. Politicians in all countries throughout the 20th century have occasionally attempted, sometimes on a totalitarian scale, to impose on artists one of the old courtly or bourgeois styles; and a few dictators, notably Hitler and Mussolini, have temporarily revived an imperial-court style with the par-

Tradition and class distinction

venu touch, familiar to art historians in such outsized forms as those of Darius's palace at Persepolis in ancient Persia and Napoleon's church (originally temple) of the Madeleine in Paris.

Here two concessions seem called for. The first is that in talking about social styles it is often impossible, even for the purpose of cold analysis, to keep morphological aspects separate from content. The second is that throughout history the artist as such has been remarkably, sometimes depressingly, available for the "expression" or "characterization" of a social stratum other than that of the man himself. Perhaps the artist as such—shaman, bard, craftsman, entertainer, 19th-century demiurge, 20th-century iconoclast—has been rather more of a performer than has already been suggested.

Ethnic styles. Like social styles, ethnic styles have become less noticeable in the modern era. Their formerly high level of recognizability, their complex morphological aspects, and their surprising persistence over long periods have made them, however, favourite subjects for study among both archaeologists and aesthetics. Good examples are plentiful in the Indian arts of North America; the sculpture, painting, and music of black Africa; the pottery of the ancient Middle East and east Asia; and the surviving decorated objects, mostly metalwork, of the so-called barbarian peoples who moved across Asia and Europe between roughly the 6th century BC and the time of Charlemagne, at the end of the 8th century AD. In some instances scholars have been able to trace the borrowing of motifs; the Germanic animal styles of the migration period, for example, show the influence of Roman figurative art and Mediterranean ribbon ornament. But the borrowed motif is invariably transformed by a repeated deviation into one of the morphological constants of the borrowing tribe, and the reasons for this enduring assimilative capacity are not well understood.

Theories of stylistic disposition

The notion of a physically inherited stylistic disposition has long since been discredited, and archetypes in a collective unconscious, as postulated by the 20th-century Swiss psychologist Carl Jung, have been dismissed by the majority of professional historians. Among the more attractive theories are some that point to analogies with the inertia and the assimilative capacity of a language; yet even these seem inadequate before such a fact as that the Eskimo ethnic style has lasted for about two millennia. Another attractive theory links the lack of stylistic change to a general lack of history—or at least to a general lack of awareness of history. But if this theory is plausible in a North American or an African context, it is much less so in the context of the Asian and European migrating peoples, who managed to pass through an immense amount of history without making important changes in the design of their crowns, buckles, and other useful objects.

Regional and national styles. The problem of stylistic stability re-appears in the consideration of regional and national styles, which are partly just ethnic styles that have settled down. But they are always more than that. The high—and polar—recognizability of Oriental and Occidental styles cannot be satisfactorily accounted for by references to ancient tribes and by linguistic analogies; nor can the long preoccupation in the Mediterranean basin with human forms; the long preoccupation in northern Europe with animal, zoomorphic, fantastic, symbolic, and abstract forms; the rigidity of Egyptian pictorial and sculptural forms during some 3,000 years; the persistently emotional, romantic, and expressionist tendencies in German music, painting, and literature; the French emphasis on structure in architecture, painting, and poetry; the English linear and decorative tendency that runs through medieval miniature painting, Perpendicular Gothic architecture, the drawings of William Blake, and Art Nouveau. Common sense, of course, is needed in thinking about such styles; the theorizer who sets out to show that all French art is rational and all German art emotional will be in trouble immediately. Also, a certain vagueness is often suitable, for the distinctive features of a regional or national style may be of the sort better described as qualities than as morphological aspects. Recurring differences, however, finally add up to recognizability, and a quality hard to define can be strongly felt.

Climate and landscape were formerly popular as explanations for regional and national styles (for all styles, as a matter of fact); Gothic architecture was supposed to have emerged in northern Europe because of the many forests. Evocations of some kind of permanent national outlook were also frequent; Russian musical styles were thought to be inhabited by a Slav soul. Such ideas are now perhaps too much out of fashion; it is not quite unthinkable that an English stone carver's affection for linear pattern was encouraged by the frequent absence of bright, shadow-casting sunlight in England and that the symmetry of much French architecture and painting corresponds to the average Frenchman's enduring attachment to order. But eventually, of course, an explanation must include an entire regional or national culture and environment and at the same time take note of the fact that an art may have an existence of its own. To neglect this latter possibility would be to repeat on a large scale Miss Spurgeon's confusion of men with writers.

Ecological styles. A subvariety of the regional variety is perhaps distinct enough to be classed separately as the ecological: here, that is, the style in question evokes in fairly specific ways the relationship of human organisms to their environment. Here also there are likely to be strong polarities; typical contrasting pairs are urban–rural, mountain–plain, hunting–farming, inland–seaboard, nomadic–sedentary, capital–provincial (in some ways), and (at least in comic strips and science fiction) earth–space. Such styles are most recognizable in architecture, painting, and the making of such useful objects as furniture, tools, and weapons. But they may be recognized more distantly in the dance, in music, and in poetry: the imagery of the Parisian poet Charles Baudelaire is urban; that of the New England farmer-poet Robert Frost is rural; while Homer's is seaboard, at least in the *Odyssey*. It is probable, too, that a well-defined ecological pattern affects in subtle respects the appreciation of foreign styles and hence the emergence of new local ones; psychologists have found that the Zulus of South Africa, for instance, who live in a "circular culture" of windowless round huts and meandering paths, are relatively immune to visual illusions seen by people who have been conditioned by the right-angled, straight-perspective, so-called carpentered world of European and American cities.

Influence on appreciation

Religious styles. Conditioning also undoubtedly affects recognition of religious styles. The problem of hard-to-define "qualities" that contribute to recognizability, mentioned above in connection with regional and national styles, also returns here to vex a conscientious historian. So does the problem of the separation of form and content, plus the false problem of the sincerity of the artist as such—of the artist as mere "performer." Strictly speaking, in terms of forms and their combinations, one must grant that a vast number of profoundly moving works of religious art are not recognizably in religious styles. One can plausibly argue, for example, that there has been no religious style at all in Western painting since about the 15th century; later painters of religious subjects, such as Van Eyck, Raphael, Rubens, and their successors, painted the Virgin much as they painted their wives and mistresses. A similar point can be made about Western post-Renaissance music; Bach's sacred works sound much like his secular ones, and Giuseppe Verdi's magnificent *Requiem* (1874) has sounded to many ears like his operas. Even in the European Middle Ages and in the worlds of the great Eastern religions, the evidence is not always clear; the style that seems to yearn toward God in the 13th-century Gothic cathedral at Chartres, France, served also for town halls and ivory combs.

The point, of course, should not be exaggerated, partly because other styles often make use of borrowed forms and principally because many religious stylistic elements do exist in their own right, even though their secular ancestry can sometimes be traced. In architecture, examples of religious style can be seen in the basilican plan

Style conditioned by content

of Christian churches, the cosmic-mountain form of Hindu temples, the rectangular and cruciform plans of mosques, the needle shape of the minaret; in music, the single vocal line and free rhythm of Gregorian chant, the florid melody of Islāmic chant; in painting, the free brushwork of Zen Buddhists, the abstract arabesques of Islām, the nonillusionistic kinds of pictorial space favoured by medieval Christians. But when the list of such elements is completed, the fact still remains that the average appreciator recognizes a religious style primarily because it is tinged by long association with religious texts, ritual, and iconography. In brief, it has for him a quality that seems to emanate from content.

Period styles. A style belonging to one of the single-culture varieties may be divided, conventionally or arbitrarily, into periods. Beethoven's personal style is usually split into early, middle, and late; the ethnic style of the migrating Germanic peoples may be referred to as Animal I, II, and III. The adjective "period" is often reserved, however, for a distinct variety of style: the one in which the metaphorical "windows" afford "views" inhabited by generations whose cultural and other activities appear to constitute definable units of history. Familiar examples are the Gothic period style, the Renaissance, and, in general, all styles that bear the names of rulers or dynasties: the Victorian style, the Carolingian, the Sung. Frequently the period style is itself "periodized," the Gothic is Early and Late, the Renaissance is Early and High. The issues raised by periodization of all sorts will be discussed in the last section of this article (see below *III. The dynamics of style*). But period styles have their place in this part the classification, for they are clearly, often emphatically, of the single-culture sort.

CROSS-CULTURAL VARIETIES

Uninhabited, unmetaphorical styles

Cross-cultural styles are usually uninhabited, in contrast to personal styles. They do not, in any event, call up into the foreground of the appreciator's imagination a John Donne, a Metaphysical school of versifiers, a Murasaki Shikibu in an ancient Japanese court, a Germanic tribe on the march, a succession of English stone carvers expressing their Englishness, a Baudelaire in an urban twilight, a medieval monk singing for God, or a group of eminent Victorians. In a sense, then, cross-cultural styles are relatively unmetaphorical; they tend to focus attention on themselves. They have modes of existence that lie outside of history, and they often are not so much aesthetic terms as general-utility adjectives: almost anything, and not just works of art, can be "classical," for instance, or "organic," or "abstract." Cross-cultural styles in art, however, are by no means without implications; they can evoke outlooks, contexts, methods, and professions.

Outlook styles. Any style, of course, can be said to express an outlook or an attitude along with whatever else it expresses. Michelangelo's personal style expresses the outlook of Michelangelo; the Biedermeier style in furniture and decoration expresses the outlook of the German and Austrian middle classes between the Congress of Vienna in 1814–15 and the Revolution of 1848. Certain cross-cultural styles, however, can be said to specialize in general outlooks and attitudes that recur everywhere in all arts in all eras. There is a Classical style—a deviation toward clear, logical, nobly impersonal, carefully proportioned forms—that is recognizable not only in works that are ordinarily labelled "Classical" or "Neoclassical," such as the sculpture of the Parthenon of 5th-century-BC Athens and the plays of Pierre Corneille of 17th-century France, but also in the 13th-century sculpture of Reims Cathedral, the 20th-century Symbolist poetry of Paul Valéry, the 18th-century English portraits of Sir Joshua Reynolds, a Japanese screen, a Chinese vase, and some buildings on Fifth Avenue in New York City. There is a Romantic style—a deviation toward restless, alogical, nobly personal, intensely expressive forms—that is recognizable not only in works of the Romantic period of the early 19th century, such as the music of Frédéric Chopin or a poem by Heinrich Heine, but also in the Italian baroque architecture of Francesco Borromini, the fantasy painting of Paul Klee, Shakespeare's *King Lear*, a Hindu stone relief, a Hellenistic mosaic, and an avant-garde dancer. The same sort of thing can be said about such styles as the realistic, the fantastic, the expressionistic, and the idealistic, each of which implies an attitude toward life and the world. And perhaps the list should be rounded out with the academic style, into which the other outlook styles tend to sink when they are reduced to a set of teachable rules.

Classical and Romantic styles

Since these styles cross every aesthetic, geographical, or temporal frontier, embrace common human attitudes, and exhibit strong polar aspects, they have tempted many critics into an effort to arrange them all into two contrasting groups: the classical and the romantic, the conservative and the liberal, the idealistic (here the classical joins forces with the romantic) and the realistic, the rational and the emotional, and so on. Such arrangements have the advantage of breaking up traditional divisions and freshening awareness along with the disadvantage of not exhausting the evidence.

Contextual styles. At first glance the outlook styles may seem capable of annexing the contextual, which include the traditional genres and kinds of art—such styles as the tragic, the comic, the satiric, the pastoral, the heroic, and the melodramatic. But here the outlooks, when they exist, are at one or two removes from those of real life and the real world. A tragic style evokes not a death in the family but a work by Sophocles, Shakespeare, or some other playwright; a pastoral style evokes not herdsmen in the Alps but swains in a mythical Arcadia; a heroic couplet evokes not a brave warrior but the practice of John Dryden and others of using units of two rhyming lines of iambic pentameter in their "heroic" drama. Moreover, when the context—the genre, or kind of art—is not literary, the contextual style may evoke nothing that can properly be called an outlook. The operatic style in Verdi's *Requiem*, cited above, is a contextual style; so is the landscape style, with its hints of land, horizon, and sky, that is sometimes recognizable, like a ghost from Dutch paintings of the 17th century, in modern abstract paintings. So are the sculptural styles recognizable in the bizarre 20th-century architecture of the Spaniard Antonio Gaudí, the painterly styles in the Italian baroque sculpture of Gian Lorenzo Bernini and in certain Cubist sculpture of the 20th century, the musical styles of the poetry of Edgar Allan Poe and Alfred Lord Tennyson.

Procedural styles. Procedural cross-cultural styles are similar to but not quite the same as the contextual. For the term "procedural" is here meant to refer to ways of making art, to varieties of *poiēsis*, that may be shared by several arts and, in fact, are often shared by all. Hence, when the point of view is that of appreciators, procedural styles are commonly called "formalistic." Familiar examples of such styles are the mimetic, the representational, the naturalistic, the abstract, the geometric, the organic, the linear, the optical, the haptical (or tactile), and the plastic; some of these can be regarded, not always profitably, as subvarieties of others, and all of them can be regarded as members of polar pairs.

Formalistic members of polar pairs

Traditionally, certain procedural styles are associated with certain arts, but they need not be. Mimetic, or representational, styles, for example, are most often associated with painting and sculpture (*e.g.*, Gilbert Stuart's lifelike 18th-century portraits of George Washington or a Sumerian animal statue of the 3rd millennium BC), but they can also be recognized in the resemblance of Frank Lloyd Wright's mid-20th-century Guggenheim Museum to a snail, in the similarities of Nikolay Rimsky-Korsakov's musical composition *The Flight of the Bumble Bee* to the buzzing of a bee, and in Tennyson's frequently quoted lines in *The Princess*:

> The moan of doves in immemorial elms,
> And murmuring of innumerable bees . . .

which employ onomatopoeia, the use of words that actually sound like what the words represent.

Similarly, linear styles can be recognized not only in such obvious places as a line engraving by the 16th-century German master Albrecht Dürer or in the linear pat-

terns of Plateresque ornament of 16th-century Spain but also in the vocal line of an Italian Renaissance madrigal or in a modern building by the American architectural firm of Skidmore, Owings, & Merrill.

Traditionally also certain procedural styles are associated with certain cross-cultural outlook styles, but again they need not be. Mimetic styles are accompanied by Classical in the serene 17th-century landscapes of Nicolas Poussin, by Romantic in the turbulent 19th-century seascapes of Turner, by Realistic in Caravaggio's works of 16th-century Italy; by fantastic in the 20th-century Surrealistic compositions of René Magritte, and by Expressionistic in the tortured figures of Matthias Grünewald's Isenheim Altarpiece of the early 16th century. Abstract styles in 20th-century painting are accompanied by Classical in Piet Mondrian's use of pure lines and rectilinear shapes, by Romantic in Wassily Kandinsky's watercolours of his Blaue Reiter period, by Realistic (of a sort) in Frank Stella's striped, "objective" works, by the fantastic in Yves Tanguy's Surrealistic vistas, and by Expressionistic in Willem de Kooning's often violent canvases. The point is worth some emphasis, for even the most careful critic is likely to slip into the error of assuming that the classical, the romantic, and especially the realistic styles call for eternally fixed methods of production.

Professional styles. When the context of a contextual style or the procedure of a procedural style is derived from somewhere outside the arts in a sufficiently recognizable way or form, the result may be called—at the risk of some confusion—a professional style. Every artist, of course, has a professional style insofar as he is indeed an artist. But the critical occasions for talking about an "art" style in a meaningful fashion are rare; they are likely to arise only when an emphatic distinction is needed between two levels of artistry or when, as during the 19th-century English Aesthetic movement, artists and allied nonartists are waging war against alleged Philistines. On the other hand, artistic styles derived from nonartistic professions have never been rare and have recently been plentiful. An engineering style is easily recognizable in Roman, Gothic, and modern architecture and also in several kinds of modern sculpture and painting. Many novelists, from the beginning of the genre, have borrowed the styles of historians, and many the styles of journalists; directors of fictional films have worked in documentary styles, composers of modern music in the styles of experimental science and electronic-age technology. Occasionally the stylistic trend among 20th-century artists toward outside professions has been noticeable enough to lead a few observers, always promptly challenged by others, to conclude that the traditional arts were running low on unmanipulated variables.

Styles from non-artistic professions

III. The dynamics of style

This article has considered styles more or less as a panorama in which all the features are contemporaneous with each other and with the appreciator. The approach, to borrow a term from modern linguistics, has been largely "synchronic"; and such an approach has evident advantages. It gives an explicator a chance to classify, to structuralize, without worrying overmuch about the randomness of history. More importantly, it corresponds in a sense with an appreciator's actual experience of works of art, which is, of course, always synchronic, always in the present; experientially speaking, the Gothic style of the 13th century is not a second older than the style of Skidmore, Owings, & Merrill of the 20th. The historical, or diachronic, approach can offer, however, its own set of advantages. Facts and speculation concerning styles as they occur or change over a period of time can have the same sort of value as political or military history, and in addition they can contribute greatly to the recognition pleasure that is basic in an appreciation of styles.

Difficulties of stylistic history

The pleasure, it should be said right away, is normally accompanied by some peculiarly daunting doubts. Like all history, stylistic history is at once a branch of knowledge that tries to explain significant past events, a chronological record of such events, and somehow the past events themselves, in all their intractable uniqueness. But whereas the members of the trinity can usually be kept reasonably distinct in political or military history, they can usually be found in a state of oneness in stylistic history; the beginning of the Renaissance style, for instance, is notoriously a blend of theoretical explanation, selective records, and such "events" as surviving paintings, poems, and buildings. Also, whereas elections and battles have their fixed times and places, many styles do not; romanticism can leave the ahistorical limbo of cross-cultural styles to become the single-culture 19th-century Romanticism and then suddenly materialize, to some historians, as a phase of late Greco-Roman single-culture Classicism. In sum, the emergent mode of existence to which styles are disposed may make impossible a clean distinction between conceptions and historical facts and may expose the most conscientious of art historians to the charge of merely manoeuvring with private abstractions.

Another difficulty, which was glanced at in connection with the measurable in style, can be mentioned again here. It is the familiar one of overspecialized interpretations presented as general theory. Much of the terminology and nearly all of the chronological framework for the history of style are inventions of specialists in the study of the visual arts. Fortunately, the terminology and the framework can be adapted, with the help of some stretched meanings, to the study of music and literature. In what follows, a strong possibility of visual-art bias should be kept in mind during the discussion, which will concern origins, diffusion, and change and duration of art styles.

HISTORICAL ORIGINS

The question of the origins of styles has already been raised. From an appreciator's strictly synchronic point of view, to raise such a question is often to be irrelevant, or to get into a round of tautologies and to confuse the merely relational with the causal. The origins of the courtly style, of the Eskimo style, and of personal styles are by definition courts, Eskimo, and artists, respectively. There is something to be said, however, for the elaborated tautology as a form of analysis and also for the fact that adding the space-time of history makes a difference: it is one thing to point to an emphasis on order in the French national style and another thing to account for the advent of the Gothic style at the abbey of Saint-Denis in 1140. Moreover, art historians do not usually mean anything as determinative as a cause when they refer to the origins of a style; they merely mean a set of conditioning historical factors. And when they disagree they usually do so over the relative importance to be assigned to one of four sorts of factors: the politico-economic, the cultural, the technical, and the artistic.

The relational and the causal

Politico-economic factors. The most evident sort is the politico-economic. If the ancient Greeks had lost the Persian Wars in the 5th century BC, the columns of the temples on the Acropolis would probably be taller. If the Great Depression in the United States had begun earlier, the Empire State Building and other skyscrapers in New York of the 1920s and 1930s would probably have been shorter. The styles of Chinese blue and white porcelain were perceptibly influenced by export possibilities. Films are an industry as well as an art. The Italian Renaissance cannot be fully understood without references to the personality cults of monarchs, petty tyrants, and usurping mercenary leaders, to the capital accumulated in the bank of the Medici family of Florence and the coffers of the Vatican, to the expansion of international trade, and to the loosening of feudal ties. Such examples, along with more subtle ones, can be multiplied into a solidly factual kind of stylistic history.

Cultural factors. Nothing quite so solid is likely to come out of a consideration of contributing cultural factors; here one is obliged to weigh such imponderables as changes in manners, morals, psychologies, philosophies, and that pervasive sensibility system known as "the spirit of the age" and then to try to match these changes with the arrival times and the formal elements of presumably new styles. Thus it can be argued that the new period

The question of the "spirit of the age"

style gradually apparent at the close of the Greco-Roman era—the lack of naturalism, for instance, in painting—reflects Christian otherworldliness. Similarly, it might be claimed that the exuberantly enumerative prose style of the French satirist François Rabelais has roots in the widespread 16th-century zest for learning and discovery. To take still another example, the development of the orchestral crescendo in the middle of the 18th century, principally at Mannheim, Germany, has been said to reflect a new kind of human self-assertion, one destined for importance in all the arts with the arrival of 19th-century Romanticism.

Technical factors. The presence of politico-economic or of cultural factors need not, of course, rule out the technical sort, and examples of the latter are plentiful. The stylistic change from 14th-century to 15th-century European painting coincides with a new utilization, although not the invention, of the oil technique. The refinement of volumes and of implied movement in Renaissance bronze sculpture is linked to a new skill in casting techniques acquired in the making of artillery. Chopin's personal style and significant improvements in the piano both date from the 1830s. Sometimes the technical factors are partly disguised by a change of material; the Doric order in architecture is a wood style in stone. Sometimes, too, the technical factors must be understood in a large sense that embraces the functional, and they may then be accompanied by moralizing; thus, some 20th-century architects have maintained that it is "honesty" to let a building reveal its construction and use. Usually more than one set of technical factors can be discovered; a poet's style that mixes true rhyme with eye rhyme (similarity only of spellings) has origins—rather remote—both in ancient oral mnemonics and in the silent reading fostered by printing.

Artistic factors. Explanations of stylistic origins involving politico-economic, cultural, and technical factors tend to have in common the defect of not providing very exclusive matchings of formal elements with supposed sources. Many times and places have had ambitious princes and accumulated capital equal to those of 16th-century Italy without producing anything like the smoky style, the sfumato, of Leonardo da Vinci's paintings, or the clarity and solemnity of Bramante's architecture. If a lack of naturalism goes well with early Christian otherworldliness, it also seems to go somehow with the reputed thisworldliness of the 20th century. If the perfected piano can be matched with the smooth 19th-century compositions of Chopin, it can also be matched with the staccato 20th-century music of Béla Bartók. To be sure, these are only contributing factors, not specific causes. Even so, the versatility of the factors mentioned usually leads an art historian to add some relatively direct, purely artistic factors and eventually to postulate a group of ancestors and influential cousins for a given style. Thus Leonardo's style is said to be derived partly from his early master Verrocchio's, early Christian from ancient Roman and Syrian, Chopin's from the styles of other Polish pianists and from the Irish-born John Field. Art historians of this persuasion do not, as is sometimes alleged, reject the hypothesis that an artist may think up a style partly on his own; they merely regard it as lying outside the concerns of scholarship.

Matching causes and effects

Here another difficulty in the writing of stylistic history has to be mentioned. In which painting did Leonardo's personal style become his own and not Verrocchio's? At what date did the early Christian style cease to be Roman and Syrian? Questions of this sort, if pressed far enough, can raise the suspicion that a discussion supposedly about the origins of a certain style is actually about the arbitrary periodization of a more inclusive style.

DIFFUSION

The factor of place in diffusion

Periodization is always, although the fact is not always made clear by periodizers, a spatiotemporal operation; it poses questions of where as well as of when. The Gothic period can be said, if dates are used with convenient recklessness, to end around 1420 in Italy, around 1530 in France, and a generation later in England. Moreover, the "where" may be not only a point in geographical space but also a zone in the imaginary space constituted by any classification of the arts. The Italian Renaissance can be said to begin in painting around 1300 with Giotto; in poetry around 1330 with Petrarch; in architecture around 1420 with Brunelleschi; and in music around 1525 with madrigal composers, or perhaps around 1600 with Monteverdi. In sum, staggered dates, distinct localities, and the separation of the arts combine to involve the student of period styles in a complex problem of assumed or alleged correspondences and sometimes in unexpected variants of the problem of stylistic origins.

Correspondences in real space. Like their colleagues in other branches of cultural history, art historians sometimes are obliged to choose between diffusionism and theories of spontaneous development; the striking similarity between a motif on a Chinese bronze and one on an Aztec temple, for instance, may be attributed to prehistoric migration, to analogous stages of civilization, or to chance. Normally, however, and for good reasons, art historians rely on diffusion from a creative centre in their attempts to explain stylistic correspondences in even widely separated areas; and they rely on it almost exclusively in discussing period styles. It is fairly easy to trace the movement of Andrea Palladio's style from the buildings he designed in Vicenza, Italy, in the 16th century to Inigo Jones's Banqueting House in London in the 17th century, of opera buffa from Naples to Vienna in the 18th century, or of Cubism from Paris to Prague and New York in the 20th century. A difficulty in this sort of history is that diffusion implies receptivity, and stylistic receptivity implies at least a degree of spontaneous development.

Correspondences in the arts. There are a number of problems in attempts to deal with stylistic correspondences in different arts. One of the most common is the difficulty of making a distinction between a period style and a period of time when the same label is used for both —and perhaps for each in more than one sense. Thus, a historian who refers to Baroque music without explanation may mean that the music in question has certain stylistic affinities with the theatricality of Bernini's sculpture, with the full-blooded movement and illusionism of Pietro da Cortona's painting, with the undulations of Borromini's architecture, and with the cadences of Sir Thomas Browne's prose; instead, or also, he may mean, as many musicologists would, that the music carries the technical eramark of a basso continuo part; again he may mean that the music evokes the manners, public life, science, and general culture of "the Age of the Baroque"; and finally he may mean nothing more than that the piece was composed in the 17th century. If he does mean that the music has stylistic affinities with Baroque sculpture, painting, architecture, and prose, then other difficulties appear. He must manage to translate each art into the others, construct a convincing historical pattern, and explain the correspondences.

Problems and ambiguities of correspondences

Nevertheless, there is considerable agreement among art historians that such correspondences do exist—not in every period but often enough to merit attention. In addition to those of the Baroque, which are frequently cited, one can mention those between musical and visual-art proportions during the Renaissance, those among nearly all the arts during the Gothic period, and those among painting, sculpture, architecture, and music during the 20th century. Persuasive arguments for correspondences between art and other disciplines have been advanced, notably in a comparison of Gothic architecture with scholasticism, the contemporaneous church-oriented philosophy, and by several commentators in comparisons of Leibniz's highly complex but integrated philosophy to Baroque art.

CHANGE AND DURATION

The already mentioned problem of distinguishing between the origins of a distinct style and the periodization of a more inclusive one complicates every discussion of stylistic change and duration. It can be argued that whenever a style moves into a new period or phase it becomes

a new style; and a few modern historians have drawn the conclusion that the concept of style should be kept out of art history and be reserved for synchronic structuring and critical analysis. They feel that one can usefully take a cross section of Baroque, for example, and study it horizontally, out of time, but that to talk about the historical antecedents of Baroque, or about phases of Baroque, is to fall into logical contradiction. Such historians, however, are very much a minority. Most of their colleagues are committed not only to stylistic change and duration but also, if often only implicitly, to various theories about supposedly typical stages in the stylistic historical process.

Cyclical theories. In the preface to his *Lives*, Vasari remarked that the arts, like human beings, "are born, grow up, become old, and die." Four centuries later, the infinitely more knowledgeable and sophisticated art historian Henri Focillon entitled an essay on style *The Life of Forms in Art* (1934). André Malraux, in his *Voices of Silence* (1951), refers to

> these imaginary super-artists we call styles, each of which has an obscure birth, an adventurous life, including both triumphs and surrenders to the lure of the gaudy or the meretricious, a death-agony and a resurrection.

References to the "maturity" or the "decay" of styles are part of nearly everybody's art criticism. In sum, the biological metaphor for style has become commonplace. But it once was a serious theory of the historical process, and as such it perhaps still deserves some denunciation for having been frequently both a misrepresentation of facts and a source of prejudice against estimable works of art —Hellenistic, late Gothic, Mannerist, Baroque, and so on —which were believed to belong to the "decadent" part of "the life cycle" of a style.

In an effort to get closer to the observed course of events, some historians have favoured, in whole or in part, a cyclical theory that pictures each major single-culture style as going through the same irreversible series of cross-cultural styles, which is usually given as archaic, classic, baroque, impressionist, and archaistic. Thus the first, or archaic, phase of the Graeco-Roman cycle is supposed to correspond with the first phase of the Gothic and of the Renaissance cycles; and one can speak of third-phase baroque Greek, third-phase baroque Gothic, even third-phase baroque Baroque. The theory has some fascination, and occasionally it works well enough to shed light on the actual historical behaviour of styles. But often it works only because one has been careful to start the cycle in the right place, and far too often it does not work at all.

Dialectical theories. The notion that history is a pendulum, with left inevitably followed by right, occurs to everyone, and it is especially likely to occur to art historians because of the polar aspects of style. Sometimes the pendulum is imagined as swinging forever between cross-cultural outlook styles: classical then romantic, idealistic then realistic. Historians who are inclined to feel that outlook styles are too vague and too laden with value implications may favour the idea of swings between cross-cultural procedural styles: mimetic then abstract, organic then geometric. The influential theory devised by the Swiss art historian Heinrich Wölfflin (1864–1945) postulates an elaborate dialectic involving pairs such as linear–painterly and closed–open. A theory devised by the German critic Paul Frankl (1878–1962) combines a movement between styles of being and becoming with a cyclical development through preclassic, classic, and post-classic stages. Another theory, devised by the Austrian scholar Alois Riegl (1858–1905), postulates cycles of evolution within a long swing across the centuries from an early haptic to a later optic phase. All these models of the stylistic historical process break down eventually, but they can provide useful categories for organizing the flux of artistic events.

The pendulum notion of history

Sequential theories. Historians who refuse to believe that styles are haunted by destiny may be willing to grant that within short periods and mostly in terms of technique something that looks like a preordained evolution can occur. Thus it is possible to discern in Gothic architecture a predictable movement toward lighter construction, in Early Renaissance painting a movement toward a desired illusionism, in 16th-century European music a movement toward the tonal system that emerged in the following century. But beyond this concession, such historians are likely to rely on simple sequences devoid of any trace of determinism. Their works often consist of a sequence of biographies of artists and a sequence of centuries. Do such methods suggest mere chronicling rather than history? Perhaps partly to answer such a question, a theory advanced by the contemporary American historian James S. Ackerman presents stylistic change as a "process" of a sort, although not one that is preordained; art history is envisaged as a series of steps away from the past, but not toward the future. "Each step," Ackerman says, "for the artist who takes it, is final and definitive; he cannot consciously make a transition to a succeeding step, for if he visualizes something he regards as preferable to what he is doing, he presumably will proceed to do it. . . ."

Satiation theories. Why does the artist take a step away from the past? A common explanation is that he, or his audience, has become overly familiar with the prevailing style. The phenomenon is similar to what psychologists call semantic satiation—the loss of meaningfulness in repeated words. According to one theory, stylistic satiation is hastened by comparison; thus when much was changing at the beginning of the 20th century, many painters felt that their old styles were intolerably familiar. According to another theory, satiation is always merely temporary. To revert to part of the biological metaphor, styles are born, but they never really die.

BIBLIOGRAPHY

Theoretical surveys: An excellent critical résumé of early 20th-century speculation, with emphasis on the visual arts, is M. SCHAPIRO, "Style," in A.L. KROEBER (ed.), *Anthropology Today* (1953). J.S. ACKERMAN, "Theory of Style," *Journal of Aesthetics and Art Criticism*, 20:227–237 (1962), stresses relational values and offers a lucid critique of evolutionary notions. R.L. SCRANTON, *Aesthetic Aspects of Ancient Art* (1964), analyzes in sharp detail the structure of style and applies his conclusions to specific works, including literature. HERBERT READ, in the introduction to *The Styles of European Art* (1965), emphasizes the psychological aspects of the subject. RENE WELLEK and AUSTIN WARREN, *Theory of Literature*, 3rd rev. ed. (1966), have much to say about style in passing, from the standpoint of Anglo-American New Criticism. GEORGE KUBLER, *The Shape of Time* (1962), wages brilliant war on conventional art history and presents a formal substitute for the concept of style. BRUCE ALLSOPP, *Style in the Visual Arts* (1956), reviews the history of the concept and stresses the allegedly damaging effects of style on contemporary artistic activity.

Special studies: ERWIN PANOFSKY, *Renaissance and Renascences in Western Art*, 2nd ed., 2 vol. (1965), examines the idea of rebirth in Italy and discusses the significance of medieval renascences; his *Meaning in the Visual Arts* (1955), although focussed on iconography, has extensive reflections on style; his *Gothic Architecture and Scholasticism* (1951) is a study in correspondence. ETIENNE SOURIAU, *La Correspondance des Arts* (1947), presents a basis for comparative aesthetics. WYLIE SYPHER, *Four Stages of Renaissance Style* (1955) and *Rococo to Cubism in Art and Literature* (1960), make good use of stylistic polarities in an analysis of painting, sculpture, architecture, and literature. RUDOLF WITTKOWER, *Architectural Principles in the Age of Humanism*, 3rd ed. rev. (1962), deals in depth with problems of symbolism, optics, and harmony in Renaissance buildings. NIKOLAUS PEVSNER, *The Englishness of English Art* (1956), offers one of the few instances of a serious historian taking the idea of a national style seriously. G.M.A. GRUBE, *The Greek and Roman Critics* (1965), discusses stylistic theories from Homer to the 3rd century AD. WALTER WIORA, *Die vier Weltalter der Musik* (1961; Eng. trans., *The Four Ages of Music*, 1965), presents a panorama of the world's musical styles from ancient times to the 20th century.

Classical studies: FRANZ BOAS, *Primitive Art* (1927, new ed., 1962); H. FOCILLON, *Vie des formes* (1934; Eng. trans., *The Life of Forms in Art*, 2nd ed., 1948); P. FRANKL, *Das System der Kunstwissenschaft* (1938); A. RIEGL, *Stilfragen: Grundlegungen zu einer Geschichte der Ornamentik* (1893; 2nd ed., 1923); HEINRICH WOLFFLIN, *Renaissance und Barock* (1888; Eng. trans., *Renaissance and Baroque*, 1964); *Die klassische Kunst* (1899; Eng. trans., *Classic Art*, 1952, reprinted 1968);

and *Kunstgeschichtliche Grundbegriffe* (1915; Eng. trans., *Principles of Art History*, 1932).

Bibliographies: Useful short lists of theoretical works on style may be found in the above-mentioned publications of Schapiro and Scranton. Longer lists, mixed with other material, may be found in those of Panofsky, Sypher, and Wellek and Warren.

(R.McMu.)

Asceticism

Asceticism, in religion, is the practice of the denial of physical or psychological desires in order to attain a spiritual ideal or goal. The term and its cognates are derived from the Greek verb *askeō*, which originally meant to manufacture technically or artistically and later to exercise or to train. The noun *askēsis* means "exercise" and "training." Hardly any religion has been without at least traces or some features of asceticism. In the history of religions and of civilization, the extent to which the practice of asceticism has enabled men to develop their inner (spiritual) powers over their instinctive urges and the influences of the external world has been notable. The practitioners of asceticism believe that the ascetic person has access to powers that uphold moral standards and enliven spirituality.

The origins of asceticism. *Asceticism in athletics, intellectual endeavours, ethics, and technological endeavours.* The origins of asceticism lie in man's attempts to achieve various ultimate goals or ideals: development of the "whole" person, human creativity, ideas, the "self," or skills demanding technical proficiency. Athletic *askēsis*, involving the ideal of bodily fitness and excellence, was developed to ensure the highest possible degree of physical fitness in an athlete. Among the ancient Greeks, athletes preparing for physical contests (*e.g.*, the Olympic Games) disciplined their bodies by abstaining from various normal pleasures and by enduring difficult physical tests. In order to achieve a high proficiency in the skills of warfare, warriors also adopted various ascetical practices. The ancient Israelites, for example, abstained from sexual intercourse before going into battle.

Goals of asceticism

As values other than those concerned with physical proficiency were developed, the concept expressed by *askēsis* and its cognates was applied to other ideals—*e.g.*, mental facility, moral vitality, and spiritual ability. The ideal of training for a physical goal was converted to that of attaining wisdom or mental prowess by developing and training intellectual faculties. Among the Greeks such training of the intellect led to the pedagogical system of the Sophists—itinerant teachers, writers, and lecturers of the 5th and 4th centuries BC who instructed in return for fees. Another change in the concept of *askēsis* occurred in ancient Greece when the notion of such training was applied to the realm of ethics in the ideal of the sage who is able to act freely to choose or refuse a desired object or an act of physical pleasure. This kind of *askēsis*, involving training the will against a life of sensual pleasure, was exemplified by the Stoics (ancient Greek philosophers who advocated the control of the emotions by reason). The goal of such moral and volitional asceticism was viewed as enhancement of the dignity of man. These ideas reveal the various shades of meaning of the word *askēsis*. They range from the negatively oriented, in which the object of *askēsis* is the avoidance of evil, the suppression of vicious tendencies, or the moderation of excessive passions, to positively oriented trends, in which the object is the strengthening of the virtues that perfect moral and spiritual life and that enable one to cultivate inner powers.

The view that one ought to deny one's lower desires—understood as sensuous, or bodily—in contrast with one's spiritual desires and virtuous aspirations, became a central principle in ethical thought. Plato believed that it is necessary to suppress bodily desires so that the soul can be free to search for knowledge. This view was also propounded by Plotinus, a Greek philosopher of the 3rd century AD and one of the founders of Neoplatonism, a philosophy concerned with hierarchical levels of reality. The Stoics, among whom asceticism was primarily a discipline to achieve control over the promptings of the emotions, upheld the dignity of human nature and the wise man's necessary imperturbability, which they believed would become possible through the suppression of the affective, or appetitive, part of man. In a similar manner, the value of asceticism in strengthening man's will and his deeper spiritual powers has been a part of many religions and philosophies throughout history. The 19th-century German philosopher Arthur Schopenhauer, for example, advocated a type of asceticism that annihilates the will to live; his fellow countryman and earlier contemporary philosopher Immanuel Kant held to a moral asceticism for the cultivation of virtue according to the maxims of the Stoics. In 20th-century society, in the face of criticism that the mechanical, commercial, and technological equipment of modern life—with a momentum and enticement and stimuli of the senses all its own—has enslaved men, critics, such as the Roman Catholic theologian Romano Guardini, have advocated asceticism as a means of rescuing human existence. Man, having lost his identity because of the effects of the growing technological process upon him, must once more renounce the world in order to regain his spiritual position. And, according to Arnold Gehlin, the possibilities that the spiritual discipline of asceticism offers are to be understood as important aids in humanization.

Asceticism in modern societies

Asceticism in religion. The concept of *askēsis* as a means of achieving various spiritual goals was also applied to religion, in which it assumed a significant position. Many factors were operative in the rise and cultivation of religious asceticism: the fear of hostile influences from the demons; the view that one must be in a state of ritual purity as a necessary condition for entering into communion with the supernatural; the desire to invite the attention of divine or sacred beings to the self-denial being practiced by their suppliants; the idea of earning pity, compassion, and salvation by merit because of self-inflicted acts of ascetical practices; the sense of guilt and sin that prompts the need for atonement; the view that asceticism is a means to gain access to supernatural powers; and the power of dualistic concepts that have been at the source of efforts to free the spiritual part of man from the defilement of the body and physically oriented living. Among the higher religions (*e.g.*, Hinduism, Buddhism, and Christianity), still other factors became significant in the rise and cultivation of asceticism. These include the realization of the transitoriness of earthly life, which prompts a desire to anchor one's hope in otherworldliness, and the reaction against secularization that is often coupled with a belief that spirituality can best be preserved by simplifying one's mode of life.

Types of religious asceticism. Though the demarcation lines seldom can be drawn clearly and many elements mingle and overlap, certain types of asceticism are still distinguishable: ritual, disciplinary (physical and moral), and mystical. According to an agelong idea that has dominated primitive religions, engagement in the profane world is believed to be the antithesis of a relationship to the sacred realm, a relationship that can be recovered by ritual sacrifices. Survivals of ritual asceticism (*e.g.*, fasting) are reflected in many of the higher religions.

Disciplinary asceticism, with its goal of training the soul, is graphically depicted in various techniques aimed at accomplishing this goal. The methods of *yoga* (a Hindu psychological–physical meditation system) consist of injunctions involving the regulation of breathing, posture, sitting, and meditation. In Buddhism this system is further developed into what is known as the Eightfold Path, a kind of guide to moral or righteous living that may be practiced by both monks and laymen. In Christianity, the spiritual exercises taught by St. Ignatius of Loyola, the 16th-century founder of the Roman Catholic monastic order known as the Jesuits, are designed to actuate the will toward the pursuit of moral and spiritual perfection. Hyperasceticism (excessive disciplinary asceticism), in its ultimate goal, goes beyond the training of the soul and ends in the ultimate annihilation of the body so that the soul may become free. This stems either

Influence of dualistic ideas

from metaphysical dualism, which separates the material from the spiritual (*e.g.*, the body from the soul) into sharply contrasted realities, or from quasi-dualistic views. Metaphysical dualism is contained in the thought of Plotinus, in Neoplatonism, and in Gnosticism, a Christian movement that taught that matter is evil and the spirit good. It is also found in the teachings of Marcion, a 2nd-century Christian who founded his own church, and Mani, a 3rd-century founder of a dualistic religion that bears his name. Quasi-dualistic ideas that produce a hatred of the body are found, for example, in Syrian Christian asceticism, which pressed relentlessly toward a life of mortification, along with an insatiable longing for the accomplishment of extravagant aims. Turning farther toward the East, one finds that Siddhārtha Gautama, the 6th–5th-century-BC founder of Buddhism, rejected asceticism, but some Buddhists have practiced disciplinary asceticism. An extreme example is found in the legend of Hui-k'o, a student of the 6th-century Japanese Zen Buddhist guru (teacher) Bodhidharma, who cut off his arm in the presence of the master to demonstrate the sincerity of his desire to study the *dharma* (law) with Bodhidharma. Gorakhnāth, medieval founder of the Hindu order of religious ascetics, the Kānphaṭa Yogis, belonged to the tradition common to both Hinduism and Buddhism of the 84 *mahasiddhas*, or "great perfect ones," who acquired magical powers through the practice of rigorous forms of spiritual discipline.

An extreme form of dying to the world is practiced by some members of Jainism in India who, in their attempts to reach sainthood, starve themselves to death.

The goal of mystical asceticism is union with the divine —the mystical experience. By means of dying to the world, of extinguishing the self and the will, and by traversing through various stages in the purification of the soul, one climbs up the ladder of mystical ascent toward the divine union. Among the Syrian Christians, who greatly influenced Ṣūfism, an Islāmic form of mysticism, mystical asceticism is graphically described in a work by Stephanos bar Ṣūdailē, 5th–6th-century monk from Edessa, entitled *Hierotheos* ("Holy God").

For Christians, still another type of asceticism, stemming from the desire to experience the sufferings of Christ, was exemplified by the emergence in the Middle Ages of the beggar-monks. Devotion to Christ and his sufferings has enabled these monks (and others) to enter into the experience of Christ's sorrows with new insights, such as finding a Christ in all who suffer. In this instance, asceticism has taken on a new meaning—*i.e.*, to serve Christ by becoming united with all whose lives are filled with pain and sorrow and by entering the service of compassion and redemption.

Forms of religious asceticism. *Celibacy and denial of material goods.* In all strictly ascetic movements, celibacy has been regarded as the first commandment (see also CELIBACY). Virgins and celibates emerged among the earliest Christian communities and came to occupy a prominent status. Among the earliest Mesopotamian Christian communities, only the celibates were accepted as full members of the church, and in some religions only celibates are permitted to be priests (*e.g.*, Aztec religion and Roman Catholicism). Abdication of worldly goods is another fundamental principle. In monastic communities there has been a strong trend toward this ideal. In Christian monasticism this ideal was enacted in its most radical form by Alexander Akoimetos, a founder of monasteries in Mesopotamia (died *c.* 430). Centuries before the activities of the medieval Western Christian monk St. Francis of Assisi, Alexander betrothed himself to poverty, and, through his disciples he expanded his influence in Eastern Christian monasteries. These monks lived from the alms they begged but did not allow the gifts to accumulate and create a housekeeping problem, as occurred among some Western monastic orders, such as the Franciscans. In the East, wandering Hindu ascetics and Buddhist monks also live according to regulations that prescribe a denial of worldly goods.

Abstinence and fasting. Abstinence and fasting are by far the most common of all ascetic practices. Among the primitive peoples, it originated, in part, because of a belief that taking food is dangerous, for demonic forces may enter the body while one is eating. Further, some foods regarded as especially dangerous were to be avoided. Fasting connected with religious festivals has very ancient roots (see also DIETARY LAWS AND CUSTOMS). In ancient Greek religion, rejection of meat appeared particularly among the Orphics, a mystical, vegetarian cult; in the cult of Dionysus, the orgiastic god of wine; and among the Pythagoreans, a mystical, numerological cult. Among a number of churches the most important period of fasting in the liturgical year is the 40 days before Easter (Lent), and among Muslims the most important period of fasting is the month of Ramadān. The ordinary fasting cycles, however, did not satisfy the needs of ascetics, who therefore created their own traditions. Among Jewish-Christian circles and Gnostic movements, various regulations regarding the use of vegetarian food were established, and Manichaean monks won general admiration for the intensity of their fasting achievements. Christian authors write of their ruthless and unrelenting fasting, and, between their own monks and the Manichaeans, only the Syrian ascetical virtuosos could offer competition in the practice of asceticism. Everything that could reduce sleep and make the resultant short period of rest as troublesome as possible was tried by Syrian ascetics. In their monasteries Syrian monks tied ropes around their abdomens and were then hung in an awkard position, and some were tied to standing posts.

Personal hygiene and limitation of location. Personal hygiene also fell under condemnation among ascetics. In the dust of the deserts—where many ascetics made their abodes—and in the blaze of the Oriental sunshine, the abdication of washing was equated with a form of asceticism that was painful to the body. With respect to the prohibition against washing, the Persian prophet Mani seems to have been influenced by those ascetic figures who had been seen since ancient times in India, walking around with their long hair hanging in wild abandonment and dressed in filthy rags, never cutting their fingernails and allowing dirt and dust to accumulate on their bodies. Another ascetic practice, the reduction of movement, was especially popular among the Syrian monks, who were fond of complete seclusion in a cell. The practice of restriction in regard to contact with human beings culminated in solitary confinement in wildernesses, cliffs, frontier areas of the desert, and mountains. In general, any settled dwelling place has been unacceptable to the ascetic mentality, as noted in ascetical movements in many religions.

Psychological and pain-producing asceticism. Psychological forms of asceticism have also been developed. A technique of pain-causing introspection was used by Buddhist ascetics in connection with their practices for meditation. The Syrian Christian theologian St. Ephraem Syrus counselled the monks that meditation on guilt, sin, death, and punishment—*i.e.*, the pre-enactment of the moment before the Eternal Judge—must be carried out with such ardour that the inner life becomes a burning lava that produces an upheaval of the soul and torment of the heart. Syrian monks striving for higher goals created a psychological atmosphere in which continued fear and dread, methodically cultivated, were expected to produce continual tears. Nothing less than extreme self-mortification satisfied the ascetic virtuosos.

Role of meditation

Pain-producing asceticism has appeared in many forms. A popular custom was to undergo certain physically exhausting or painful exercises. The phenomena of cold and heat provided opportunities for such experiences. The Hindu fakirs (ascetics) of India provide most remarkable examples of those seeking painful forms of asceticism. In the earliest examples of such radical forms of self-mortification that appeared in India, the ascetic stared at the sun until he went blind or held up his arms above the head until they withered. Syrian Christian monasticism was also inventive in regard to forms of self-torture. A highly regarded custom involved the use of iron devices, such as girdles or chains, placed around the loins, neck, hands, and feet and often hidden under

garments. Pain-producing forms of asceticism include self-laceration, particularly castration, and flagellation (whipping), which emerged as a mass movement in Italy and Germany during the Middle Ages and is still practiced in parts of Mexico and the southwestern United States.

Variations of asceticism in world religions. *Primitive and ancient religions.* In the primitive religions, asceticism in the form of seclusion, physical discipline, and the quality and quantity of food prescribed has played an important role in connection with the puberty rites and rituals of admission to the tribal community. Isolation for shorter or longer periods of time and other acts of asceticism have been imposed on medicine men, since severe self-discipline is regarded as the chief way leading to the control of occult powers. Isolation was and is practiced by young men about to achieve the status of manhood in the Blackfoot and other Indian tribes of the northwestern United States. In connection with important occasions, such as funerals and war, taboos (negative restrictive injunctions) involving abstinence from certain food and cohabitation were imposed. For the priests and chiefs these were much stricter. In Hellenistic culture (*c.* 300 BC–*c.* AD 300), asceticism in the form of fasting and refraining from sexual intercourse was practiced by communities of a religiomystical character, including the Orphics and Pythagoreans. A new impetus and fresh approach to ascetic practices (including emasculation) came with the expansion of the Oriental mystery religions (such as the cult of the Great Mother) in the Mediterranean area.

Eastern religions. In India, in the late Vedic period (*c.* 1500 BC–*c.* 200 BC), the ascetic use of *tapas* ("heat," or austerity) became associated with meditation and *yoga*, inspired by the idea that *tapas* kills sin. These practices were embedded in the Brahmanic (ritualistic Hindu) religion in the *Upaniṣads* (philosophical treatises), and this view of *tapas* gained in importance among the Yogas and the Jainas, adherents of a religion of austerity that broke away from Brahmanic Hinduism. According to Jainism, liberation becomes possible only when all passions have been exterminated. Under the influences of such ascetic views and practices in India, Siddhārtha Gautama himself (see above) underwent the experiences of bodily self-mortification in order to obtain spiritual benefits; but since his expectations were not fulfilled, he abandoned them. But his basic tenet, which held that suffering lies in causal relation with desires, promoted asceticism in Buddhism. The portrait of the Buddhist monk as depicted in the *Vinaya* (a collection of monastic regulations) is of one who avoids extreme asceticism in his self-discipline. The kind of monasticism that developed in Hinduism during the medieval period also was moderate. Asceticism generally has no significant place in the indigenous religions of China (Confucianism and Taoism) and Japan (Shintō). Only the priests in Confucianism practiced discipline and abstinence from certain foods during certain periods, and some movements within Taoism observed similar rules. Shintō in Japan, however, does include ascetics.

Western religions. Judaism, because of its view that God created the world and that the world (including man) is good, is nonascetic in character and includes only certain ascetic features, such as fasting for strengthening the efficacy of prayer and for gaining merit. Though some saw a proof of the holiness of life in some ascetic practices, a fully developed ascetical system of life has remained foreign to Jewish thought, and ascetic trends could, therefore, appear only on the periphery of Judaism. Such undercurrents rose to surface among the Essenes, a monastic sect associated with the Dead Sea Scrolls, who represented a kind of religious order practicing celibacy, poverty, and obedience. The archaeological discovery (1940s) of their community at Qumrān (near the Dead Sea in an area that was a part of Jordan) has thrown new light on such movements in Judaism.

In Zoroastrianism (founded by the Persian prophet Zoroaster, 7th century BC) there is officially no place for asceticism. In the Avesta, the sacred scriptures of Zoroastrianism, fasting and mortification are forbidden, but ascetics were not entirely absent even in Persia.

Significance for Christianity

In Christianity all of the types of asceticism have found realization. In the Gospels asceticism is never mentioned, but the theme of following the historical Christ gave asceticism a point of departure. An ascetic view of the Christian life is found in the First Letter of Paul to the Corinthians in his use of the image of the spiritual athlete who must constantly discipline and train himself in order to win the race. Abstinence, fasts, and vigils in general characterized the lives of the early Christians, but some ramifications of developing Christianity became radically ascetic. Some of these movements, such as the Encratites (an early ascetic sect), a primitive form of Syrian Christianity, and the followers of Marcion, played important roles in the history of early Christianity. During the first centuries ascetics stayed in their communities, assumed their role in the life of the church, and centred their views of asceticism on martyrdom and celibacy. Toward the end of the 3rd century, monasticism originated in Mesopotamia and Egypt and secured its permanent form in cenobitism (communal monasticism). After the establishment of Christianity as the official religion of the Roman Empire (after AD 313), monasticism was given a new impetus and spread all over the Western world. In Roman Catholicism new orders were founded on a large scale. Though asceticism was rejected by the leaders of the Protestant Reformation, certain forms of asceticism did emerge in Calvinism, Puritanism, Pietism, early Methodism, and the Oxford Movement (an Anglican movement of the 19th century espousing earlier ecclesiastical ideals). Related to asceticism is the Protestant work ethic, which consists of a radical requirement of accomplishment symbolized in achievement in one's profession and, at the same time, demanding strict renunciation of the enjoyment of material gains acquired legitimately.

Islām in its beginnings knew only fasting, which was obligatory in the month of Ramaḍān. Monasticism is rejected in the Qur'ān (the Islāmic sacred scripture). Yet ascetic forces among Christians in Syria and Mesopotamia, vigorous and conspicuous, were able to exercise their influence and were assimilated by Islām in the ascetic movement known as *zuhd* (self-denial) and later in that of Ṣūfism, a mystical movement that arose in the 8th century and incorporated ascetic ideals and methods.

BIBLIOGRAPHY. JAMES HASTINGS (ed.), "Asceticism," in the *Encyclopaedia of Religion and Ethics*, vol. 2 (1910); and "Askese," in the *Reallexikon für Antike und Christentum*, vol. 1 (1950), are two summary articles, the latter in German. OSCAR HARDMAN, *The Ideals of Asceticism* (1924), presents a treatment of wider scope. EDMOND DEMAITRE, *Fakirs et yogis des Indes* (1936; Eng. trans., *The Yogis of India*, 1937), offers a matter-of-fact treatment of the phenomenon. For Buddhist asceticism, the best work is G.C.A. EVOLA, *La dottrina del risveglio* (1943; Eng. trans., *The Doctrine of Awakening*, 1951); for Platonism, I.G. WHITCHURCH, *The Philosophical Bases of Asceticism in the Platonic Writings and in Pre-Platonic Tradition* (1923), is still usable. F. MARTINEZ, *L'Ascétisme chrétien pendant les trois premiers siècles de l'Église* (1913), presents a comprehensive account of Christian asceticism in early centuries, which may be supplemented by OWEN CHADWICK (ed.), *Western Asceticism* (1958), unfolding the phenomenon through the translation of selected classical original records. A. VOOBUS, *History of Asceticism in the Syrian Orient*, 2 vol. (1958–60), based on manuscript research, deals with the autochthonous origin of Syrian monasticism and its development in Syria, Mesopotamia, and Persia. For the later history of Christian asceticism, the best treatment is LOUIS GOUGAUD, *Dévotions et pratiques ascétiques du moyen âge* (1925; Eng. trans., *Devotional and Ascetic Practices in the Middle Ages*, 1927).

(A.V.)

Aschelminthes

The phylum Aschelminthes (also known as Nemathelminthes) comprises a diverse group of invertebrate animals. Adults possess a space—the pseudocoel—between the gut and the body wall, have bilateral symmetry (*i.e.*, a plane through the centre of the body divides it into mir-

ror-image halves), and lack division of the body into segments.

Many aschelminths are less than one millimetre (0.04 inch) long; some are less than half that length. The smallest forms, in fact, are no larger than certain protozoans, with which they were confused by early microscopists. The bodies of aschelminths, however, contain several complete organ systems and thus are more complex than the protozoans, which lack comparable organs.

The group contains about 17,000 described species, a large number of which are common in salt water or freshwater. Many are of economic importance as parasites in man and domestic animals. Adults are generally of academic interest because they demonstrate eutely; *i.e.*, they tend to be composed of a fixed number of cells.

GENERAL FEATURES

Size range and diversity of structure. Six classes of Aschelminthes are recognized: Nematoda, Rotifera, Gastrotricha, Kinorhyncha (or Echinodera), Nematomorpha (or Gordiacea), and Priapulida.

Nematodes, also called eelworms, pinworms, threadworms, and roundworms, range in length from about one millimetre (about 0.04 inch) to more than 20 centimetres (about eight inches), and their elongated bodies tend to be pointed at both ends. Rotifera, so-called wheel animalcules, are microscopic in size, usually from 100 to 500 microns (about 0.004 to 0.02 inch) long, and typically have a ciliated (*i.e.*, with ever-moving hairlike structures) disk at the anterior, or front, end. Gastrotrichs and kinorhynchs are minute, wormlike forms. Gastrotrichs are ciliated on the ventral, or lower, side. The kinorhynch has a spiny ringed body and a retractable head. Nematomorphs, or hairworms, are thin, elongated animals and attain lengths exceeding 80 centimetres (about 30 inches). The outer layer (cuticle) is thick and tough. Priapulids are warty and superficially (externally) ringed, or segmented; the largest are about eight centimetres (about three inches) long.

Distribution and abundance. Nematoda, the largest class, comprises some 13,000 species. Free-living species occur in salt water and freshwater and in sand and soil; parasitic forms are found in animals and plants throughout the world. Most Rotifera, numbering about 1,800 species, are free-living forms in both freshwater and salt water.

The Gastrotricha number about 1,800 species and live in salt water and freshwater. The Kinorhyncha, which includes about 100 known species, is a marine group. Members of the Nematomorpha, containing 250 to 300 species, are parasitic in arthropods (*e.g.*, insects, spiders, centipedes, and crabs) as juveniles and free-living in salt water or freshwater as adults. Priapulida, which numbers less than ten species, is a salt-water group.

NATURAL HISTORY

Reproduction and development. Most aschelminth species either have separate sexes or are hermaphrodites (*i.e.*, possessing functional reproductive organs of both sexes) derived from males or females.

The eggs are fertilized inside the female, but development usually takes place outside the female's body. Sperm are injected by male rotifers through the female body wall. In nematodes, sperm are injected through the female genital pore (vulva) with the assistance of the copulatory spicules (needlelike structures) of the male (see Figure 2). Copulation in kinorhynchs presumably involves setae, or bristles, on the male copulatory organ. Gastrotrichs are mainly hermaphroditic. Male nematomorphs deposit sperm near the female cloaca (*i.e.*, a cavity that functions as both a reproductive and an excretory duct); the sperm move through the cloaca before entering a storage sac called the seminal receptacle. In some rotifers and nematodes the eggs hatch into larvae within the female; the larvae then move to the outside.

The embryonic development of aschelminths, although peculiar in several ways, is basically determinate and spiral—that is, the early divisions of the fertilized egg produce cells called blastomeres that form definite organs or tissues in the adult, and certain parts (spindle axes) of the dividing egg are oriented in a definite way. It is therefore possible to identify each blastomere and to determine its final form in the adult.

During embryonic development a hollow ball of cells, a coeloblastula, is produced, whose central space, the blastocoel, forms the adult pseudocoel, in which lie the internal organs. Early in development, cell division ceases, except in the reproductive organs, so that the number of cells in the adult of a species is more or less constant. Even in adult rotifers and nematodes, whose tissues usually consist of masses of cytoplasm containing nuclei rather than individual cells with one nucleus each, the number of nuclei remains constant.

Larval development

Larval development involves periodic shedding (molting) of the cuticle in kinorhynchs, nematodes, nematomorphs, and priapulids. The external, or superficial, segments of newly hatched kinorhynch larvae are either lacking or indistinct; during molting, which occurs at least five times, the number of external segments increases, and associated modifications of the cuticle occur until the adult form is attained.

Molting in the larger nematodes is controlled by a mechanism involving an organic substance called a molting hormone and specialized cells in the nervous system called neurosecretory cells. Four molts occur between the four larval stages and the adult stage; in some parasitic nematodes one or two molts occur before the egg hatches. Larvae of most free-living nematodes differ from adults only in size and the absence of sex organs and sex characteristics. The larvae of many parasitic nematodes, on the other hand, differ markedly from the adults. These differences, apparently associated with the parasitic mode of life, usually involve only the head and the esophagus.

The nematomorph larva, free-living in water at the time of hatching, is small and divided into two regions—presoma and trunk. Within a short time it enters the body of the host arthropod—usually an insect but sometimes a centipede, a millipede, or a crab in the case of the marine nematomorph *Nectonema*. The larva grows in the body cavity of the host, gradually losing certain larval structures associated with feeding and derived from the presoma (proboscis and stylets); after several weeks or months, the larva leaves the host and molts to form a lethargic adult that is free-living in or near water. Only males in search of females actually swim.

Priapulids hatch as tiny larvae enclosed in plates derived from the cuticle. A series of circular spines at the front end is followed by a narrow neck; both regions can be withdrawn into the lorica, a thick, hard covering. The last circle of spines in larvae of *Halicryptus* serves as an outlet for gland cells resembling the adhesive tubes of gastrotrichs. Larvae of the genus *Priapulus* possess lateral tactile spines resembling lateral antennae found in rotifers and a terminal foot similar to that in rotifers. The internal structure of the larvae resembles that of the adults. Priapulid larvae apparently live and feed on the ocean bottom for about two years before shedding the lorica; they molt throughout their lives.

Behaviour and ecology. *Habitat.* Free-living aschelminths are aquatic animals; the aquatic medium, however, may occupy only a tiny crevice in soil or sand. Rotifers are abundant in freshwaters everywhere; a few live in brackish water or the sea. Gastrotrichs live primarily in freshwater, although there are also a few marine forms. Kinorhynchs are wholly marine; nematomorphs, except for the marine genus *Nectonema*, live only in freshwater. Nematodes occur in virtually every aquatic habitat, including vinegar; priapulids are marine.

Locomotion. Most aschelminths remain close to the substrate, defined here as the ground or the surface of an object with which an organism associates, when moving. Some rotifers have a terminal posterior foot that is elongated and stalklike; the foot serves to attach the animal permanently. In other rotifers the foot is shorter and is used either to grasp objects as the animal creeps or, in the swimming forms, as a rudder. Gastrotrichs glide across a substrate on cilia; kinorhynchs move by protruding the head, locking it in the substrate by the

scalids, or hooked spines, and pulling the rest of the body forward. Nematodes usually move in a way similar to that of snakes—*i.e.*, by waves of contraction that move down the body; they can also move or remain stationary by using an adhesive cement produced by the caudal glands. Priapulids bury themselves in the soft bottom mud of the sea at depths as great as 500 metres (over 1,600 feet).

Food. Aschelminths generally feed on small particles, such as bacteria, protozoans, and organic detritus; but many nematodes and rotifers and all priapulids are carnivorous and can kill and ingest larger animals. Digestion is believed to be extracellular (*i.e.*, occurring outside the cells) in the forepart of the gut, with absorption taking place in the hindpart; in some rotifers and nematodes, however, intracellular digestion may occur.

Adaptations to dry conditions. Many aschelminths live in temporary pools and have evolved eggs that are resistant to drying and other means of surviving dry conditions. Some rotifers can survive desiccation for three to four years then become active again when placed in water. Some bdelloids (rotifer order Bdelloidea) secrete a protective cyst. Others produce three types of eggs, of which a thick-shelled dormant type can survive desiccation and low temperatures for months; such eggs invariably produce females. The gastrotrich order Chaetonatoidea produces two types of eggs, one of which is dormant and resistant to desiccation. Nematode eggs are covered by a protein membrane that occasionally has a heavily sculptured appearance; in some cases it contains chitin, a tough carbohydrate substance that strengthens the case. The egg shells may have lids through which the young hatch or, particularly in some mermithids, so-called byssus threads by which the eggs become attached to plants. In many groups of nematodes, particularly parasitic forms, only one type of egg is normally produced; it is highly resistant to desiccation and to the action of chemicals.

Resistance of rotifers

Parasitic and free-living nematodes. Larvae of parasitic nematodes are either free-living or parasitic in a wide range of intermediate hosts: annelid worms, arthropods, mollusks, vertebrates of all types, and plants. Some species have parasitic larvae and free-living adults. Adults also occur in a wide range of hosts, which may be infected by eating nematode eggs or larvae; in some cases, the larvae penetrate the host's skin. In the medically important filariids the larvae enter the bloodstream of their vertebrate host, from which they may be sucked by blood-sucking insects. In the insect host further development takes place before the larvae are injected into another vertebrate host, in which they grow to adulthood. The free-living larvae of many plant-parasitic nematodes first attack the roots of the host plant; they remain attached as adults on the outer surface of the root or penetrate the plant tissues, through which they may migrate.

FORM AND FUNCTION

External structure. *General characteristics.* Separation of the head as a distinct body part is seldom apparent in aschelminths. The body is usually elongated; some rotifers, however, are nearly spherical. Spines, scales, and setae (bristles) of various kinds occur on the body surface. Gastrotrichs are almost invariably wormlike, with an expanded head region; the body bears cilia, particularly on the head lobe and along the ventral surface. The kinorhynch body, relatively constant in diameter, is superficially divided into segments and bears a series of regularly arranged spines.

Nematodes and nematomorphs are usually wormlike; some of the free-living marine nematodes may be conspicuously ringed and bear many long setae; a few have rows of ventral stiltlike bristles on which they move.

The warty body of the priapulids is divided into the presoma and a longer annulate (ringed) trunk irregularly covered by small spines and papillae (fingerlike projections). The anterior tip of the body is modified as a retractable, spined, barrel-shaped structure (proboscis).

The head and related structures. The anterior end of an aschelminth, although called a head, does not usually form a distinct one. In rotifers it is typically broad, with the mouth on a so-called apical field, which is centrally located, nonciliated, and surrounded by the corona (a ciliated region; see Figure 1). The apical field usually

The corona of rotifers

From L. Hyman, *The Invertebrates.* (1959); used by permission of McGraw-Hill Book Co.

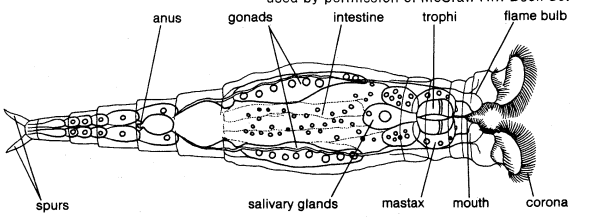


Figure 1: Body plan of rotifer (*Philodina roseola*).

bears a series of sensory hairs and the ducts of the retrocerebral organ, which is of uncertain function. The shape of the corona varies with its function. The cilia are frequently fused to form cirri (tentacle-like organs), membranelles (flat plates), and other complex structures; such structures, however, may be entirely lacking. In general the corona is an organ of locomotion; it is reduced in sessile (nonmoving) and creeping forms, but the former may use it to produce water currents for passing food into the mouth.

The anterior end of nematomorphs, called the calotte, is usually white and bordered posteriorly by a ring of dark pigment. The mouth is terminal or ventral within the calotte but may be totally lacking in the adults of some species.

Kinorhynchs are virtually the only aschelminths with a more or less definite head. It can be withdrawn into the anterior region by muscles attached to the outer surface of the esophagus or pharynx. The roughly spherical head includes the first of the 13 or 14 external segments and bears a central terminal mouth and five to seven circles of scalids (hooked spines) that may bear bristles. The mouth is on the tip of an extendable mouth cone armed with spines called oral styles; there may be other spines on the mouth cone.

Posterior to the head in some kinorhynchs is a region divided into 14 or 16 plates, or placids, which close over the anterior end when the head is retracted. Other kinorhynchs have eight or ten slight thickenings that can be withdrawn with the head.

The gastrotrich head is typically a swollen lobelike region with one or two pairs of red pigment spots; sometimes with one or two pairs of palps, or tentacles, on each lobe; and with long tufts of head cilia, particularly in chaetonotids, that probably act as sense organs. The chaetonid head also contains paired ciliated pits. The circular mouth opening is usually surrounded by bristles or hooks derived from the lining of the mouth cavity.

The nematode head typically bears a central, terminal mouth surrounded by a series of sense organs that are probably tactile (*i.e.*, pertaining to touch) and a pair of lateral sensory structures called amphids. The latter are characteristic features of nematodes. The mouth opening may be bounded either by a fringe derived from the cuticle supported by 12 rods (rugae), or by two, three, or six muscular lips; in some species, the mouth is rather large and permanently open. Basically the sense organs form three circles: an internal circle of six organs bordering the mouth opening, an intermediate circle of six organs, and an outer circle of four. The last two circles in some groups unite to form one circle of ten, in which the components of the original circles can be distinguished.

In other groups, the intermediate circle apparently is lacking, and only an outer circle of four remains. The sense organs of the head, other than the amphids, are often hairlike in free-living marine species and papillose, or fingerlike, in parasitic species and in those inhabiting the soil. The lateral amphids, of uncertain function, may be spiral-shaped grooves in the cuticle, pocket-like and elongated from front to back, or—in parasitic species—papillose.

A hypothetical primitive nematode head has been described as having the mouth bounded by six lips with two papillae—one related to the inner lips, the other to the intermediate lips—plus an outer circle of four additional papillae. From such a hypothetical arrangement other lip and sense organ arrangements could have been derived. Although such speculation is of some value, the marked similarity between the circular, rugae-bordered mouth openings of some members of the order Chromodorida and the mouth of many gastrotrichs suggests that the earliest nematodes did not possess lips.

The cheilostome and esophostome

The nematode mouth opens into a buccal cavity, or cheilostome, lined by cuticle similar to that covering the outer surface of the body. The anterior region of the esophagus may also form a chamber, the esophostome. Toothlike structures of various sizes and number may develop from the walls of the cheilostome and esophagus; spear-like structures in the esophagus are characteristic of certain nematodes parasitic in plants and insects.

The mouth in the Priapulida opens in the centre of the retractable proboscis and is surrounded by spines arranged in concentric five-membered groups.

The tail. Most aschelminths, except kinorhynchs, have a tail, frequently with structures that may aid in locomotion, may enable the animal to attach to a substrate, or—in males—may function during mating. In rotifers this tail is called a foot; it may be divided at the tip to form spurs (see Figure 1) or toes, at the ends of which pedal glands extrude an adhesive cement. Similarly, the tail of chaetonotoid gastrotrichs has tubes that produce an adhesive substance; in the gastrotrich order Macrodyasoidea, as many as 250 such tubes and associated adhesive glands occur over the entire body. In kinorhynchs, which have no true tail, a pair of similar adhesive tubes occurs in the ventral surface.

Many free-living nematodes, particularly marine forms, have a pore on the tail through which a cement substance produced by three internal caudal (tail) glands is extruded (see Figure 2); several rows of pores of uncertain function are found along the body of many, largely marine groups. In many species, particularly parasitic ones and those inhabiting the soil, the cuticle of the male tail is expanded into thin lateral sheets called caudal alae. Priapulids have no true tail, although one or two hollow caudal appendages of uncertain function occur in two species. The posterior end of the body also bears the anus and two urogenital (excretory and reproductive) pores.

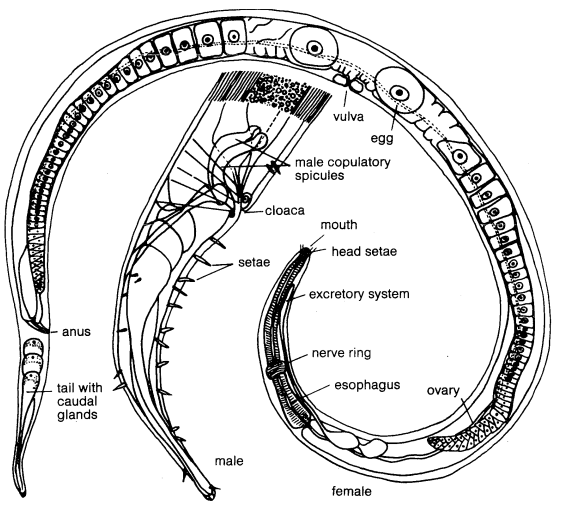


Figure 2: Body plan of nematode (*Axonolaimus paraspinosus*). General view of a female and the tail of a male.

Body structure. *The first tube: the cuticle.* The body of all aschelminths can be considered a series of tubes, one within the other. The outermost tube consists of the epidermis ("skin") and the cuticle. The cuticle may form rings; in rotifers, however, it forms thick plates, and in kinorhynchs it forms external segments. The gastrotrich cuticle is covered with spines, and that of various nematodes bears structures such as hooks, rings, and bristles.

Cuticle structures and modifications

The cuticle is derived from and attached to the underlying syncytial epidermis—*i.e.*, epidermis without cell walls. This epidermis, structurally simple in rotifers and gastrotrichs, bulges in nematodes into the pseudocoel to form three or four longitudinal ridges, or chords, which are dorsal, ventral, and lateral in position. In kinorhynchs the chords are dorsal and lateral; in nematomorphs they are dorsal and ventral.

The cuticle of large nematodes may have as many as eight layers. Some small species have several layers with complex systems of rings, rods, and articulatory ("hinging") processes. Generally, the cuticle appears to consist of three layers: an outer cortical layer, a middle plastic matrix layer, and an inner basal layer.

The surface of the nematode cuticle is usually marked by transverse grooves (striations) and may appear to be covered by dots and lines that indicate the presence of rings or rods within the cuticle. They do not open to or project from the surface. Four structural types of cuticle occur in the adult. In one type, hardened cortical and basal layers are separated by the usually plastic, sometimes fluid, matrix layer; through the middle layer pass rigid columns connecting the cortical with the basal layer. The second type of cuticle consists of transverse rings, or annules, of rigid material separated by narrow regions of more flexible material. The rigid annules have overlapping edges or a series of comblike processes that extend from one annule to another. The third type of cuticle has several distinct layers of crossing, spiraling fibres that form a flexible lattice, which usually consists of an annular cortical layer and a median, matrix layer that is soft and easily deformed. The fourth type of cuticle is very thin, with no distinguishing features other than striations and dots.

The cuticle of nematomorphs consists of an outer homogeneous layer and an inner layer that may have as many as 45 fibrous laminations. The most conspicuous feature of the cuticle in many species is the irregular pattern of wartlike processes, or areoles, projecting from the outer surface. In priapulids the cuticle has two layers: an outer homogeneous layer and a thicker, laminated inner layer.

The cuticle in rotifers, gastrotrichs, and kinorhynchs is thin; little else is known of its structure except that it contains no chitin. In kinorhynchs the cuticle forms overlapping annular segments. In rotifers it is yellowish in colour and often ringed and, in some species, forms the lorica, consisting of one or more plates; this structure may often be conspicuously marked with grooves, striations, tubercles (knoblike projections), or spines. The gastrotrich cuticle is typically covered by scales, which, at the posterior end of the body, frequently bear curved, sometimes multipronged spines.

The second tube: the musculature. The muscles form the next tube of the body. In rotifers, gastrotrichs, and kinorhynchs, they consist of a number of enucleate—*i.e.*, cells without nuclei—circular muscles running transversely around the body. Typically, another system of nucleate muscles runs longitudinally. This longitudinal system, least modified in gastrotrichs, is most extensive along the ventral surface. In kinorhynchs, the longitudinal muscles correspond to the external segmentation. Rotifers have a large number of single muscles, but they are not arranged to form a system. In nematodes and nematomorphs, only longitudinal muscles occur. Two muscle layers occur in priapulids; the longitudinal muscles apparently control the proboscis, but there are also circular muscles in that region.

The third tube: the digestive system. The third tube, the pseudocoelom, is poorly developed in rotifers, gastrotrichs, and nematomorphs but is extensive and fluid-filled in nematodes and kinorhynchs. The large body cavity of priapulids probably does not constitute a pseudocoelom.

The innermost tube of aschelminths, the digestive tract, extends the length of the body and, toward the anterior end, is always swollen and muscular. In gastrotrichs,

Adaptations of the digestive system

kinorhynchs, and nematodes, the mouth leads into a short, often elaborate buccal (mouth) cavity, followed by the swollen muscular region variously called an esophagus, pharynx, or mastax. This region consists of a mass of radial muscles, through the middle of which runs a triradiate (Y-shaped), cuticle-lined space, or lumen. In some kinorhynchs, however, the lumen may be round or flattened. In nematodes with a triradiate lumen, one arm is always midventral; it is middorsal in kinorhynchs and some gastrotrichs (macrodyasoids) but ventral in others (chaetonotoids). In certain nematodes and gastrotrichs, the esophagus forms muscular bulbs; in some nematodes the bulbs have valves. Gland cells, or salivary cells, occur on the outer surface of the esophagus in kinorhynchs and gastrotrichs. Three similar groups of cells are buried within the nematode esophagus and empty by ducts into the lumen. In kinorhynchs the esophagus has an epithelial cell layer below the lining; the latter is derived from the cuticle.

In rotifers, the mouth is often on the ventral side. It may lead either directly to the esophagus, or mastax, or indirectly through a ciliated buccal tube. The mastax is usually a muscular organ, and the arrangement of the muscles may be complex, with an internal chewing apparatus made up of several cuticle-derived structures—trophi—that are characteristic of the different groups of rotifers. Usually, a pair of salivary glands empties anterior to the trophi. The esophagus is swollen anteriorly into a stomach. The intestinal wall, which is not lined with cuticle, consists of one layer of granular cells, which may bear cilia. The posterior end may terminate in a rectum.

The esophagus of priapulids differs from that of other aschelminths in having both circular and longitudinal muscles covered externally by a thin layer of longitudinal muscles and internally by an epidermis and a spiny cuticle. The esophagus leads into a nonmuscular intestine.

The nematomorph digestive tract is simple; both sexes have a cloaca, and there is frequently no mouth. The anterior part of the gut usually is a solid rod of cells, presumably representing the esophagus; the intestine, a small, simple tube, may be an excretory organ.

Functional features. *Water balance and excretion.* Aschelminths have no special respiratory or circulatory systems. Water balance and excretion are carried out by tubes closed at one end; at the other end they empty into the posterior end of the gut. This arrangement prevails in most rotifers, some gastrotrichs (order Chaetonotoidea), and all kinorhynchs. In priapulids the tubes are connected to the ducts of the gonads, forming excretory organs called protonephridia; these do not occur in nematodes and nematomorphs. The protonephridia in rotifers and gastrotrichs have flame bulbs, cup-shaped organs into the lumen of which project cilia. In kinorhynchs and priapulids the flame bulbs are replaced by solenocytes, long tubular cells with one whiplike hair, or flagellum. In priapulids, several large clusters of solenocytes close to the gonads are supported by mesenteries, or sheets of cells, within the body cavity.

In most marine nematodes water balance and some excretion are performed by one large ventral gland cell, the renette (see Figure 2). In most soil, freshwater, and parasitic species, however, these functions are performed by lateral tubes buried in the lateral cords of the epidermis. Both systems empty to the exterior through an anterior, ventral pore. A definite excretory system is lacking in all nematomorphs, some rotifers, and some nematodes.

Reproductive system. The reproductive system is relatively simple. The ovaries and testes are often double in both sexes, but most rotifers and some nematodes have one gonad. The male and female sex cells in rotifers and most gastrotrichs empty through the cloaca; kinorhynchs have terminal lateral gonopores, or reproductive openings, in each sex. In nematodes the sperm leave by the subterminal cloaca in most males, and in most females the eggs leave via an opening at about the middle of the body (Figure 2). The tail of male nematodes usually has files of sensory papillae, setae, or rays and a pair of copulatory, rodlike spicules (Figure 2). In some groups, caudal alae, which assist in copulation, may develop around the cloacal region. A pair of small porelike sense organs (phasmids) occur on the tail of some nematodes.

Components of the nervous system

Nervous system. The aschelminth nervous system is relatively simple. It consists of an anterior nerve mass or nerve ring and a series of major longitudinal and numerous smaller nerves. There are two major ventral nerves in rotifers; a pair of similar lateral nerves in gastrotrichs; a single indistinct ventral nerve in kinorhynchs; a midventral nerve cord in nematomorphs; and one ventral nerve in priapulids. The remainder of the priapulid nervous system is closely associated with the epidermis. In nematodes nerves extend anteriorly from a nerve ring that passes around the esophagus (see Figure 2); posteriorly are a middorsal nerve, a midventral nerve, and from one to three pairs of lateral nerves.

EVOLUTION

On largely structural evidence—since there is no fossil record—the Rotifera appear to be rather primitive animals that evolved more or less directly from the Turbellaria, a class of flatworms (phylum Platyhelminthes). The strongest evidence is provided by developmental similarities of the two groups and by the presence of virtually identical protonephridia in rotifers and in certain turbellaria.

The Gastrotricha are considered by some authorities to be related to rotifers because of the form of the body musculature and, in some species, the presence of protonephridia with flame cells. These features are common to most aschelminth groups, however; gastrotrichs resemble nematodes—to which they are probably most closely related—in the structure of esophagus and the digestive tract and in the presence of cuticular spines and bristles; in addition, the gastrotrichs have an organ similar to the adhesive tube found in some marine nematodes. Many of these resemblances, however, may be only structural similarities and thus may have no evolutionary significance. It is doubtful that gastrotrichs are a natural group since the location of the esophageal lumen of members of the order Macrodyasoidea differs from that of other gastrotrichs and nematodes (see above *Tubular structure*). In addition, the Macrodyasoidea have pores that lead from the lumen of the esophagus to the exterior, and they lack protonephridia.

Both Kinorhyncha and Rotifera have been considered as relatives of arthropods or as links between arthropods and annelids, largely because of their segmentation and the addition, during development, of segments anterior to the terminal segment of the body. The segmentation is only superficially similar, however, since in kinorhynchs it develops from the cuticle inward, and in annelids and arthropods it starts in the mesoderm and develops outward.

In contrast, the similarities between kinorhynchs and other aschelminths are extensive: no true coelom, protonephridia with flame cells, a close association between the nervous system and epidermis, presence of adhesive tubes and copulatory spicules, and similar form of the esophagus. On the basis of these similarities, kinorhynchs are generally considered relatives of rotifers, gastrotrichs, and nematodes.

Nematodes are generally regarded as pseudocoelomate aschelminths. The only strong evidence against such a relationship is the absence of protonephridia. Nematodes, elongated cylinder-shaped animals, have longitudinal muscles that move the body in a snakelike fashion. The "skeleton" of a nematode is the pseudocoelomic fluid, which is under pressure. The esophagus, an efficient anterior pump, sucks food into the gut against the pressure of the fluid in the pseudocoelom. The absence in nematodes of flame cells and solenocytes may be attributable to the fact that cilia and flagella probably cannot function in the presence of such internal pressure. Because mechanical problems are imposed on nematodes by the structural nature of the body, most of the features characteristic of nematodes are functionally interrelated. Any animal that evolved such a unique body would nec-

essarily have such complex morphological features. It is therefore difficult to determine phylogenetic relationships on the basis of structural features. Even the triradiate esophageal lumen common to most aschelminths is not necessarily indicative of a relationship among the various groups; it happens to be the most simple and efficient shape for a lumen.

Although the taxonomic position of the nematomorphs is somewhat complicated by the unique nature of their digestive system, the free-living larvae have certain structural similarities to the kinorhynchs, and the nematomorphs are more easily accommodated within the aschelminths than elsewhere. The relationship of the Priapulida to other groups is even more uncertain, since it has not yet been definitely established that it is a pseudocoelomate group. Although treated here as aschelminths for convenience, the priapulids are probably more appropriately treated as a distinct phylum.

Aschelminthes thus can be considered a grouping of convenience. None of its subgroups is clearly derived from other members, and all are at about the same organizational level. Aschelminthes are important, nevertheless, in demonstrating the morphological problems involved in acquiring a body cavity and an anus; they also reveal the limitations that solutions to such problems can impose on the design of the invertebrate body.

CLASSIFICATION

Distinguishing taxonomic features. The classification used here is based on *The Invertebrates*, by L.H. Hyman, still an excellent survey of the Aschelminthes, although published in 1951. According to this system, the Aschelminthes are considered a phylum and each of the six groups a class, although Priapulida could justifiably be considered an independent phylum. Structural similarities allow the inclusion of Rotifera, Gastrotricha, Kinorhyncha, and Nematoda within the phylum, although some authorities are hesitant about such a grouping. The taxonomy of Aschelminthes presents difficulties that have not been fully resolved.

Annotated classification. The classification of Rotifera, Gastrotricha, and Kinorhyncha used here is basically unchanged from that of Hyman; that of the Nematoda has been rearranged.

PHYLUM ASCHELMINTHES

Many-celled, bilaterally symmetrical, unsegmented, invertebrate animals with a pseudocoel between the gut and the body wall; about 17,000 known species; many forms aquatic; often parasitic; mostly wormlike, generally microscopic or very small; no circulatory or respiratory organs; nervous system usually consists of anterior ganglia and longitudinal nerve cords; reproductive system simple.

Class Rotifera

Microscopic and aquatic; anterior corona (area of cilia surrounding anterior end); internal jaws or trophi (a set of small, hardened jaws); about 1,800 species.

Order Seisonacea or Seisonidea. Marine, epizoic (*i.e.*, living on or in another animal but not parasitic); elongated forms; corona poorly developed; sexes similar in appearance; gonads paired.

Order Bdelloidea. Swimming or creeping; retractable anterior end; males unknown; 2 germovitellaria (organs that produce ova and yolk cells separately); mastax (esophagus) ramate (stout with large platelike unci, or toothlike structures).

Order Monogononta. Swimming or nonmoving (sessile); 1 germovitellarium; males small with 1 testis; mastax not ramate.

Class Gastrotricha

Microscopic, aquatic; without corona; cilia on body, particularly on ventral surface; adhesive tubes present; no trophi or mastax; about 1,800 known species.

Order Macrodyasoidea. Elongated marine forms; adhesive tubes full length of body; no protonephridia (excretory organs).

Order Chaetonotoidea. Mostly fusiform (spindle-shaped) freshwater forms; 2 to 4 pairs of adhesive tubes posteriorly; protonephridia present.

Class Kinorhyncha (Echinodera)

Microscopic, marine forms; 13 or 14 external segments (zonites); head retractable; protonephridia present; about 100 known species; orders used below correspond to Hyman's informal groups.

Order Heterorhages. Closing apparatus of 3rd external segment without a definite dorsal plate; relatively spiny.

Order Homalorhages. Closing apparatus of 3rd external segment with a definite dorsal and 3 ventral plates; relatively less spiny.

Class Priapulida

Small to medium-sized (to about 8 cm, or 3 in.), warty marine forms with superficial rings and a retractile proboscis (structure at anterior end); 2 genera; *Priapulus* and *Halicryptus;* few species.

Class Nematomorpha (Gordiacea) (hairworms)

Very long and thin, no rings on cuticle ("skin"); juveniles parasitic in arthropods; adults free-living in seawater and freshwater; about 250 to 300 species.

Order Gordioidea. Freshwater forms, without body bristles; larvae parasitic in terrestrial or aquatic arthropods; pseudocoel reduced.

Order Nectonematoidea. Open-sea dwelling; double row of bristles on body; larvae parasitic in Crustacea; pseudocoel not reduced.

Class Nematoda

Largest and most diverse aschelminth class; about 13,000 known species; classification below recognizes 18 orders in 3 groups that should be treated as subclasses in a formal classification; the first 2 groups comprise what are often called the Adenophorea (or Aphasmidia); the 3rd group is often called Secernentea (or Phasmidia).

Group 1 (Adenophorea [Aphasmidia])

Amphids (lateral sense organs at head region) generally obvious, pocket-like, below the cuticle surface, typically opening through a horizontal slit; excretory system a ventral cell; cuticle with cross spiral fibres, rarely with strong annulations (rings); sense organs surrounding mouth, typically setae (bristles); body setae and caudal glands common; no phasmids (porelike sense organs); some species parasitic.

Order Enoplida. Marine, free-living; buccal (mouth) cavity (esophostome) surrounded by esophageal tissue; caudal glands usually present.

Order Oncholaimida. Marine, free-living; large cuplike buccal cavity not surrounded by tissue, usually contains large teeth; caudal glands usually present.

Order Dorylaimida. In freshwater or soil; rarely marine or plant parasites; papillae (fingerlike projections) at head region; buccal cavity large and toothed or with spear-shaped structure in esophagus; caudal glands usually absent.

Order Mermithida. Smooth, filiform (threadlike); either larvae parasitic in insects and adults free-living in soil or water or adults parasitic in insects and larvae free-living; intestine modified for food storage; no anus.

Order Trichurida. Parasitic in vertebrates; body usually thin anteriorly; mouth minute; esophagus reduced; stichosomes (internal glandular columns) present; 1 male spicule (needlelike structure used in copulation); larval buccal cavity with spear-shaped structure.

Order Dioctophymatida. Parasitic in vertebrates; larval buccal cavity with spear-shaped structure; 1 long spicule and a terminal sucker-like organ in male.

Group 2 (Adenophorea [Aphasmidia])

Free-living, typically small marine forms, occasionally found in soil and freshwater; amphids open grooves in cuticle, forming spirals, circles; excretory system a ventral cell; cuticle frequently ringed; body setae common; caudal glands present; no phasmids.

Order Araeolaimida. Amphids spiral or looplike; sense organs of head region usually in 3 circles; cuticle simple, sometimes ringed.

Order Monhysterida. Amphids typically circular; sense organs of head region generally in 2 circles; cuticle sometimes ringed.

Order Chromadorida. Sense organs of head region in 2 or 3 circles; amphids simple or multispiral, sometimes reniform (bean-shaped); cuticle always heavily ornamented, complex in some genera.

Order Desmodorida. Wholly marine; amphids spiral or crook-shaped; cuticle strongly annulate; locomotory setae in 1 subgroup (Draconematina).

Group 3 (Secernentea [Phasmidia])

Typically in freshwater and soil or parasitic in wide range of vertebrates, invertebrates, and plants; amphids rarely obvious.

often papillate; excretory system consists of lateral canals; cuticle superficially structureless; no body setae; phasmids present; no caudal glands.

Order Rhabditida. Most free-living, some parasitic; buccal cavity well-developed; posterior, valved esophageal bulb present; 2 spicules in male.

Order Rhabdiasida. No definite esophageal bulb; complex life history, with reproduction occurring in parasitic and free-living stages; parasitic stages hermaphroditic (both sex organs on 1 animal) or parthenogenetic (reproduction without fertilization); free-living stage may be bisexual.

Order Oxyurida. Typically small; valvulated posterior bulb, at least in larvae; males with one or no spicules; larvae like miniature adults; species are parasites mainly of terrestrial hosts, invertebrates and vertebrates, including man.

Order Strongylida. Mouth large without lips but frequently with cuticular leafs (corona); males with broad caudal alae (sheets of cuticle) supported by rays; life histories variable, usually direct (no intermediate host); parasites of mammals, rarely of birds or reptiles.

Order Ascaridida. Three prominent lips; buccal cavity not obvious; no posterior esophageal bulb in adults; life history usually indirect (*i.e.*, with an intermediate host or hosts); parasites of vertebrates.

Order Spirurida. Typically 2 lateral lips; esophagus muscular anteriorly, glandular posteriorly, without bulb; life history indirect; parasites of vertebrates.

Order Dracunculida. No lips or buccal cavity; esophagus as in Spirurida; vulva (female genital pore) usually nonfunctional; male spicules filiform; life history indirect; parasites of vertebrates.

Order Filariida. Long and thin, without lips; buccal cavity minute or lacking; esophagus muscular anteriorly, glandular posteriorly, without bulb; vulva anterior; young hatch from egg outside or inside the female; life history frequently involves removal of young organisms (microfilaria) from host's blood by blood-sucking insect; parasites of vertebrates except fish.

BIBLIOGRAPHY. B.G. and M.B. CHITWOOD, *An Introduction to Nematology*, rev. ed. (1950), an out-of-print classic of great importance; E.C. DOUGHERTY (ed.), *The Lower Metazoa: Comparative Biology and Phylogeny* (1963), a technical, thorough, authoritative work; L.H. HYMAN, *The Invertebrates*, vol. 3, *Acanthocephala, Aschelminthes, and Entoprocta* (1951), a useful general work; P.P. GRASSE (ed.), *Traité de zoologie*, vol. 4, *Némathelminthes ou Aschelminthes* (1965), a monumental treatise. An immense body of literature is available on Aschelminthes and on nematodes in particular. Many basic works, however, are not available in English. Of special interest are the sections (in German) by A. REMANE in H.G. BRONN (ed.), *Klassen und Ordnungen des Tier-reichs*, vol. 4 (1929–33); and in G. GRIMPE and E. WAGLER (eds.), *Die Tierwelt der Nord- und Ostsee* (1929).

(W.G.I.)

Ashʿarī, al-

Abū al-Ḥasan al-Ashʿarī was a famous Muslim theologian, Arabic in language and culture, who founded a theological school that later claimed as members such celebrated authors as al-Ghazālī (died AH 505 [AD 1111] —AH denotes dates according to the Muslim calendar) and Ibn Khaldūn (died AH 808 [1406]).

He was born about AH 260 (AD 873/874) in the city of Basra, at that time one of the centres of intellectual ferment in Iraq, which, in turn, was the centre of the Muslim world and the seat of a world civilization. It is generally agreed that he belonged to the family of the celebrated Companion of the Prophet Abū Mūsā al-Ashʿarī (died AH 42 [AD 662/663]), though some theologians opposed to his ideas contest the claim. Since this would have made him by birth a member of the Arab-Muslim aristocracy of the period, he must have received a careful education. A contemporary recorded that the wealth of al-Ashʿarī's family permitted him to devote himself entirely to research and study.

Early work

His works, especially the first part of *Maqālāt al-Islāmīyīn* ("Theological Opinions of the Muslims"), and the accounts of later historians record that al-Ashʿarī very early joined the school of the great theologians of that time, the Muʿtazilites. He became the favourite disciple of Abū ʿAlī al-Jubbāʾī (died AH 303 [AD 915/916]), head of the Muʿtazilites of Basra in the final decades of the 3rd century AH (late 9th and early 10th centuries AD).

Despite the lacunae in the documentation concerning these theologians, certain characteristic traits of their culture and social position can be singled out. Anxious to mark the originality of Islām in contrast to all dualist doctrines and in contrast to Christianity and Judaism as interpreted by the Qurʾān, the Muʿtazilites concentrated their efforts on underlining the absolute transcendence of the one God. To accomplish this they drew principally from their own Arab-Muslim tradition and remained relatively impervious to the foreign cultures (especially Greek and Iranian) that invaded Baghdad from the beginning of the 3rd century AH. But the Arab-Muslim culture was not popular: during the period of al-Ashʿarī's studies, the rupture between the intellectual elite of Muʿtazilite theologians and the common people was all but complete.

A disciple of al-Ashʿarī describes that period of his master's life in this way: "Al-Ashʿarī was a disciple of Jubbāʾī. He faithfully went to hear him and take lessons from him, never leaving him for all of 40 years. In the sessions devoted to controversy, he showed his gift for argumentation and was bold in confronting his adversary; but he was not gifted for writing. When he took up a pen, at times he never finished, and at times what he wrote was not satisfactory."

That testimony, at least in its negative aspect, should be corrected by what Ibn ʿAsākir reports: "His works are very well organized, the expressions and the developments are very exact." In any case, it was during that period of his life that al-Ashʿarī, the brilliant and faithful disciple of the Muʿtazilites, undertook the composition of a work in which he gathered the opinions of the diverse schools on the principal points of Muslim theology. That work, the first volume of the current edition of the *Maqālāt*, is valuable for what it records of Muʿtazilite doctrines. It remains one of the most important sources for retracing the history of the beginnings of Muslim theology.

At the same period al-Ashʿarī composed *Risālah ilā ahl ath-thaghr* ("Treatise for the Men of the Frontier") for the Muslims of Bāb al-Abwāb (Derbent, between the Caucasus and the Caspian). The occasion was the renewed interest of the central administration in the security of the northern frontier of the empire.

Later, at the age of 40 (*c.* AH 300 [AD 912/913]), when he had become a specialist in theology and was well known for his oral controversies and his written works, al-Ashʿarī quit his master al-Jubbāʾī and abandoned Muʿtazilite doctrine. This conversion was spectacular. It made al-Ashʿarī the focus of attention of Basra for some time and merited a certain number of accounts, which, though impossible to confirm in their details, are certainly correct in their general lines. What happened was apparently this: following a crisis of conscience, accompanied, perhaps, by dreams, al-Ashʿarī saw clearly the limits of Muʿtazilism and was led back to a closer attachment to the sources of Muslim faith, the Qurʾān (Islāmic scripture) and the *Sunnah* (tradition). Reading texts of his master, al-Jubbāʾī, it is possible to ascertain the defects that may have struck al-Ashʿarī. In those texts al-Jubbāʾī seems to have no particular audience in mind, nor does he try to convince; he only demonstrates. It appears that, for him, the reality of God as well as that of man has been so sterilized and desiccated that it has become little more than matter for rational manipulation.

Later ideas and works

Al-Ashʿarī, more conscious than all others of these limits and of the premature desiccation of Muʿtazilite theology, did not hesitate to proclaim his new faith publicly. Certain reports even speak of him dramatically declaring his new stand in the middle of the Friday prayer in the Cathedral Mosque of Basra. From that day, the former Muʿtazilite started combatting his colleagues of yesterday. He even attacked his old master, refuting his arguments in speech and writing. It was then, perhaps, that he took up again his first work, the *Maqālāt*, to add to the objective exposition rectifications more conformable to his new beliefs. In this same period, he composed the

work that marks clearly his break with the Mu'tazilite school: the *Kitāb al-Luma'* ("The Luminous Book").

It was not until his former master Abū 'Alī al-Jubbā'ī died at Basra in AH 303 (AD 915) that al-Ash'arī decided to make Baghdad his centre. Arriving in the capital, he soon became aware of the importance assumed by a group of faithful of the *Sunnah*, the disciples of Ibn Ḥanbal. Their leader, al-Barbahārī, was a dynamic person with a touch of the demagogue. Al-Ash'arī visited him to explain his doctrinal views, insisting on the fact that his previous theological formation enabled him to attack the Mu'tazilites with their own weapons thus making his treatises models of apologetic in support of the truth. The response was disappointing: there was no encouragement, no approval, no acceptance into the ranks of the Ḥanbalites. Al-Barbahārī curtly replied that he had no interest in any of this; he had only one master—Ibn Ḥanbal.

It must have been after that interview that al-Ash'arī composed, or perhaps put the last touches to, one of his most famous treatises, the *Ibānah 'an uṣūl ad-diyānah*, ("Statement on the Principles of the Religion"), which contains some passages venerating the memory of Ibn Ḥanbal.

In the years that followed, al-Ash'arī, now installed in Baghdad, began to group around himself his first disciples. Focussing his theological reflection on certain positions of the mystic al-Muḥāsibī (died AH 243 [AD 857]) and of two theologians, Ibn Kullāb (died *c.* AH 250 [AD 864/865]) and Qalanisī (died *c.* AH 300 [AD 912/13]), al-Ash'arī laid the bases for a new school of theology distinct from both the Mu'tazilites and the Ḥanbalites. His three best known disciples were al-Bāhilī, aṣ-Ṣu'lūkī, and Ibn Mujāhid, all of whom transmitted the doctrines of their master to what later became the flourishing school of Khorāsān.

Al-Ash'arī died sometime around AH 320 (AD 932/933) but most probably in AH 324 (AD 935/936). He was buried at the southwest of the city in a place called the Wharf of the Water Jars. A mausoleum erected over his tomb was later destroyed by fanatic Ḥanbalites.

The opposition that was first aroused by al-Ash'arī himself through his spectacular conversion continued to assail his disciples, but through constant dialogue with their opponents they slowly disentangled the main lines of doctrine that became the stamp of the Ash'arite school.

BIBLIOGRAPHY. M. ALLARD, *Le Problème des attributs divins dans la doctrine d'al-As'ari et de ses premiers grands disciples* (1966), is a complete study of the life and works of al-Ash'arī, including an exposition of his conception of God. An indispensable work for those who wish to read al-Ash'arī is R.J. MCCARTHY (ed.), *The Theology of al-Ash'ari* (1953), containing the texts of two of his theological treatises along with their English translation. Three appendixes give in translation the writing of Arab authors on the life and work of al-Ash'arī, and a fourth gives the main elements of his credo. W.M. WATT, *Islamic Philosophy and Theology* (1962), serves to situate al-Ash'arī in the chronological development of Muslim thought. A.J. ARBERRY, "Al-Ash'arī's Tract on Faith," *BSOAS*, 19:160–162 (1957), is a learned note showing how a short treatise of al-Ash'arī was handed down among the traditionists.

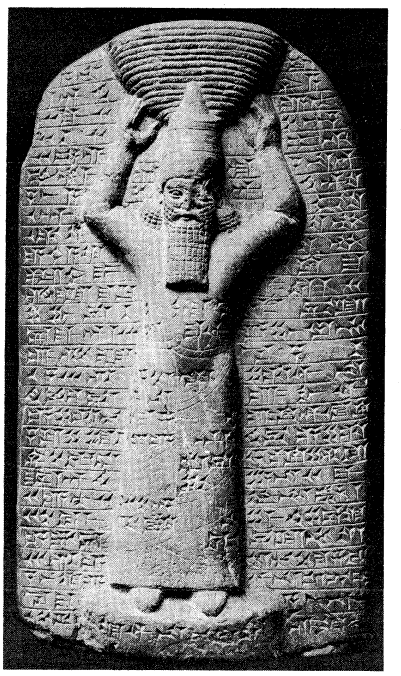
(M.A.Al.)

Ashurbanipal

Ashurbanipal was the last of the great kings of Assyria and collector of a great library at Nineveh. He reigned from 668 to 627 BC.

The life of this vigorous ruler of an empire ranging initially from the Persian Gulf to Cilicia, Syria, and Egypt can be largely reconstructed from his autobiographical annals and royal correspondence. His father, Esarhaddon, appointed him crown prince of Assyria in May 672 BC with the intention of averting dynastic struggles. Shamash-shum-ukin, a son of equal status by another wife, was appointed crown prince of Babylonia. Probably due to the influential queen mother Naqi'a-Zakutu, Ashurbanipal was given responsibility earlier.

Ashurbanipal was involved in administration and versed in the problems of controlling the northern hill tribes.



Ashurbanipal carrying a basket in the rebuilding of the temple, stone bas-relief from the Esagila, Babylon, 650 BC. In the British Museum.

By courtesy of the trustees of the British Museum

His tutors were Nabu-shar-usur, a general, and Nabu-ahi-eriba, who interested him in history and literature. Like few Mesopotamian kings before him, he mastered all scribal and priestly knowledge and was able to read Sumerian and obscure Akkadian scripts and languages. His athletic powers were shown in hunting, archery, and horsemanship. Though there is little evidence of his experience on the actual battlefield, there is no reason to doubt Ashurbanipal's claim that his father favoured him for his bravery and intelligence.

He soon shouldered heavy responsibilities, having to command the court and nobles. No governor or prefect was appointed without consulting him, and he had authority over many state building projects. His reports to his father showed such qualities of statesmanship that he was left in charge of all affairs while his father was en route to Egypt. When Esarhaddon died at Harran in December 669 BC, Ashurbanipal transferred full power to himself without incident. The queen mother exacted an oath of allegiance from both family and courtiers.

Ashurbanipal's reign

Ashurbanipal's first concern was to quell a revolution in Egypt, where Taharqa (Tarku; biblical Tirhaka), an Egyptian king, had invaded the Nile Delta and won support. Swift Assyrian military action enforced his withdrawal, and Ashurbanipal appointed local princes supported by Assyrian garrisons. Some of the princes intrigued with Taharqa, and the Assyrians deported them to Nineveh. Keeping to his plan to have native administrators, Ashurbanipal chose Necho I as supreme ruler of the delta and made a treaty with him. Further pressure from Taharqa's successor Tanutamon (Tandamane) led to another Assyrian intervention in 664–663, when the Assyrians seized control of Memphis and sacked Thebes. When Necho died in 663, Ashurbanipal held to his policy and accepted the succession of another local ruler, Psamtik (Psammetichus I); he was rewarded by a peace that enabled him to campaign elsewhere. In 654 BC the Assyrian garrisons were expelled from Egypt, but trade continued so that this loss resulted in little weakening of his position.

He next turned to the Phoenician city of Tyre, which had supported both Egyptian and Lydian bids for independence. A successful siege of Tyre led to the resubmission of the rulers of Syria and Cilicia and to a request for Assyrian help from Gyges of Lydia against Cimme-

rian intruders. Because Lydian mercenaries had assisted Egypt, this help was refused. A swift display of military might against the Mannaeans and an alliance with Madyes, the Scythian chief, repulsed Cimmerian advances and left Ashurbanipal free to attend to Babylonia, his southern neighbour.

Ashurbanipal had confirmed his half-brother Shamash-shum-ukin as local ruler of Babylonia, but with restricted powers. Assyrian garrisons and officials there continued to report to the Assyrian king, and he continued to appoint governors both in the Sea-lands (Persian Gulf) and Ur. Babylonians petitioned him direct and received land grants. For 16 years, relations with his brother were peaceful. When Tept-Humban, a usurper in Elam, entered Assyrian territory and was killed, the Assyrian action was primarily in support of the Elamite princes Humbanigash and Tammaritu, who were given specific regions in Elam with no attempt at direct Assyrian rule. Ashurbanipal's actions probably aimed also to assist his own brother, whom he still trusted. Ashurbanipal received a deputation of Babylonians about this time, and he punished the Gambulu tribe for complicity in the Elamite affair.

Shamash-shum-ukin's long stay in Babylon had imbued him with the traditional local spirit of nationalism and resistance. He may have interpreted his brother's policy of appeasement as weakness and as an opportunity for him to increase his own status. In any event, he contrived a coalition with other outlying peoples of the Assyrian Empire—Phoenicia, Judah, Elam, Egypt, Lydia, and the Arab and Chaldean tribesmen; and had these groups risen simultaneously Assyria would have fallen. When Ashurbanipal discovered the plots, he appealed directly to the Babylonians and perhaps tested their loyalty by imposing a special tax; only upon their refusal did he take military action. He seemed to move in ways that avoided direct danger to his brother, and he worked more through siege warfare than through direct action; the Babylonian Chronicle records that for three years "the war went on and there were perpetual clashes." Elam, suffering from internal dissension, was unable to help the rebels; and gradually, through starvation, the Arabs who had retreated into Babylon deserted as the famine became intense. Shamash-shum-ukin committed suicide in his burning palace in 648 BC. Ashurbanipal's own feelings toward the city are shown by his work of restoration and by his appointment of a Chaldean noble, Kandalanu, as his viceroy there.

Ashurbanipal had to take further action to quell the rebellion. Raiding the Arab tribes, he defeated the Nabataean Uate and his allies and isolated the Qadar tribe. The struggle with Elam was harder; war there dragged on until 639 BC, when the Assyrians sacked Susa. That year Ashurbanipal celebrated his triumph; he had "the whole world" under his sway, and four captive kings drew his chariot in the procession.

The military action required to maintain order must not overshadow Ashurbanipal's ability as an administrator. The empire prospered economically despite the threatened closure of the northern and eastern trade routes due to Lydian and Median expansions. Unfortunately, the sources are too scanty to follow his reign after 631 BC. Ashurbanipal's death is nowhere recorded, but it seems that he followed his father's precedent in bringing his sons Ashur-etel-ilani and Sin-shar-ishkun into coregency, each with a separately defined authority. It is no indictment of his rule that his empire fell within two decades after his death; this was due to external pressures rather than to internal strife.

Personality and significance

Ashurbanipal was a person of religious zeal. He rebuilt or adorned most of the major shrines of Assyria and Babylonia, paying particular attention to the "House of Succession" and the Ishtar Temple at Nineveh. Many of his actions were guided by the omen reports, in which he took a personal and informed interest. He celebrated the New Year Festival, and one of his reliefs, showing him dining in a garden with his queen Ashur-sharrat, may illustrate this event. His younger brothers were priests in Haran and Ashur.

Ashurbanipal's outstanding contribution resulted from his academic interests. He assembled in Nineveh the first systematically collected and catalogued library in the ancient Near East (of which 20,720 Assyrian tablets and fragments are now in the British Museum). At royal command, scribes searched out and collected or copied texts of every genre from temple libraries. These were added to the basic collection of tablets culled from Ashur, Calah, and Nineveh itself. The major group includes omen texts based on observations of events; on the behaviour and features of men, animals, and plants; and on the motions of the sun, moon, planets, and stars. Lexicographical texts list in dictionary form Sumerian, Akkadian, and other words, all essential to the scribal educational system. He collected many incantations, prayers, rituals, fables, proverbs, and other "canonical" and "extracanonical" texts. The traditional Mesopotamian epics such as the stories of Creation, Gilgamesh, Irra, Etana, and Anzu have survived mainly due to their preservation in his library. The presence of handbooks, scientific texts, and some folk tales (*The Poor Man of Nippur* was a precursor of one of the *Thousand and One Nights* tales of Baghdad) show that this library, of which only a fraction of the clay tablets has survived, was more than a mere reference library geared to the needs of diviners and others responsible for the King's spiritual security; it covered the whole range of Ashurbanipal's personal literary interests, and many works bear the royal mark of ownership in their colophons.

The King was patron of the arts; he adorned his new and restored palaces at Nineveh with sculptures depicting the main historical and ceremonial events of his long reign. The style shows a remarkable development over that of his predecessors, and many bas-reliefs have an epic quality unparalleled in the ancient world, which may well be due to the influence of this active and vigorous personality.

BIBLIOGRAPHY. M. STRECK, *Assurbanipal und die letzten assyrischen Könige bis zum Untergange Ninivehs* (1916), a reliable study and discussion of the historical, religious, and epistolary evidence for the king and his family; T. BAUER, *Das Inschriftenwerk Assurbanipals* (1933), further discussion with additional texts; A.C. PIEPKORN (ed.), *Historical Prism Inscriptions of Ashurbanipal* (1933), a literary analysis of a historical text with English translation; S.S. AHMED, *Southern Mesopotamia in the Time of Ashurbanipal* (1968), a study of Ashurbanipal's relations with his brother in Babylon; R.D. BARNETT, *The Sculptures of Ashurbanipal* (1971), a modern illustrated presentation of the bas-reliefs from the palace at Nineveh depicting the royal campaigns, hunting, and other activities.

(D.J.W.)

Asia

Asia is more a geographical term than a homogeneous continent. The most diverse of all continents, it extends over a latitudinal range of 92° from north to south, has the greatest range of land height of any continent, is subject to climates ranging from Arctic to tropical, and produces the most varied forms of vegetation and animal life in consequence. Similarly, its patterns of human adaptation range from the life-styles of the nomads of Arabia and Central Asia to those of the crowded cities of the Yangtze Basin and the Gangetic Plain.

Asia has to be described mainly in superlatives. It is the largest of the continents, occupying 30 percent of the world's land area, with a mainland area of approximately 17,000,000 square miles (44,000,000 square kilometres). While to the east the Pacific Ocean forms its natural boundary, the chains of islands that include the component territories of Japan, Taiwan, the Philippines, and Indonesia also form part of Asia. To the west the boundary is more difficult to define but is generally regarded as running southward along the eastern foot of the Ural Mountains, after which it turns approximately southwestward to the northern shore of the Caspian Sea, from where it again runs generally southwestward to the Caucasus Mountains, which form the boundary until the Black Sea is reached; from the Black Sea, the coast of Asia Minor and the Mediterranean coast of the Levant

form Asia's western limits, after which the boundary runs south across the Isthmus of Suez and along the coast of the Arabian Peninsula.

Asia is the most populous of the continents. Its population, in the early 1970s, was estimated to be 2,164,000,000, representing more than half the human race. The continent includes the two most populous countries in the world—China and India—in addition to Japan and Indonesia, each of which has a population of more than 100,000,000; among non-Asian states these last two countries are surpassed in numbers only by the populations of the Soviet Union (the territory of which also extends into Asia) and the United States. In the 20th century the population of Asia has been growing faster than the world average, its growth rate in the early 1970s being a little more than 2 percent a year. Despite the general awareness in Asia, especially in the most populous countries, of the need for some regulation of this growth, it appears that—barring some unexpected breakthrough in technology that would make large-scale limitation of population possible—the population of Asia may well reach 3,800,000,000 by the year 2000.

Origin of the name Asia

The name Asia is very ancient, and its origin in antiquity has been variously explained. The Greeks used it to designate the lands situated to the east of their homeland. It is also believed that the name may be derived from the Assyrian word *asu*, meaning "east." Another possible explanation is that it was originally a local name given to the plains of Ephesus and gradually extended to include Anatolia (contemporary Asia Minor) and the rest of the continent.

Geological development has resulted in a configuration that includes the Arabian massif (mountainous mass) and the plains of Iraq; the mountain belts of Turkey, Iran, and Afghanistan; the great Himalayan mountain chain stretching from Afghanistan to the Burmese peninsula; the Indo-Pakistan subcontinent; Central Asia; the vast Siberian lowlands extending from the Urals to the Pacific; and the great island chains that sweep in arcs from Japan to Indonesia.

As a result of this configuration, Asia's population is unevenly distributed. Thus, while there has been a concentration of population on the Arabian plains and river valleys, in the Indo-Pakistan subcontinent, in Central Asia, and especially China, and to some extent in the Pacific borderlands and on the islands, there are vast areas with a low density of population.

Siberia, which represents approximately one-third of the land area of Asia, has an estimated population of fewer than ten persons per square mile (four persons per square kilometre), compared with the average density of 180 persons per square mile in the rest of Asia.

The mountain systems of Central Asia not only have provided the great rivers with water from their melting snows but also have formed a natural barrier that has made the movement of peoples possible only through mountain passes. As a result the historic movement of population has been broadly from the arid zones of Central Asia through the mountain passes into the Indo-Pakistan subcontinent. Another movement has been from China through Southeast Asia to modern Indonesia and Malaysia. There has also been movement from the Arabian Peninsula and from India across the Bay of Bengal into Indonesia and Malaysia. The Japanese people and, to a lesser extent, the Chinese have remained ethnologically more homogeneous than the populations of other Asian countries.

Asia's religious heritage

Asia has been the birthplace of all the great world religions, including Buddhism, Christianity, Hinduism, Islām, Judaism, Sikhism, Taoism, and Zoroastrianism. Of these, only Christianity moved westward; it subsequently exerted little influence, in its religious aspects, on Asia, although many Asian countries have Christian minorities. Buddhism has had a greater impact outside its birthplace in India and is prevalent in various forms in China, Korea, Japan, the Southeast Asian countries, and Sri Lanka (formerly Ceylon). Islām has spread out of Arabia eastward to Afghanistan, Pakistan, and India, and thence to Malaysia and Indonesia, as well as west and south to several areas of Africa. Hinduism, basically a non-proselytizing religion, has been mostly confined to the Indian subcontinent.

Before the major industrial revolution that occurred in the 18th and 19th centuries, most of the principal technical achievements began in Asia. Three millennia before Christ, Asians knew the arts of cooking and pottery and the use of fire for the smelting of ores. They had already progressed from the nomadic stage to the established life of cultivators, using irrigation and practicing crop rotation. They had learned to domesticate animals and had invented the wheel, the harness, the saddle, and the chariot. They had begun to use a form of paper and had developed elaborate scripts. Familiar with wood carving, stonecutting, and the casting of metals, they have left monuments of stone and metal that to this day evoke admiration and astonishment. They used the art of calculation, including the decimal system, and employed means of measurement. Various indigenous systems of medicine were developed, which have stood the test of time and are still in use. In ancient times major Asian empire-states arose, employing intricate systems of laws and regulations and delegating power and authority to institutions of government at various levels.

In the 15th century, Asia's material advance was envied by people in Europe. The port of Venice was hailed as Europe's window to the East. The fabled prosperity of Asia prompted the movement of adventurous spirits from Western nations to Asia and led to the eventual conquest of many Asian countries. The Western impact on the civilizations of these Asian countries, on their mental attitudes, and on the development of their political institutions and economic systems can only be referred to briefly here. Even though almost all Asian countries have attained their independence, the era of European conquest and colonization has left an indelible mark on much of Asia. The Asian countries received many benefits by way of the establishment of an infrastructure, a system of posts and telegraphs, a railway and road network, ports and harbours, public health systems, modernized educational systems, and also the rudiments of an apparatus of modern government—civil, judicial, and military—and ideas for publishing newspapers and magazines. On the other hand, there was little progress in industrialization. The effect of the neglect of industrial development during these years of colonization is still noticeable in commercial relations among Asian countries. Broadly, it may be said that the role of the Asian countries—with the exception of Japan, which did not feel the impact of conquest and colonization—was that of primary producers exporting their products to the metropolitan countries, where they were turned into finished products and re-exported back to the Asian countries. Today, as a result, most Asian countries have economies that are non-complementary; recent efforts at the international and regional levels to produce greater economic integration and coordination among Asian countries have not entirely succeeded.

The impact of the West

In addition to this factor, the lack of any ethnic, religious, or cultural homogeneity among the Asian peoples has made it difficult for a common Asian political consciousness to arise. In the struggle for independence most of the Asian countries developed a political philosophy favouring nationalism. Furthermore, since achieving independence, many Asian countries seem to have lost the unifying force of emergent nationalism and are groping to find a substitute for it. Even the doctrine of nonalignment, which seemed to be common to several Asian countries, has lost its impetus, and in the early 1970s the centre of interest in this direction appeared to have moved from Asia to Africa.

In the economic field, Asian countries are caught between two conflicting forces. On the one hand, the need to ensure improved levels of living has led to a desire to move toward orderly and planned economic and social development. On the other hand, in many countries there has been a long tradition of other-worldliness and austerity. This conflict of forces may produce a new philosophy of economic development in Asia that will reconcile the

need for material progress with the imperative of not abandoning spiritual and religious values.

Asia's entry into the nuclear age

The political forces that are emerging in Asia cannot be ignored. China has begun to play an important role in world affairs; it is as yet the only nuclear power in Asia. Japan has emerged as one of the great industrial nations of the world and clearly has the capability to develop into a nuclear power, although the Japanese, in the light of their own experience at the end of World War II, have accepted a constitution renouncing war as a sovereign right of the nation and abjuring the threat or use of force as a means of settling international disputes. India perhaps has the technology to become a nuclear power, but to do so would involve such an enormous sacrifice of resources needed for economic development that it is doubtful whether it will be done.

The recent political history of the continent is to some extent the result of the political vacuum created by the withdrawal of the colonial powers from Asia. In the early 1970s the disturbed political situation at one end of the continent, in the Middle East, and at the other end the equally confused and dangerous political situation in the countries of Indochina formerly under French administration provided two illustrations of this fact. The difficulties that have persisted between India and Pakistan since their independence in 1947 and between Indonesia and Malaysia between 1963 and 1968 are clearly part of the same pattern.

With a new awakening of political consciousness, with a new spirit of regional cooperation, and with a new urge at the grassroots level for the improvement of living standards, there are positive factors at work that should make possible a resurgence of Asia in the coming decades. The rate of population growth constitutes the only negative factor. Several Asian countries have shown, however, that even under existing circumstances great progress can be made in the limitation of population growth. If this trend were to become more general and widespread, the chance that Asia will enter the 21st century economically stronger and socially better developed would be correspondingly increased. (This article covers the general physical and human geography of Asia. For historical aspects, see COLONIALISM and INNER ASIA, HISTORY OF, as well as historical articles on individual nations. There are separate articles on the major geographical features of Asia, as well as on Asian nations and major Asian cities. For related geographical articles, see PRECAMBRIAN TIME and CONTINENTS, DEVELOPMENT OF.) (C.V.N.)

This article is organized as follows:

I. Geological history
  Elements and processes in the making of the continent
  The territorial formation of Asia
  The pattern of Asia's paleogeographic development
II. Physical geography
  Relief
  Climate
  Drainage
  Soils
III. Vegetation and animal life
  Vegetation
  Animal life
IV. Natural resources
  Mineral resources
  Water resources
  Biological resources
V. Human resources
  Evolution of the ethnic pattern
  Population distribution and regional ecology
  Forms of ethnic administration
VI. Political geography
  Historical development
  The contemporary pattern
VII. Resource development
  Industries
  Power
  Agriculture, irrigation, and land use
VIII. Asian commerce
  Transportation
  Internal trade
  External trade
  Commercial prospects
IX. Demographic patterns
  Present population patterns
  Future trends

I. Geological history

ELEMENTS AND PROCESSES IN THE MAKING OF THE CONTINENT

Platforms, shields, and geosynclines. Asia consists, in part, of several platforms that have been subjected to almost no folding since the Precambrian Era, which lasted from 4,600,000,000 to 570,000,000 years ago. It also consists of vast regions of folding, formed later in various periods or ages, that originated in geosynclines (large and generally linear troughs that gradually subsided over long periods of time and subsequently became filled with thick accumulations of sediments). Contemporary geosyncline systems, still not consolidated by folding, join the shores of the Asian mainland to the Malay Archipelago, underlie bordering seas such as the Sea of Japan and the Sea of Okhotsk, and extend to the island arcs of the Pacific Ocean.

The four continental platforms

The principal continental platforms of Asia are the Siberian Platform in the north, the Chinese Platform in the east, the Indian Platform in the south, and the Arabian Platform in the southwest. Those parts of the platforms where the Precambrian crystalline bedrock is not covered by a sedimentary deposit are called shields; almost all the shield areas were dry land during the last 500,000,000 years and include the Aldan Shield in Eastern Siberia, the North Chinese Shield, the Indian Shield in peninsular India, and a large part of the Arabian Shield. In addition, the Russian Platform has played an important role in the geological history of western Asia. Along its western edge, the folded structures of the Urals and of Kazakhstan were formed at the end of the Paleozoic Era (which lasted from 570,000,000 to 225,000,000 years ago).

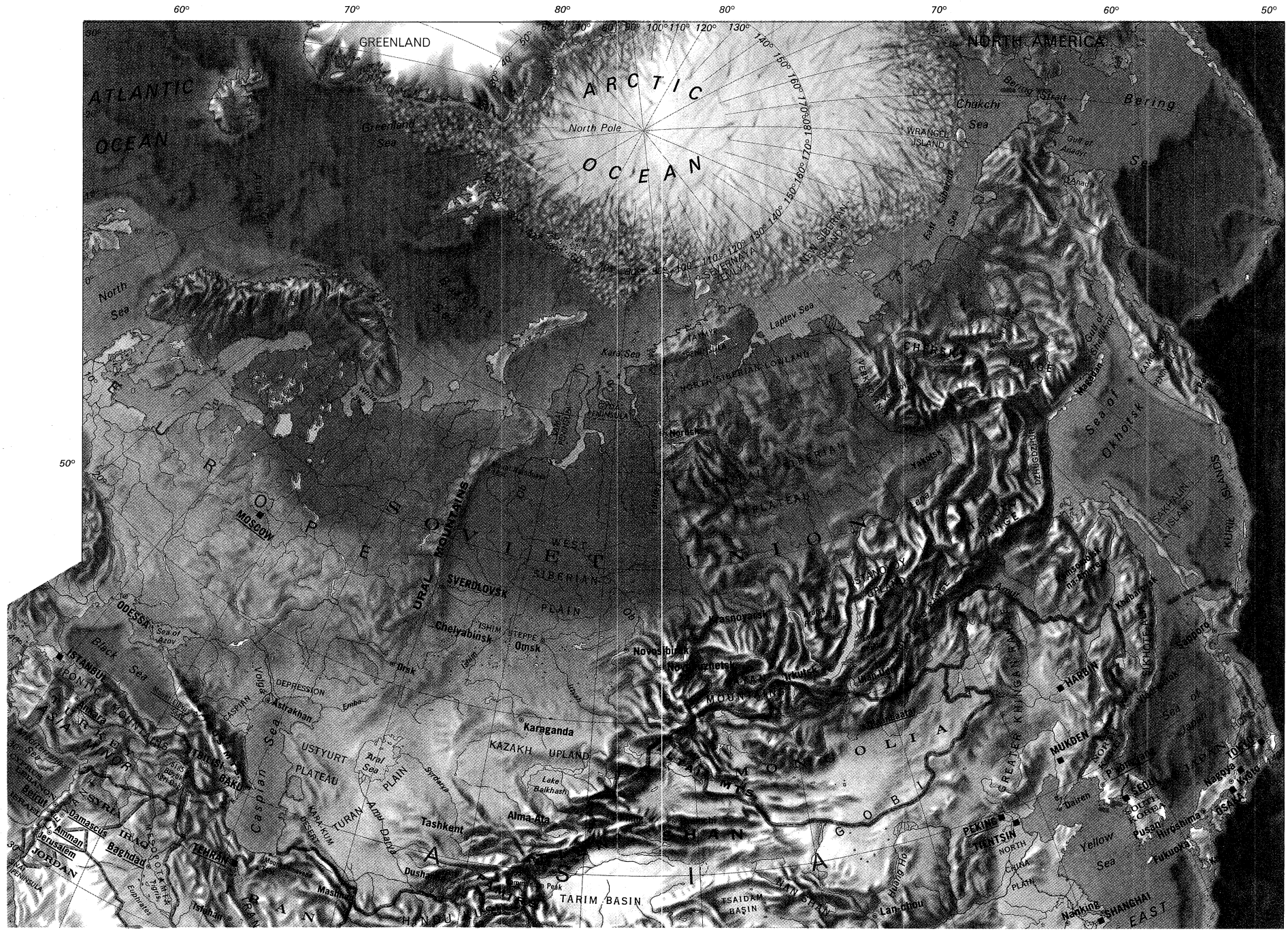
Endogenetic and exogenetic forces. Two forces have been at work to mold the Asian continent into its present configuration. On the one hand are the endogenetic forces, which represent vertical or lateral forces originating deep within the Earth and that produce fissures in the Earth's crust through which magma, or molten rock, wells up in the form of lava or forms intrusive bodies; when it solidifies, this rock, basalt, is described as igneous (solidified from the molten state). Exogenetic forces, on the other hand, represent those endless processes of erosion and sedimentation that take place on the surface, where rocks are subjected to weathering and denudation. Associated with these two forces is a principle of equilibrium called isostasy, which represents a theoretical balance maintained between large sections of the Earth's crust, which act as though they were floating on a denser underlying layer. In maintaining this balance, less dense material rises vertically while denser material sinks downward.

The mountain-building process. Mountain building, or orogenesis, is the result of folding or thrusting, which occurs when thousands of feet of sediment accumulate in geosynclines, causing them to sink deeper, thus either activating, or interacting with, vertical or lateral endogenetic forces. In this way the geosynclines themselves, representing zones of structural weakness between crustal blocks or platforms, are folded upward into new mountain systems—a process associated with the upwelling of granitic magma and the deformation of metamorphic rock (rock that has been altered in composition, texture, or internal structure as a result of heat and pressure).

THE TERRITORIAL FORMATION OF ASIA

Blocks and zones of folding

As a result of folding, the intrusion of granite, the swelling of the Earth's crust in response to pressures, the isostatic uplifting (uplifting caused by equal pressure on all sides) of mountains, and the drying up of marine basins that lay in the geosynclinal regions between the continental platforms, ancient blocks of the Earth's crust and weaker zones of folding were united to form the Asian continent. From the formerly inundated geosynclines,


60° 70° 80° 80° 70° 60° 50°
GREENLAND
NORTH AMERICA
ATLANTIC OCEAN
ARCTIC OCEAN
North Pole
Greenland Sea
Chukchi Sea
Bering Strait
Bering Sea
WRANGEL ISLAND
NEW SIBERIAN ISLANDS
SEVERNAYA ZEMLYA
East Siberian Sea
Laptev Sea
Kara Sea
North Sea
White Sea
Barents Sea
TAIMYR PENINSULA
NORTH SIBERIAN LOWLAND
CHERSKIY RANGE
Gulf of Anadyr
Gulf of Shelikhov
Sea of Okhotsk
KAMCHATKA PENINSULA
KURIL ISLANDS
SAKHALIN ISLAND
50°
40°
30°
EUROPE
ASIA
MOSCOW
URAL MOUNTAINS
WEST SIBERIAN PLAIN
CENTRAL SIBERIAN PLATEAU
Norilsk
Yakutsk
Lena
Aldan
Stanovoy Range
Amur
Khabarovsk
Komsomolsk
Nikolayevsk
Sea of Japan
SVERDLOVSK
Chelyabinsk
TSHIM STEPPE
Omsk
Novosibirsk
Novokuznetsk
Krasnoyarsk
Irtysh
Ob
Yenisey
MONGOLIAN PLATEAU
HARBIN
MUKDEN
Dairen
PEKING
TIENTSIN
NORTH CHINA PLAIN
Yellow Sea
SEOUL
Pusan
Hiroshima
Nagoya
TOKYO
Fukuoka
ISAKI
SHANGHAI
Nanking
EAST CHINA
Hwang Ho
Lan-chou
ODESSA
Sea of Azov
Black Sea
ISTANBUL
ANATOLIAN PLATEAU
Volga
Astrakhan
CASPIAN DEPRESSION
Caspian Sea
Ural
Emba
Aral Sea
USTYURT PLATEAU
KARA-KUM DESERT
TURAN PLAIN
Amu-Darya
Syrdarya
Karaganda
KAZAKH UPLAND
Lake Balkhash
Alma-Ata
Tashkent
TIEN SHAN
Dushanbe
Pamir Peak
TARIM BASIN
TSAIDAM BASIN
Gobi Desert
GREATER KHINGAN RANGE
Black Sea
Caspian Sea
Aral Sea
Lake Balkhash
Yellow Sea
East China Sea
Sea of Japan
Sea of Okhotsk
Sea of Azov
Gulf of Anadyr
Gulf of Shelikhov
Bering Strait
Greenland Sea
White Sea
Barents Sea
Kara Sea
Laptev Sea
East Siberian Sea
North Sea
Ural Mountains
Caspian Depression
Ustyurt Plateau
Kara-Kum Desert
Turan Plain
Tien Shan
Tarim Basin
Tsaidam Basin
Gobi Desert
Greater Khingan Range
North China Plain
Anatolian Plateau
West Siberian Plain
Central Siberian Plateau
Mongolian Plateau
Kamchatka Peninsula
Kuril Islands
Sakhalin Island
Taimyr Peninsula
North Siberian Lowland
Cherskiy Range
Stanovoy Range
Ural
Volga
Ob
Yenisey
Lena
Aldan
Amur
Irtysh
Hwang Ho
Yellow
Sea
Sea of Japan
Sea of Okhotsk
Sea of Azov
Black Sea
Caspian Sea
Aral Sea
Lake Balkhash



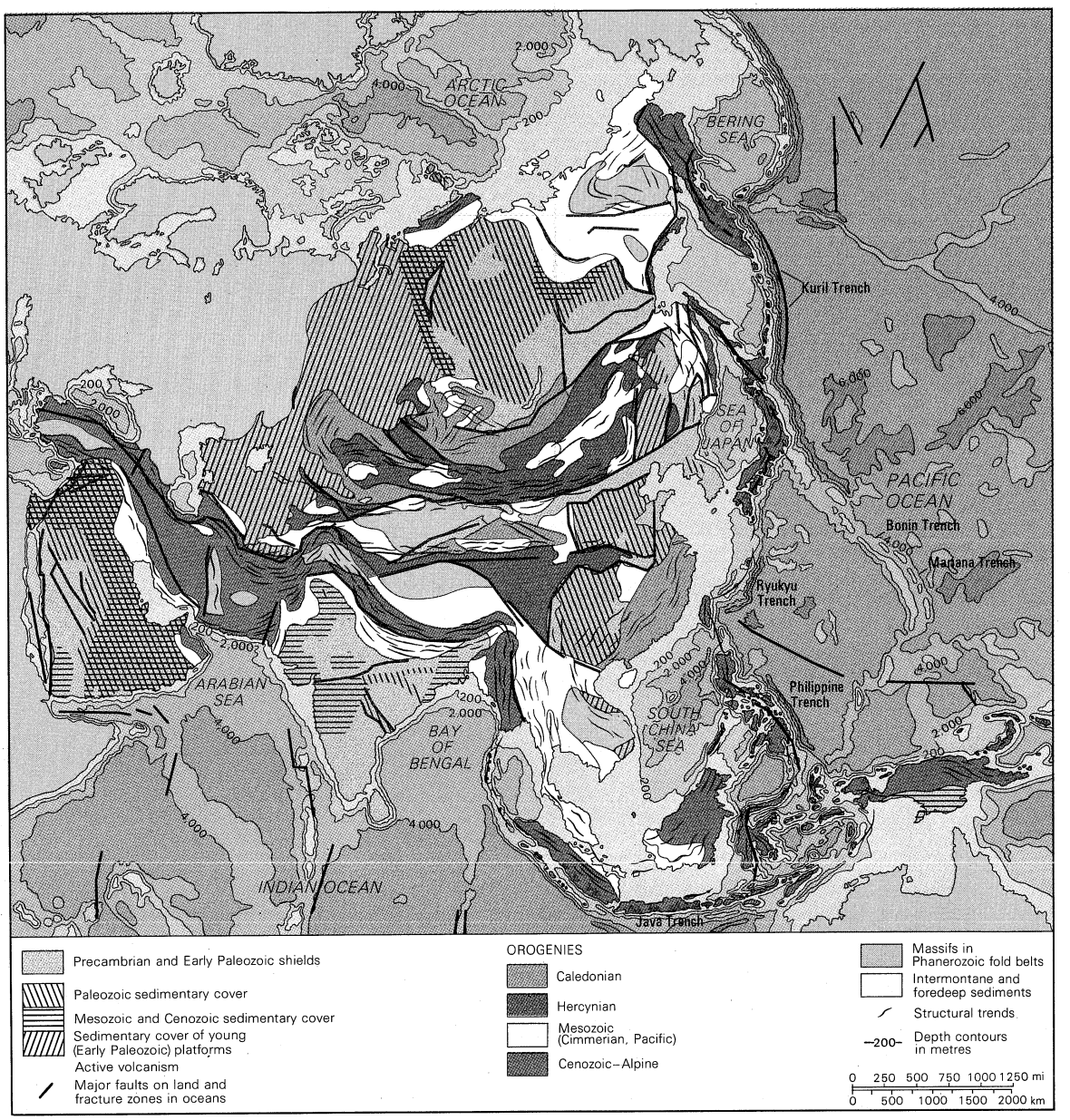
ASIA

Size of symbol indicates relative size of town • ◉ ■

Elevations in metres

0 200 400 600 800km

0 200 400 600 800mi



Structural features of Asia.

several belts of folding—including the Alpine–Himalayan belt—were formed in the Late Paleozoic Era, the Mesozoic Era (from 225,000,000 to 65,000,000 years ago), and the Cenozoic Era (which is dated from 65,000,000 years ago to the present time).

The composition of the continental platforms. *The Siberian Platform.* Greatly deformed schists (metamorphic rocks containing micaceous minerals), gneisses (rocks in which bands rich in granular minerals alternate with bands containing schistose materials), and granites of the Precambrian form the basement of the Siberian Platform. Extending between the river basins of the Yenisey and the Lena, it surfaces within the boundaries of the Aldan Shield, north of the Stanovoy Mountains, and the Anabar Massif (mountainous mass), west of the Lena River and south of the Arctic. In other areas the crystalline bedrock is overlain almost horizontally by sedimentary rock, including marine and salt lagoon deposits from the Cambrian Period (from 570,000,000 to 500,000,000 years ago), marine deposits from the Ordovician Period (from 500,000,000 to 430,000,000 years ago) and the Silurian Period (from 430,000,000 to 395,000,000 years ago), as well as deposits from the Permian Period (from 280,000,000 to 225,000,000 years ago) containing strata of coal. In the western part of the platform, primarily along the edges and in the centre of the broad Tungus syncline (a basin formed by a downward bend of rock strata), east of the Yenisey River, the Paleozoic strata are cut by numerous intrusive sills (layers) of a dark, fine-grained, basaltic rock known as trap-rock, locally associated with copper and nickel ores and volcanic pipes (cylindrical veins of volcanic origin) containing diamond-bearing kimberlite (an ultrabasic rock of a subcrustal origin). In the southern part of the platform, the lower layers of the sedimentary cover consist of thick layers of Precambrian marine sediments analogous to the Sinian System in China. They were slightly disturbed at the time of the Baikal folding, which occurred at the end of the Precambrian and during the Cambrian Period. Along the northern and eastern edges of the platform, there is widespread distribution of coal-bearing deposits from the Jurassic Period (from 190,000,000 to 136,000,000 years ago) and the Cretaceous Period (from 136,000,000 to 65,000,000 years ago). Mesozoic sedimentary series reach particularly great thickness in the depressions of the Vilyuy River and in the depression near the Verkhoyansk Mountains, which run parallel to the Lena River, both areas containing deposits of fuel gas.

The Chinese Platform. The Chinese Platform consists of four separate massifs—the North Chinese, the South

The four Chinese massifs

Chinese, the Tarim, and the Tibetan—which probably constituted a single platform in the Precambrian. This platform appears to have broken up at the beginning of the Paleozoic Era, at which time geosynclinal downwarps were formed between the separated blocks. As a result of the folding that took place in these downwarps during the Paleozoic and Mesozoic eras, mountain ranges arose, including the Kunlun and the Tsinling (both in China). The Precambrian bedrock is exposed on the surface primarily within the boundaries of the North Chinese massif, which contains large deposits of ferruginous (iron-bearing) quartzite.

During the Silurian Period and the Devonian Period (from 395,000,000 to 345,000,000 years ago), the southern part of the platform—as well as almost all of it during the later part of the Cretaceous Period—was subjected to folding and faulting deformations that were considerably more intense than those experienced by platforms elsewhere in the world. These deformations were accompanied by intrusions of ore-bearing granite.

The sedimentary rocks overlying the folded bedrock (basement) of the Chinese Platform represent, for the most part, marine sediments of the Precambrian and Lower Paleozoic (Cambrian, Ordovician, and Silurian) eras; marine and continental deposits from the Devonian, Carboniferous (from 345,000,000 to 280,000,000 years ago), and Permian periods, some of which are coal-bearing; and thick layers of rocks of the Mesozoic Era, as well as layers from the Cenozoic Era in some of the depressions. Among the Mesozoic deposits there is an abundance of red fragmented rocks that were formed when the prevailing climate was hot and dry. The sedimentary cover is thickest in the synclines of the Ordos Desert region of Inner Mongolia, of the Chinese province of Szechwan, and of the Lower Huang Ho, as well as in the Mesozoic downwarp of southeastern Korea.

The Arabian Platform. The Arabian Platform, like the Indian Platform, is usually considered to have formed a part of the continent of Gondwana, which began to split apart during the Paleozoic Era. Gondwana is believed to have included Africa, Australia, South America, and possibly Antarctica, and the two platforms do have much in common with these continents. The ancient basement of Arabia is composed of Precambrian crystalline rocks, which crop out in the western part of the peninsula as the Arabian Shield. To the north and east the bedrock is buried under thick layers of horizontal or only slightly disturbed Paleozoic rocks; these are mainly sediments from the Jurassic and Cretaceous periods and from the earlier part of the Cenozoic Era. These sedimentary strata lie within the bounds of the Mesopotamian downwarp, which occurs in the Persian Gulf region; here the bedrock has sunk to a depth of about two and a half miles and contains large amounts of petroleum.

The crystalline bedrock

The Indian Platform. The Indian Platform covers a large part of peninsular India and Sri Lanka. It consists of a Precambrian crystalline shield in which broad grabens (blocks that have been downthrown along faults in the rock on either side) along the Godāvari River and along the Dāmodar River farther to the northeast are overlain by horizontal deposits of the Gondwana system. These were laid down between early Carboniferous and late Jurassic time. The Deccan Plateau, which forms the backbone of peninsular India, is itself overlain by basaltic lavas (traprock from the Late Cretaceous Period and the Early Cenozoic Era). Along the edges of the platform the bedrock is covered by marine deposits from the Jurassic, Cretaceous, and Tertiary (from 65,000,000 to 2,500,000 years ago) periods; in the Gulf of Cambay region, off the west coast of India, these contain petroleum deposits. To the north, the bedrock lies under Cenozoic deposits that reach their greatest thickness in the northern part of the Ganges River Basin, in the marginal downwarp at the foot of the Himalayas, and in the Indus Basin. Deposits of gas and petroleum are associated with the Cenozoic marine and continental sediments, while the Precambrian rocks of the shield contain iron and manganese ores. A small Precambrian outcrop, the Shillong Plateau massif, occupies the eastern part of the platform.

Mountain folding. *The Late Precambrian and the Paleozoic eras.* Those Asian zones that were folded during the Late Precambrian and Early Paleozoic eras consist of various kinds of weakly and strongly metamorphosed volcanic and sedimentary rocks. They are cut by intrusions of granite, granodiorite (a quartz-bearing rock formed at great depth), gabbro (a coarse-grained, dark igneous rock), diabase (a rock of basaltic composition), and serpentinous ultrabasic rocks (*i.e.*, igneous rocks containing less than 45 percent of silica). Deposits of gold, copper, tungsten, polymetallic (*i.e.*, containing many metals), and other ores are associated with these rocks in many areas.

Regions of Late Precambrian and Early Paleozoic folding border the Siberian Platform to the west and south. They compose the Yenisey Ridge, the Eastern and Western Sayan ranges, the Kuznetsk Alatau, the higher (alpine) part of the Soviet Altai, the Mongolian Altai, and a considerable part of the Khangay and Tannu-Ola ranges. Caledonian folding—which is to say, folding that took place during the Caledonian (*i.e.*, Silurian–Devonian) orogeny (mountain-building process)—is also found in the Taymyr Peninsula and on the Severnaya Zemlya Islands, in central Kazakhstan, in the northern chains of the Tien Shan, in the Kunlun Range, and in southeastern China. In the portion of the Caledonian orogeny that occurred during the Devonian Period, a number of large depressions were formed. These included the Kuznetsk and Minusinsk basins and the depressions in the central part of Tuva. Large reserves of coal are concentrated in the deposits laid down in these regions in the Carboniferous and Permian periods.

Areas of depression

The fold structures of the Hercynian orogeny were formed as the result of tectonic activity (movements of the Earth's crust) that occurred during the Middle Paleozoic (Devonian) and Late Paleozoic (Carboniferous and Permian) eras; these constitute a broad arc traversing the central part of the continent. The Hercynian fold struc-

Picturepoint—Publix



Jabal Tuwayq, a prominent escarpment that parallels the bulge in the Arabian Shield near Riyadh, in central Saudi Arabia.

tures have a northeasterly and southerly trend between the Russian and Siberian platforms, and a northwest trend in Kazakhstan, along the Salair Ridge, and in the southern Altai; run east to west in the Tien Shan, Kunlun, and Tsinling ranges; and run northeast in northeastern China. According to geophysical data, the Hercynian fold system of the Urals, covered toward the south by a horizontal layer of Mesozoic and Cenozoic deposits, is joined with the Tien Shan. In addition, a branch proceeds from it which passes through the Mangyshlak Peninsula toward the Donets Basin.

A considerable area of the original Paleozoic fold zones was levelled by erosion and subsequently sank, forming large depressions that were filled to an overall depth of between one and four miles with virtually undisturbed marine and continental sediments of the Mesozoic and Cenozoic eras. Associated with these depressions are the vast West Siberian Plain; the Turgay Downwarp, located to the east of the southern Urals; the Turan Platform in the Amu Darya and Syr Darya basins; the depressions of Lake Balkhash and of the lower course of the I-li Ho (river); the Dzungarian Basin; the Fergana and Turfan depressions; the Tsaidam Basin; the Tungliao or Manchurian lowlands in the Sungari River Basin; and others. Only the intermontane depressions, those of Fergana and Turfan, were affected by slight folding; in the other depressions the strata lie almost horizontal. Deposits of petroleum and gas are found in almost all these depressions, from Western Siberia to the Tsaidam Basin and Manchurian Plain in China.

Mesozoic folding zones

The Mesozoic Era. The belt of Mesozoic folding includes northeastern Siberia (the Verkhoyansk and Chersk ranges), the Sikhote-Alin Range in the Amur region, a large part of Indochina, and possibly the Trans-Himalayas. Folded structures, intersected by numerous granite intrusions, were formed in all these areas as a result of the crumpling of thick layers of geosynclinal deposits of the Permian, Triassic (from 225,000,000 to 190,000,000 years ago), Jurassic, and Early Cretaceous periods. Associated with these folding belts are gold, tin, and polymetallic ores. Moderately large blocks—such as the Kolyma, Indosinian, and other massifs, which resemble the Precambrian or Paleozoic platforms in their structure—are enclosed between the branches of the broad Mesozoic folding zones.

Mesozoic folding and the upwelling of magma (molten rock) also affected adjacent areas, such as the Hercynian mountain zone of Mongolia and the Transbaikalia, the Caledonian zone of South China, and a considerable part of the Chinese Platform. Deposits of tungsten, tin, mercury, and other metals are found in these areas.

The Cenozoic Era. During the Early Cenozoic Era, a volcanic belt was formed that extends from the Chukotsk Peninsula and the shores of the Sea of Okhotsk through the eastern slopes of the Sikhote-Alin Range, in South Korea, to the southeast coast of China. It runs approximately along the border of the Mesozoic and Cenozoic fold zones. This volcanic belt is composed of massive accumulations of basaltic, andesitic, and acidic lavas dating from the Cretaceous and Early Tertiary periods, with many granite intrusions. It would seem that a series of volcanoes extended through this belt in a long arc, resembling the volcanic island chains of East Asia.

The two Cenozoic fold belts

Areas of Cenozoic folding are confined to two zones—the Alpine–Himalayan belt that traverses Asia from west to east, and the Pacific Ocean belt that runs through the island arcs to unite with the Alpine–Himalayan belt in Indonesia. The Alpine–Himalayan belt was essentially formed on the site of the extensive Tethys Geosyncline. This geosynclinal ocean, the remains of which are to be seen in the Mediterranean and Black seas and in the marine basins of Indonesia, divided two distinct continents—Gondwanaland and Angara, or Angarida—during the Paleozoic, Mesozoic, and Early Cenozoic eras. In the Alpine–Himalayan belt itself there may be distinguished two—and in places three—curving folded mountain ranges; at times these are close together, and at times they diverge. These ranges represent slabs or blocks that were overthrust onto one another or onto the edges of the neighbouring platforms, forming gigantic and complex anticlines (convex folds). The internal parts of their folded structures were sometimes already formed during the Mesozoic Era from Mesozoic and Paleozoic sediments that had accumulated in the Tethys Geosyncline, while some slopes and foothills were formed of Tertiary deposits as a result of less violent folding. The thick sand–clay strata of the Cenozoic Era, which accumulated at the base of the mountains and in depressions in front of them, often contain gas and petroleum.

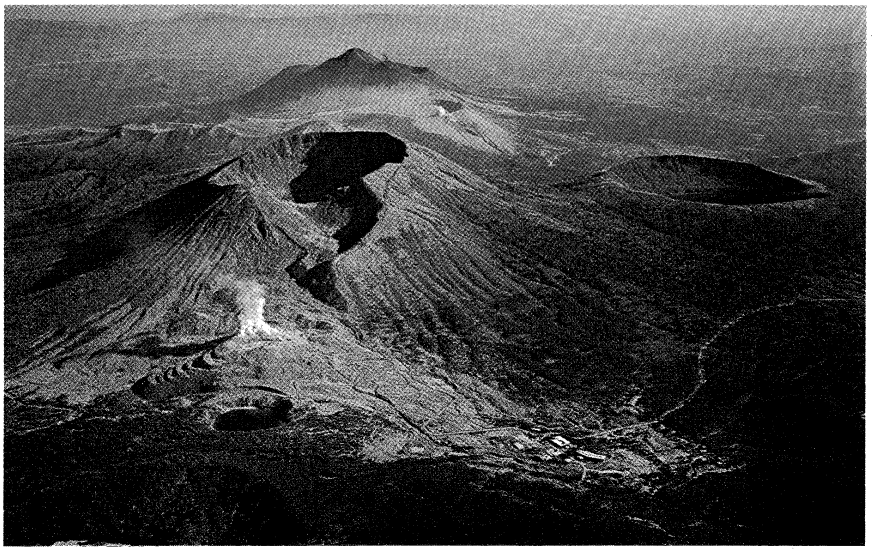
Northern and southern anticlinoria

The northern series of alpine anticlinoria (*i.e.*, series of anticlines and synclines so arranged that together they form a general arch or anticline) is formed by the Greater Caucasus Mountains between the Black and Caspian seas; the Turkmen–Khorasan ranges east of the Caspian; and the Safīd Kūh Selseleh-ye (Paropamisus), Pamir, Gissar, and Alai ranges. West of this system of ranges runs still another series of folded structures, separated from the first series by deep troughs. These consist of the Pontic Mountains, the Lesser Caucasus, and the Elburz Mountains. In the related depressions—the Black Sea, the Kolkhida (Rion) and Kura-Aras Lowland, and the South Caspian basin—are concentrated layers, from three to six miles thick, of Cenozoic deposits, which are folded along the edges of the depressions. The southern series of anticlinoria is composed of the ranges of the Tavr, Zagros, Makran and Soleymān mountains, the Hindu Kush, and the Himalayas. Between the northern and southern series of folded mountain ranges are situated the massifs of Menderes and Kirshekhir in Turkey, and the central Iranian massif. The small Georgian block occupies a similar position between the Greater and Lesser Caucasus. All these massifs represent regions consolidated by Precambrian or Paleozoic folding, with surface outcrops of bedrock in some places; in others, the bedrock is covered by weakly folded sedimentary deposits of the Paleozoic, Mesozoic, and Cenozoic periods.

The youngest geosyncline system

Mountain folding in progress. In the folded ranges of Burma, Malaysia, and Indonesia there occurs a transition from the belt of alpine folding—formed on the site of geosynclines that were already landlocked and drained—to a contemporary geosyncline system in which folding is not yet complete. This system, the youngest, occupies the area between the Asian continent itself and the Pacific Ocean and includes the folded systems of the Koryaken Range; the Kamchatka Peninsula; Sakhalin Island; Borneo; Celebes Island; the archipelagoes of the Komandorskiye (Commander), Aleutian, and Kuril islands; and the islands of Japan, the Ryukyus, Taiwan, the Philippines, the Moluccas, and the Sundas. Within its structure, geoanticlinal uplifts, consisting of island arcs and mountainous peninsulas, may be distinguished, as well as intermontane troughs, the contemporary geosynclinal downwarps of the seas bordering Asia, and deep-sea trenches on the periphery of the Pacific and Indian oceans. The axial parts of the folded ranges and island arcs in this zone are usually composed of Mesozoic and Upper Paleozoic deposits; they are cut by recent (Tertiary) intrusions and are crowned with a series of active volcanoes. The Tertiary deposits of the intermontane troughs attain great thickness and contain petroleum deposits, such as those of Sakhalin, Japan, and Indonesia; they were folded during the Miocene Epoch (from 26,000,000 to 7,000,000 years ago) and the Pliocene Epoch (from 7,000,000 to 2,500,000 years ago). In some places there are folds in the deposits sedimented in the Quaternary Period (which began 2,500,000 years ago), and as a result of large and recent fractures, Cenozoic lavas consisting of basalts and andesites cover vast areas of the island regions of East Asia.

The Alpine–Himalayan belt and, in particular, the Pacific Ocean belt are both characterized by the active tectonic processes peculiar to geosynclinal systems. A comparatively rapid horizontal shifting of different parts of the Earth's crust at a speed of between one-quarter and three-quarters of an inch a year is taking place, according to geodetic measurements taken in Japan and Tadzhikistan. Also characteristic is intensive vertical movement. This occurs in the form of the uplifting of geoanticlines,



Kirishima Range on Kyushu, Japan's southernmost island; it includes several active volcanoes.
Kokunai Jigyo Kouku

accompanied by the sinking of neighbouring depressions and strong seismic activity. Disturbances of the isostatic equilibrium are concentrated in zones where the contrasting nature of the vertical movement is the most clearly manifest. The foci of earthquake shocks are not confined only to those fractures that rupture the Earth's crust but include also the deeply buried zones of folding that are associated with fractures. These zones are tilted from the deepwater trenches of the Indian and Pacific oceans toward the Asian continent at angles of from 20° to 70°. The sources of the earthquake shocks lie from 10 to 450 miles deep (about 150 miles deep in the Hindu Kush region). The characteristic movement originating from these foci is overthrust folding and upthrusting, and it indicates extreme compression of the Earth's crust. It is assumed that along the deep fractures there occurs an overthrusting of the island arcs onto the floor of the Pacific Ocean, an underthrusting of the Indian Platform under the Himalayas, and an analogous movement along the buckled edges of all the other outlying downwarps and deep-sea trenches. A considerable thickening of the Earth's crust is taking place in the high mountain regions of Central Asia and the Himalayas; here the Earth's crust is up to 40 or 50 miles thick, as opposed to 20 to 25 miles thick on the low plains and flatlands. This thickening may also be related to the lateral compression that is folding the Cenozoic strata.

Foci of earthquake shocks

THE PATTERN OF ASIA'S PALEOGEOGRAPHIC DEVELOPMENT

The Precambrian era. Although, on the crystalline shields of Asia, rocks may be found that are known to have been formed 3,000,000,000 years ago, an adequate description of the paleogeography of the continent—that is to say, of its geography in different eras and periods of geological time—can be given only for the last 1,000,000,000 years. At the beginning of this period, primitive forms of animal and plant life appeared, primarily as algae, when the vast marine basins of the Precambrian covered the geosynclinal regions of the Urals, the Tien Shan, the Altai, the Western Sayans, and South China, as well as the southern parts of the Siberian Platform and the entire Iranian Massif. At that time continental sediments, including glacial deposits, accumulated over a considerable part of the Chinese Platform, while regions of erosion where detrital material was removed were represented by the shields.

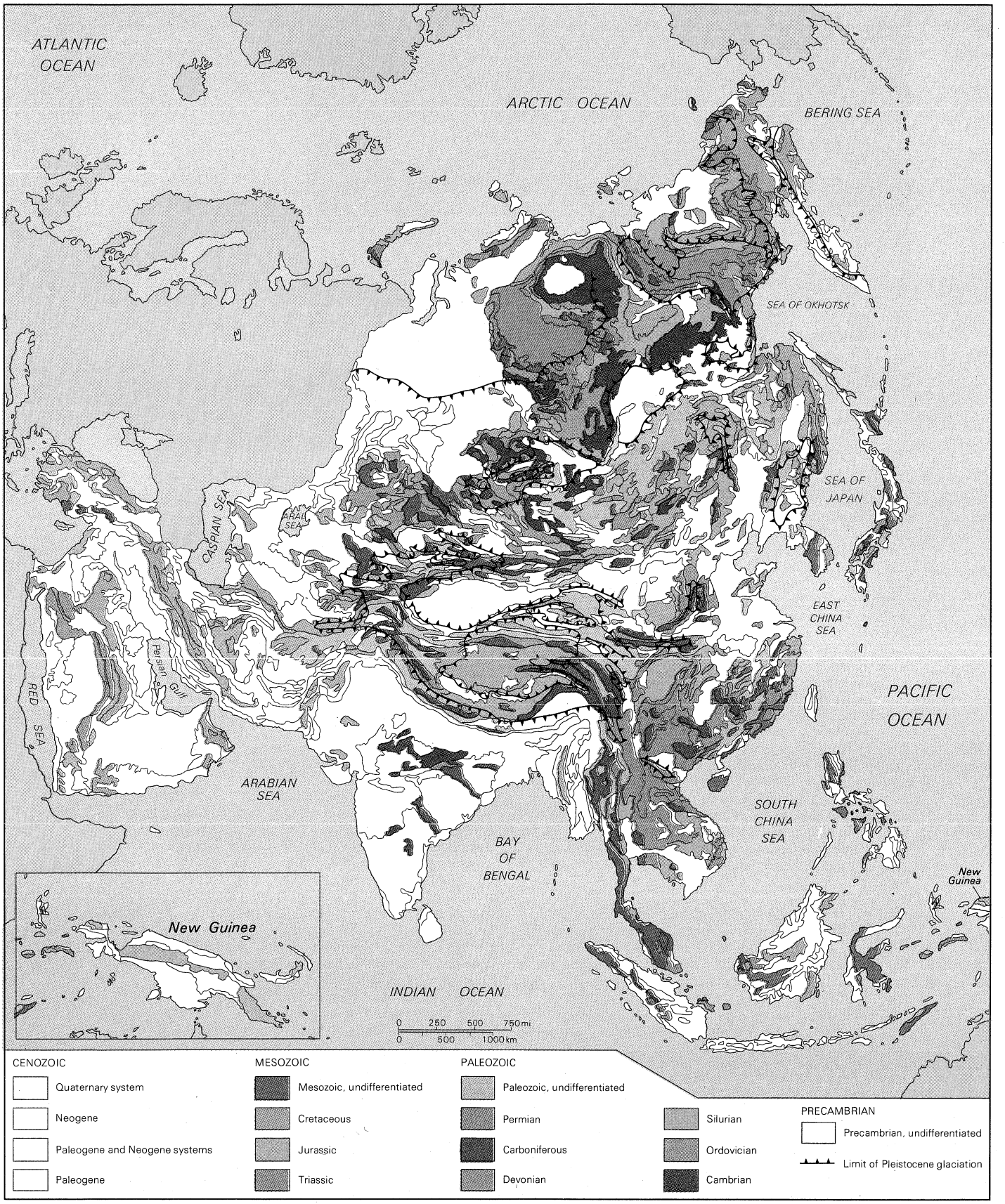
Asian seas of the Precambrian

The Cambrian and Ordovician periods. During the Early Cambrian Period, the seas began to transgress, covering almost all of the Siberian Platform, a large part of the Chinese Platform, and the northern parts of the Arabian and Indian platforms. The transgression reached its maximum extent in the middle of the Cambrian Period, and the seas began to recede during the Late Cambrian, particularly from the Siberian Platform. But the sea advanced again during the Early Ordovician Period, covering almost the entire area of the Chinese Platform. The Cambrian and Ordovician deposits on the platforms were primarily limestone, but there were also red sand–clay rocks. Deposits of rock salt were associated with the sand–clay rocks in the southern part of the Siberian Platform, in the northwestern part of the Indian Platform, and in the eastern part of the Arabian Platform. Such deposits, and the nature of the marine animal life, indicate that these three platforms were situated in a hot climatic zone; judging from the paleomagnetic data, the Equator then passed through the southern part of the Siberian Platform.

The Silurian Period. The seas receded during the course of the Silurian Period; red sediments and evaporites (sediments formed as a result of evaporation) continued to be laid down in the western half of the Siberian Platform and in other areas. At the end of the Silurian Period and in the course of the Devonian Period, all the platforms were uplifted and became dry land, except for some of the marginal areas and freshwater basins. Meanwhile, during the Cambrian, Ordovician, and Silurian periods, marine sandy-clayey and carbonate sediments, tuffs (rocks formed of compacted volcanic fragments), and lavas continued to be deposited in the geosynclines that separated the Russian, Siberian, and Chinese platforms from each other, as well as from Gondwanaland. The basins, which were separated from each other by islands, gradually dried up as the result of the Baikal and Caledonian folding; these included the basins of the Sayans, a considerable part of the Altai, the Khangarulsk, Tannu-Ola, central Kazakhstan, Kunlun, and South China. Large land areas were also subjected to volcanic activity.

The Devonian, Carboniferous, and Permian periods. The paleogeography of the Late Devonian, Carboniferous, and Permian periods is somewhat similar. In place of the present continent of Asia there existed, as before, the separate, small continental blocks of the three northern platforms, with the enormous continent of Gondwanaland to the south of them. Carbonaceous, detrital, and volcanic rocks continued to be deposited in the seas of the geosynclines separating these blocks. Calcareous sediments were also characteristic of the seas covering the massifs of Iran and Turkey, part of the Siberian Platform during the Devonian Period, and the southern half of the Chinese Platform in the Carboniferous Period. But the area of the marine geosynclinal basins was greatly reduced as a result of subsequent Hercynian folding. The basins between the Russian, Siberian, and Chinese platforms almost disappeared toward the middle of the Permian Period, at which time the area of present-day Mongolia and Western Siberia dried up. Coal-bearing strata

Virtual disappearance of the landlocked basins



Geological structure of Asia and (inset) New Guinea.

that accumulated during the Late Carboniferous and Early Permian periods on the Indian, Chinese, and Siberian platforms and in the intermontane Kuznetsk, Minusinsk, and Tuva basins provide evidence that a humid climate then prevailed. In India these strata are underlain by glacial deposits of the Carboniferous Period.

The Triassic and Jurassic periods. From the beginning of the Triassic Period, the northern platforms, together with the folded mountain ranges from the Paleozoic Era,

formed a large continent embracing three-quarters of modern Eurasia. At the same time Gondwanaland split up into several giant blocks, separated from each other by a series of basins, some of which broadened and formed the Indian Ocean. The geosynclinal basins of the Mesozoic Era survived in the Tethys Sea, in Northeastern Siberia, in the Amur region, in Indochina, and in the belt of island arcs located in East and Southeast Asia. Shallow seas or coal-bearing freshwater basins and swamps, especially numerous during the Jurassic Period, at times covered the lowlands of Western Siberia, the Turansk Plain, the Iranian Massif, and the Ordos Desert and Szechwan Basin.

The Cretaceous Period and the Cenozoic Era. Toward the middle of the Cretaceous Period, the areas of Mesozoic folding—located in what are now the Verkhoyanski Mountains, Sikhote-Alin, and Indochina—dried up, while in the course of the Late Cretaceous Period and the Early Cenozoic Era a rapid reduction in the size of the Tethys Sea took place. Finally, toward the beginning of the Early Cenozoic Era, folding in the Himalayas united India with the remainder of the continent, and Asia acquired approximately the outlines it has today. The island arcs and the basins of the marginal seas were shaped at the same time.

The joining of India and Asia

The Cretaceous Period and the Cenozoic Era were marked by mighty volcanic phenomena throughout East Asia. The Pliocene Epoch and Quaternary Period were times of vigorous tectonic movements that not only led to the uplifting of folded mountain ranges in the Alpine–Himalayan belt and in the island arcs but also rejuvenated the relief of the ancient folded mountains of the Urals, the Tien Shan, the Altai, the Sayans, and other ranges. Today an active belt of contemporary seismic activity stretches from the Hindu Kush through the Tien Shan, Mongolia, and the Baikal region to the Sea of Okhotsk.

Contemporary developments. The paleoclimatic and paleomagnetic data indicate that considerable shifting was taking place—in relation to the North and South poles as well as to each other—of those blocks from which the Eurasian continent was gradually formed. During the course of the early and middle periods of the Paleozoic Era, a considerable rapprochement (bringing together) of the Russian and Siberian platforms occurred. In the Late Carboniferous and Permian periods the Indian Platform was situated much closer to the South Pole and was partially subjected to glaciation. The Siberian Platform, on the other hand, was at that time located in the northern humid zone; in the Triassic Period the North Pole evidently lay at its northeastern edge. At the end of the Paleozoic Era the Tethys Sea was several times broader than the belt of folding that was formed within it in the course of the Mesozoic and Cenozoic eras. There are indications of the drift of the island arcs toward the shores of the Pacific Ocean and of the spreading of the bottoms of the marginal seas, such as the Sea of Japan and others, which originally developed and deepened in the middle of the Mesozoic Era. (P.N.K.)

II. Physical geography

RELIEF

Asia is the highest of the continents and contains the sharpest relief. The highest peak in the world, Mt. Everest, which is 29,028 feet (8,848 metres) high; the lowest place on the Earth's land surface, the Dead Sea, which is 1,296 feet (395 metres) below sea level; and the world's deepest continental trough, occupied by Lake Baikal, which is 5,315 feet (1,620 metres) deep and whose bottom lies at 4,250 feet (1,295 metres) below sea level, are all located in Asia.

Asia is also the most extensive of the continents. The farthest terminal points of the Asian mainland are Cape Chelyuskin in the Soviet Union (77°43′ N) to the north; the tip of the Malay Peninsula, Cape Piai, or Bulus (01°16′ N), to the south; Cape Baba in Turkey (26°04′ E) to the west; and Cape Dezhnyova, or East Cape (169°40′ W), also in the Soviet Union, overlooking the Bering Strait, to the east. The shores of Asia are washed by the Arctic Ocean on the north, the Pacific Ocean on the east, the Indian Ocean and the marginal seas of the Indian and Pacific oceans on the south, and by the seas of the Atlantic Ocean—the Mediterranean, the Aegean, the Sea of Marmara, the Black Sea, and the Azov Sea—as well as by the landlocked Caspian Sea, on the west.

The continent's extremities

Asia is separated from Australia to the southeast by the mingled waters of the Indian and Pacific oceans, and from North America on the northeast by the Bering Strait. The Isthmus of Suez unites Asia with Africa, and it is generally agreed that the Suez Canal forms the border between them.

The boundary between Asia and Europe is a historical–cultural concept that has changed more than once and is only as a matter of agreement tied to a specific borderline. The most convenient geographic boundary is a line drawn along the eastern base of the Urals, then turning west along the Emba River to the Caspian Sea; west of the Caspian, the boundary follows the Manych River and the Kerch Strait to the Black Sea. From a statistical–economic point of view, the boundary is taken to run along those political–administrative borders of the republics and *oblasti* of the Soviet Union that most closely approximate this line, which is to say, along the eastern borders of the Komi Autonomous Soviet Socialist Republic; the *oblasti* of Arkhangelsk, Sverdlovsk, and Chelyabinsk; the western border of the Kazakh S.S.R.; and along the northern borders of the *kraya* (territories) of Stavropol and Krasnodar. Some authorities consider, however, that the boundary runs along the border between the Russian Soviet Federated Socialist Republic and the Transcaucasian republics (the Georgian S.S.R. and the Azerbaijan S.S.R.).

The area of mainland Asia, including the Caucasian isthmus, amounts to about 16,750,000 square miles (43,400,000 square kilometres), of which the peninsulas—Asia Minor to the west; the Arabian Peninsula, peninsular India, Indochina, and the Malay Peninsula to the south; Korea, Kamchatka, and the Chukotsk Peninsula to the east; Taymyr and Yamal to the north—make up about 3,000,000 square miles.

The islands—Cyprus, Sri Lanka, the Andamans, the Malay Archipelago, the Philippines, Hainan, Taiwan, the Ryukyus, Japan, the Kurils, Sakhalin, Wrangel Island, the New Siberian Islands, and Severnaya Zemlya—account for another 770,000 square miles.

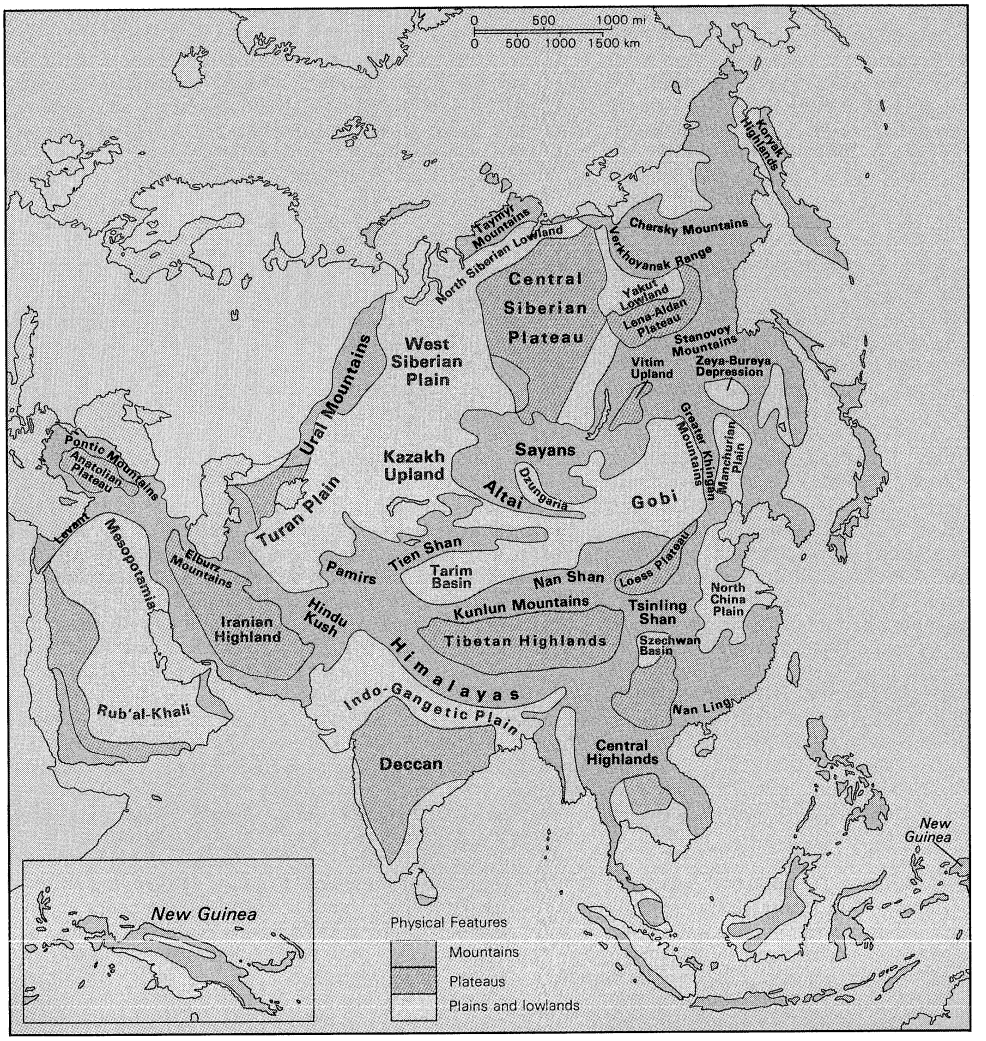
Asia's coastline is, variously, high and mountainous; low and alluvial; terraced as the result of the land being uplifted; or "drowned," where the land has subsided. The specific features of the coastline in some areas—especially in the east and southeast—are the result of active volcanism; of thermal abrasion (resulting from a combination of action by sea breakers and of thawing) by the subterranean fossilized ice (consisting of fossil ice, subsurface ice, and ice-formed rock), as in northeastern Siberia; and coral building, as in the south-southeastern area.

A characteristic of the surface of Asia is the predominance of mountains and plateaus, which form about three-quarters of the total area. The highest mountains and plateaus occur in Central Asia (Mongolia, Dzungaria, the Kashgar region, and Tibet) and Middle Asia (Turkmenia, Uzbekistan, Tadzhikistan, Kirgiziya, and Kazakhstan), which are also characterized by the vastness of their interior drainage basins.

Asia's mountain ranges

The mountain belts. The mountains are grouped in two large belts. One extends from the Chukchi Peninsula at the eastern extremity of Asia through the Kolyma Highlands, the Dzhugdzhur Mountains, and the Stanovoy Mountains to the mountains of Southern Siberia (the Stanovoy Uplands, the Sayans, and the Altai) and to the Tien Shan and Gissar-Alai. The Chersk and Verkhoyansk ranges are the Western spurs of this belt.

The second belt, in the west, runs in a latitudinal direction and includes the West Asian highlands, the Pamirs, Karakoram, the Tibetan Highlands, and the Himalayas; it then turns to the south and southeast, running through the Arakan Mountains to the islands of the Malay Ar-



Physiographic regions of Asia.

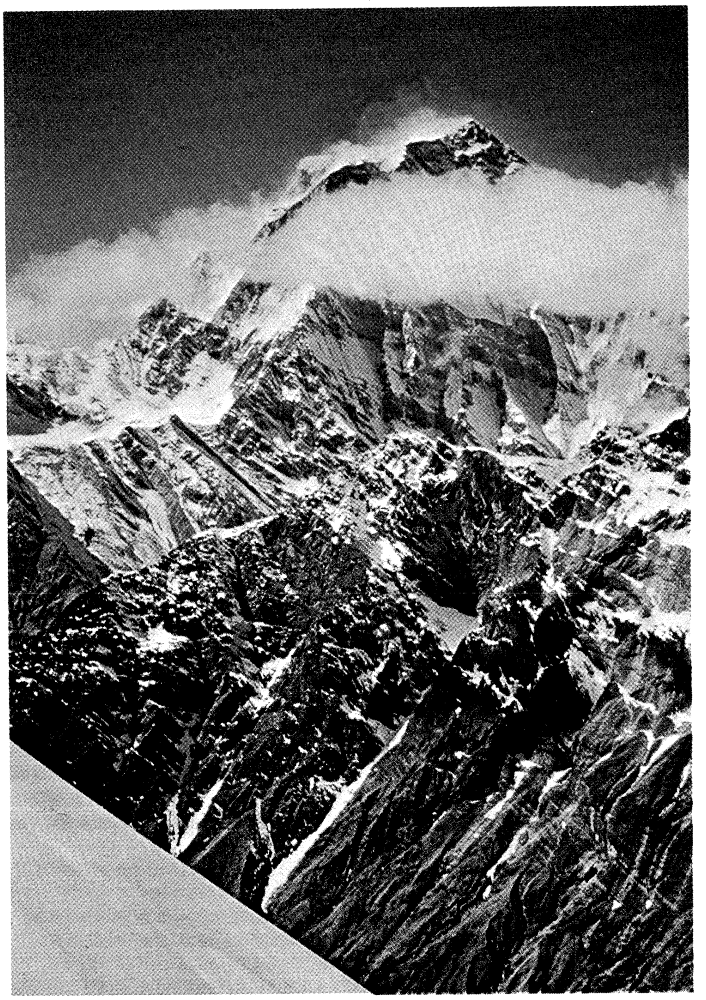
chipelago. A generally latitudinal branch springs from it in the Pamirs region and runs eastward through the Kunlun, Nan Shan, and Tsinling mountains. The western part of the second mountain belt consists, for a considerable distance, of two series of mountain chains that converge in dense knots in the highlands of Armenia, in the Pamirs, and in the southeast of the Tibetan Highlands; the two chains then diverge to encompass the interior plateaus—the highlands of Asia Minor together with the Anatolian Plateau, and the Iranian and Tibetan highlands. At the margins of the highlands the mountain chains change direction abruptly in the congested mountain knots, but in the intervening areas they curve in flowing arcs.

Along the edges of the Central Asian plateaus extend the elongated mountain chains of the Greater Khingan, T'ai-hang Shan, and the Sino-Tibetan ranges. The Khingan–Bureya mountains (Bureya and Lesser Khingan mountains) demarcate the Zeya-Bureya Depression; the Manchurian–Korean mountains and the Sikhote-Alin ranges separate the plains of the Amur–Sungari, the Lake Khanka lowland, and the Manchurian Plain. The coastal ranges in the southeast consist of the Nan Ling (South China) and Annam mountains. In the east, the Koryak Highlands rise in the northern section of the Kamchatka–Koryak arc; in its southern portion they form the Central Range (Sredinny Khrebet) of Kamchatka. The marginal seas of the Pacific Ocean are bordered by the East Asian islands, which form the line of arcs that also extends partly onto the continent, running from Borneo to Taiwan, through the Ryukyu Islands to Korea, from Honshu to Sakhalin, and from the Kuril Islands to Kamchatka.

The distribution of plains and lowlands

The plains and lowlands. Low plains occupy approximately a quarter of Asia, particularly the vast West Siberian and Turan plains of the interior. The remaining lowlands are distributed either in the maritime regions, such as Northern Siberia, the Yana-Indigirka lowland, Kolyma, and the Chinese coastal mainland, or in the piedmont depressions of Mesopotamia, the Indo-Gangetic Plain, and mainland Southeast Asia. In addition, there are the intermontane plains of Kashgaria, Dzungaria, Tsaidam, and Fergana and the plateaus of middle Siberia and the Gobi. The plateaus inside Tibet, the Tien Shan, and the Pamirs lie at altitudes as high as 12,000 feet. Many ranges in Middle and Central Asia reach elevations of 21,000 to 24,000 feet, but in other mountain chains and massifs they rarely exceed 12,000 feet. To the south of the zone of piedmont depressions, partially occupied by seas—the Arabian Sea, the Bay of Bengal, the Gulf of Oman, the Persian Gulf, and, in the Mediterranean, the Cyprus Basin—lie extensive tablelands and plateaus, including the Deccan Plateau in India and the Syrian–Arabian Plateau in the west; these are enclosed by marginal mountain ranges, such as the Western Ghāts in India; the Oman, Ḥaḍramawt, Yemen, and Hejaz mountains on the Arabian Peninsula; and the Lebanon and Anti-Lebanon mountains in the Levant. In Central Siberia are the isolated and uplifted Putorana Mountains and, situated to the north of them, the Byrrangas Mountains.

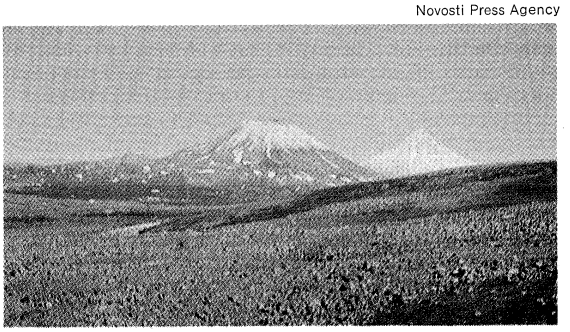
The islands. A large part of the islands of Asia are mountainous. The highlands of Sri Lanka rise to 8,279 feet (2,524 metres), Mt. Kinabalu in Malaysia reaches 13,451 feet (4,101 metres), Fujiyama on the Japanese island of Honshu has an altitude of 12,385 feet (3,776



Dhaulāgiri, in the Great Himalayas, Nepal, rising to a height of 26,810 feet.
Bernhaut—FPG

metres), and many volcanoes of Sumatra, Java, and Mindanao reach 10,000 feet. The Kuril–Kamchatka island arc, which extends onto the Kamchatka Peninsula, comprises the Vostochny (East) volcanic range. Especially high is the Klyuchevsk group of volcanoes, where the highest active volcano in Asia—Klyuchevskaya Sopka—rises 15,580 feet (4,750 metres).

Novosti Press Agency



Sopka (volcano) Krasheninnikova on Kamchatka Peninsula in Northeastern Siberia.

Forces shaping Asia's relief

Geologic and climatic influences. Mesozoic and Alpine foldings created boundaries between basic types of mountains over vast areas of Asia. The contemporary relief of Asia was molded primarily under the influences of: (1) ancient processes of planation (levelling); (2) larger vertical movements of the surface during the later Tertiary and Quaternary periods; and (3) severe erosive dissection of the edges of the uplifted highlands with the accompanying accumulation of alluvium in low-lying troughs, which were either settling downward or being uplifted more slowly than the adjoining heights.

The interior parts of the uplifted highlands, and the plateaus and tablelands of peninsular India, Arabia, Syria, and Eastern Siberia, which are relatively low-lying but composed of resistant rock, have largely preserved their ancient peneplaned (levelled) surfaces. Particularly spectacular uplifting occurred in Central Asia, where for the last 30,000,000 years the amplitude of this uplift of the mountain ranges of Tibet, the Pamirs, and the Himalayas has exceeded 13,000 feet. The eastern margin, meanwhile, underwent subsidences of up to 2,300 feet. Uplifting as a result of fractures at great depths, of which the Kopet-Dag and Ferghan mountains provide examples, and of folding over a large radius, of which examples may be seen in the Tien Shan and Gissar-Alai, played a large role.

Erosional dissection transformed many ancient plateaus into mountainous regions. Majestic gorges were carved into the highlands of the western Pamirs and southeastern Tibet; the Himalayas, the Kunluns, the Sayans, the Stanovoye Highlands, the Cherski Mountains, and the marginal ranges of the West Asian highlands were deeply cut by the rivers, creating deep superimposed gorges and canyons. In many areas, and especially in those regions with dry climates, erosion clearly exposed the structural forms, including rock layers of different erosional resistance.

Eolian relief

Vast areas of Middle, Central, and East Asia, particularly in the Huang Ho Basin, are covered with loess (a loamy unstratified deposit formed by the wind or by glacial meltwater deposition). There are broad expanses of badlands, eolian (wind-produced) relief, and karst topography (limestone terrain associated with vertical and underground drainage), and features associated with ancient glaciation.

The mantle of Quaternary glaciation embraced northwestern Asia only to 60° N. East of the Khatanga River, which flows from Siberia into the Arctic Ocean, only isolated glaciation of the mantle debris and of the mountains occurred because of the extremely dry climate that existed in the northeast even at that time. The high mountain regions experienced mainly mountain glaciation. There are traces of several periods during which the glaciers advanced—periods separated by warmer interglacial epochs. Glaciation continues in many of the mountainous areas and on the Severnaya Zemlya archipelago. Karakoram, the Pamirs, the Tien Shan, the Himalayas, and the eastern Hindu Kush are noted for the immensity of their contemporary glaciers.

There is an enormous area of permafrost in northern Asia that extends to lower latitudes than in any other part of the world. Little snowfall occurs, due to the aridity, and deep freezing of the soil takes place.

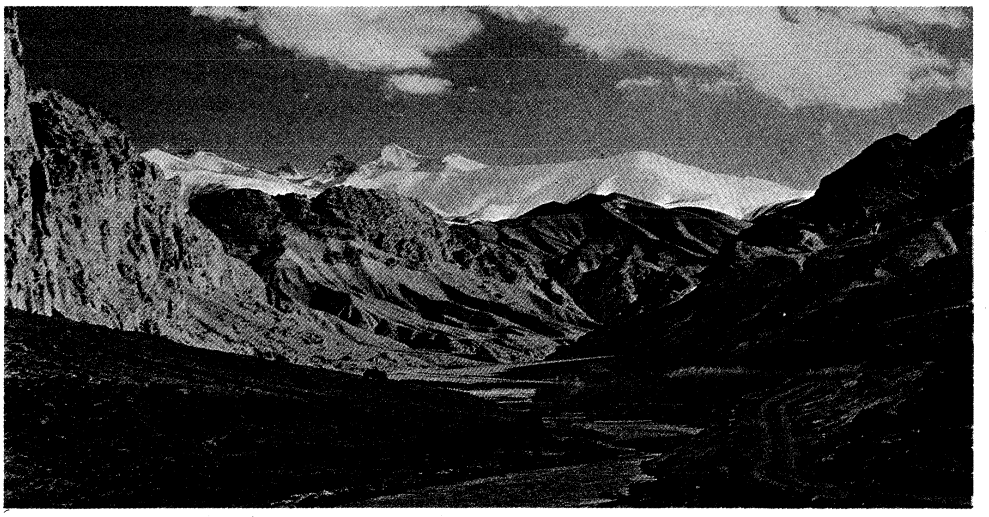
Several lowlands, primarily coastal plains, are covered with marine sediments as the result of recent advances of seas, such as the Caspian and the northern seas.

Volcanic zones

Volcanism added broad lava plateaus and chains of young volcanic cones to the relief of Asia. Ancient lavas and intrusions of magma, exposed by later erosion, cover the terraced plateaus of peninsular India and Central Siberia. Extensive zones of young volcanic relief and contemporary volcanism, however, are confined to the unstable arcs of the East Asian islands, together with Kamchatka, the Philippines, and the Greater and Lesser Sunda Islands.

Recent volcanism is also characteristic of the West Asian highlands, the Caucasus, Mongolia, the Manchurian–Korean mountains, and the Syrian-Arabian Plateau. In historic times eruptions have also occurred in the interior of the continent in the Lesser Khingan Mountains and the Anyuy highlands.

The regions of Asia. In geographical literature the practice of dividing Asia into large regions, each grouping together a number of countries, is common. These divisions usually consist of North Asia, including Siberia and the northeastern edges of the continent; East Asia, including the continental part of the southern Soviet Far East, the East Asian islands, Korea, and eastern and



Bololo Canyon, north of Kābul, Afghanistan. The Hindu Kush mountains are visible in the background.
Harrison Forman

northeastern China; Central Asia, including the Tibetan Highlands, Dzungaria and Kashgaria in the Sinkiang Uighur Autonomous Region, Inner Mongolia, the Gobi, and the Sino-Tibetan ranges; Middle Asia, including the Turanian Plain, the Pamirs, the Gissar-Alai and the Tien Shan; South Asia, including the Philippine and the Malay archipelagoes, Indochina and the Indian Peninsula, the Indo-Gangetic Plain, and the Himalayas; and West Asia, including the West Asian highlands (Asia Minor, Armenia, and Iran), the Levant, and the Arabian Peninsula. On occasion, the Philippines, the Malay Archipelago, and the Indochina peninsula, instead of being considered as part of South Asia, are grouped separately as Southeast Asia; the Arabian Peninsula and the Levant are also sometimes grouped together separately as Southwest Asia.

North Asia. The North Asia region includes platform plains, plateaus, and folded mountain ranges. Frost weathering and permafrost have influenced relief.

Northeast Siberia

In Northeast Siberia are found faulted and folded mountains of moderate height, such as the Verkhoyansk, Chersk, and Okhotsk-Chaun mountain arcs, formed of Mesozoic structures rejuvenated by neo-tectonic uplifting; the Koryak Mountains, formed of Cenozoic structures, are also in this region. Volcanic activity took place in these areas during the Cenozoic Era. Some plateaus are found in the areas of the ancient massifs, such as the Kolyma massif. Traces of several former centres of mountain glaciers remain, as well as traces of lowland originally covered by the sea, such as the New Siberian Islands. The Aldan Plateau—an ancient peneplain resting on the underlying platform that sometimes outcrops on the surface as the Aldan Shield—is located in the region. Traces of ancient glaciation are also to be distinguished.

The North Siberian plains consist of the Middle Siberian Tableland and the Lena–Vilyuy lowlands, which are platform plateaus and stratified plains that were uplifted in the Cenozoic Era. They are composed of terraced and dissected mesas with exposed horizontal volcanic intrusions; plains formed from uplifted Precambrian blocks; a young uplifted mesa, dissected at the edges and partly covered with traprock (Putorana Mountains); and the peripheral North Siberian lowland, covered with its original marine deposits.

The West Siberian Plain is stratified and is composed of Early Cenozoic sediments deposited over thicknesses of Mesozoic material, in addition to folded bedrock that is Hercynian in the west and Caledonian in the east. The northern part was earlier subjected to several periods of glaciation; in the south the predominant deposits are those laid down by glacier streams, as well as alluvial deposits.

In the northern part of the region are the mountains and islands of the Asian Arctic. The archipelago of Severnaya Zemlya is formed of fragments of fractured Paleozoic folded structures. Throughout the region vigorous contemporary glaciation has occurred.

East Asia. Mountains and plains are characteristic of the northern part of continental East Asia. The main features in the northern region include the Khingan–Burein mountains; the Sikhote-Alin ranges of Khabarovsk and Primorsky *kraya* (territories) in the Soviet Union; the Manchurian–Korean highlands running along North Korea's border with China; the East Korean range of the Korean Peninsula; the Zeya-Buryea Depression of Amur *oblast* in the Soviet Union; the Liao Ho in Liaoning Province, China; the Manchurian Plain and the North China Plain; and the Amur and Sungari rivers and the Lake Khanka lowlands. Most of these features were formed by folding, faulting, or broad zonal subsidence. The mountains are separated by alluvial lowlands in areas where recent subsidence has occurred.

The mountains of southeastern China were formed from Precambrian and Caledonian remnants of the Chinese Platform by folding and faulting that occurred during the Mesozoic and Cenozoic eras. The mountain ranges are numerous, are of low or moderate altitude, and occupy most of the surface area, leaving only small, irregular-shaped plains.

The East Asian islands

The islands off the coast of East Asia and the Kamchatka Peninsula are related formations. The Ryukyu Islands, Japan, Sakhalin, and the Kuril Islands are fragments, uplifted in varying degrees, of the Ryukyu–Korean, Honshu–Sakhalin, and Kuril–Kamchatka mountain-island arcs. Dating from the Mesozoic and Cenozoic eras, these arcs have complex knots at their junctions, represented by the topography of Kyushu and Hokkaido. The mountains are of low or moderate height and are formed of folded and faulted blocks; some volcanic mountains and small alluvial lowlands are also to be found.

Kamchatka is a mountainous peninsula, formed from fragments of the Kamchatka–Koryakskaya and Kurilo–Kamchatskaya arcs, which occur in parallel ranges. The young folds enclose rigid ancient structures. Cenozoic (including contemporary) volcanism is pronounced. Vast plains exist that are composed of alluvia with volcanic ashes.

Central Asia and South Siberia. Central Asia consists of mountains, plateaus, and tablelands formed from fragments of the Siberian and Chinese platforms, peripherally surrounded by a folded area formed in the Paleozoic and Mesozoic eras.

The mountains of Southern Siberia and Mongolia were formed by renewed uplift of old faulted and folded blocks; ranges are separated by intermontane troughs. The alpine mountains—the Altai, the Mongolian Altai,

and the Sayano–Tuvan and Stanovoy highlands—are particularly noticeable. They have clearly defined features resulting from ancient glaciation; contemporary glaciation is also very active.

The Central Asian plains and tablelands include the Takla Makan Desert, the Gobi, and the Ordos Desert. Relief features vary from surfaces levelled by erosion in the Mesozoic and Cenozoic eras to stratified plateaus with low mountains, eroded plateaus on which loess had accumulated, and vast sandy deserts covered with windborne alluvium and lacustrine deposits.

Alpine Asia—sometimes known as High Asia—includes the Pamirs and the eastern Hindu Kush, the Kunlun Mountains, the Tien Shan, the Gissar-Alai Mountains, the Tibetan Highlands, the Karakoram Range, and the Himalayas.

The Pamirs and the eastern Hindu Kush are sharply uplifted mountains dissected into ridges and gorges in the west. There is thick glacial cover; alpine deserts occur on the plateaus.

The Kunlun Mountains, the Tien Shan, and the Gissar-Alai Mountains belong to an alpine region that was formed from folded structures of Paleozoic age. There are glaciers that are of impressive size centred in this alpine region.

The Tibetan Highlands

The Tibetan Highlands represent a fractured alpine zone in which Mesozoic and Cenozoic structures that surround an older mass in the centre have experienced more recent uplifting. Some of the highlands are covered with detrital desert; elsewhere in this region, alpine highlands are dissected by erosion or are covered with glaciers.

The Karakoram Range and the Himalayas include the highest mountains in the world; they were formed by uplifting that took place in a zone of Cenozoic folds and Mesozoic partially folded areas containing outcrops of the ancient bedrock. Contemporary glaciation is vigorous.

South Asia. South Asia, in the limited sense of the term, consists of peninsular India and Sri Lanka (formerly Ceylon) and the Indo-Gangetic Plain.

Peninsular India and Sri Lanka are formed of platform plateaus and tablelands uplifted in the Mesozoic and Cenozoic eras and subjected to humid climate erosion ever since. Tablelands with uplifted margins and terraced and dissected plateaus with lava mantles or intrusions may be distinguished.

The Indo-Gangetic Plain is formed from the combined alluvial plains of the Indus, Ganges, and Brahmaputra rivers, which lie in a deep marginal depression running north of and parallel to the main range of the Himalayas. It is an area of immature subsidence, in which thick accumulations of earlier marine sediments and later continental deposits that washed down from the mountains have been transformed into sandy deserts in the western arid region.

Southeast Asia. Southeast Asia comprises the Indochina Peninsula and the islands and peninsulas to the southeast of the Asian continent.

Indochina

The mainland consists of the western mountain area and the central and eastern mountains and plains. The western mountain area of Burma is a zone of Cenozoic folding. Mountains of medium altitude are formed of folded blocks that decrease in size and altitude to the south; the valleys are alluvial and broaden out to the south. The central and eastern region of Thailand and North and South Vietnam is characterized by mountains of low and moderate height that have been moderately fractured. The region is one of Mesozoic structures surrounding an ancient mass known as the Cambodian saucer, with which are associated plateaus and lowlands filled with accumulated alluvial deposits.

Archipelagoes border the southeastern margin of Asia, consisting mainly of island arcs with which peninsulas are associated. The island arcs are bordered by very deep oceanic trenches. These arcs are characteristically very unstable and are volcanically active.

The Indian Ocean arcs—the island chains of Sumatra, Java, and the Lesser Sunda Islands—consist of fragments of alpine folds formed from materials of different ages. Cenozoic and contemporary volcanic activity is manifest, and volcanic mountains as well as alluvial lowlands may be distinguished.

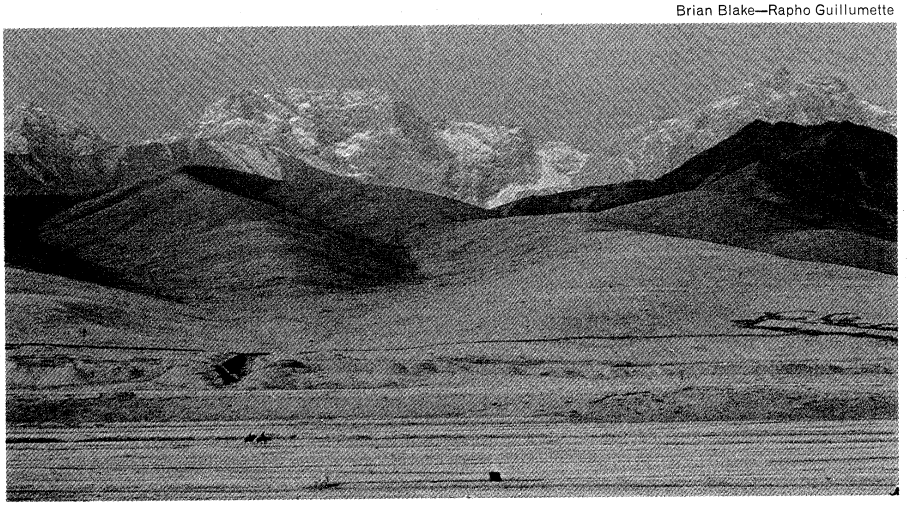
Borneo and the Malay Peninsula are formed from fractured continental land situated at the junction of the Alpine–Himalayan and East Asiatic geosynclinal regions; contemporary volcanism is absent. The mountains are composed of folded and faulted blocks; the lowlands are alluvial.

The Pacific Ocean island arcs, including Celebes, the Moluccas, the Philippine Islands, and Taiwan, are fragments of folded alpine structures that were built up by volcanic products during the Cenozoic Era. Volcanic activity and the building of coral reefs continue. Mountain areas of moderate height, volcanic ranges, alluvial lowlands, and coral reef islets may all be distinguished.

Middle Asia. Middle Asia includes the plains and hills lying between the Caspian Sea to the west and Lake Balkhash to the east.

The area between the Caspian and Lake Balkhash is composed of flat plains on continental platforms of folded Paleozoic and Mesozoic bedrock. Individual uplifted portions form low rounded hills (*melkosopochniks*) in the Kazakh region; low mountains on the Mangyshlak and Krasnovodsk peninsulas of the Caspian Sea; and mesas (isolated hills with level summits and steeply sloping sides) in areas of earlier marine sedimentation, such as the Ustyurt Plateau and the Kara-Kum Desert. Thick accumulations of alluvium have been transported by the wind, forming sandy deserts in the south. Original marine and lacustrine sediments adjoin the shores of the Caspian and Aral seas and Lake Balkhash.

Brian Blake—Rapho Guillumette



The Ch'iang-t'ang plain of northern Tibet, bordered by the Kunlun Mountains.

West Asia. West Asia includes the highlands of Asia Minor, and the Armenian and Iranian highlands.

The highlands of Asia Minor—the Pontic mountain system that parallels the Black Sea, and the Tavr and Anatolian tablelands—are areas of severe fragmentation, heightened erosional dissection, and isolated occurrences of volcanism.

The Armenian Highlands

The Armenian Highlands, which include the Little Caucasus and the Kurd mountains, are severely fragmented. Recent uplifting, in the form of a knot of mountain arcs, took place during a period of vigorous volcanism that occurred in the Cenozoic Era.

The Iranian Highlands represent a combination of mountain arcs (in the north, the Elburz and Turkmen-Khorasan mountains, the Safīd Kūh Selseleh-ye, and the western Hindu Kush; in the south, the Zagros, Makran, Soleymān, and Kirthar mountains), together with the tablelands of the interior, and the Central Iranian, Eastern Iranian, and Central Afghanistan mountains. There are isolated Cenozoic volcanoes, a predominance of accumulated remnants resulting from ancient erosion, and saline and sandy deserts in the depressions and on the tablelands.

Southwest Asia. Southwest Asia, like much of southern Asia, is made up of an ancient platform—the northern fragments of Gondwanaland—in which sloping plains occur in the marginal downwarps. Its principal components are the Arabian Peninsula and Mesopotamia.

The Arabian Peninsula is a tilted platform, highest along the Red Sea, on which the stratified plains have undergone erosion under arid conditions. Block tablelands with uplifted margins, Cenozoic lava plateaus, stratified plains, and cuestas (long, low ridges with a steep face on one side and a long gentle slope on the other) may all be distinguished. Ancient marine sands and alluvia, resulting from previous subsidence and sedimentation, now take the form of sandy deserts.

Mesopotamia consists of the Tigris and Euphrates floodplains and of the deltas from Baghdad to the Persian Gulf. The original lowland is covered with late Cenozoic and Quaternary sedimentation; the elevated plain, on the other hand, has been dissected by erosion and denudation under the continental conditions prevailing in the Late Cenozoic Era.

CLIMATE

Air masses and wind patterns. The enormous expanse of Asia and the abundance of mountain barriers and inland depressions have resulted in great differences in existing conditions of solar radiation, atmospheric circulation, and climate as a whole. A continental climate, associated with large landmasses and characterized by an extreme annual range of temperature, prevails over a large part of Asia. Air reaching Asia from the Atlantic Ocean, after passing over Europe or Africa, has had time to be transformed into continental air. As a result of the prevalent easterly movement of the air masses, as well as the isolating effect of the marginal mountain ranges, the influence of sea air from the Pacific Ocean extends only to the eastern edge of Asia. From the north, Arctic air has unimpeded access into the continent. In the south, tropical and equatorial air masses predominate, but their penetration to the centre of Asia is restricted by the ridges of the latitudinal belt of highlands; in the winter months—November through March—such penetration is further impeded by the density of the cold air masses over the interior.

The continental climate

The contrast between the strong heating of the landmasses in the summer months from May to September and the chilling in winter produces sharp seasonal variations in the atmospheric circulation and also enhances the role of local centres of atmospheric activity. Winter chilling of the Asian landmass develops a persistent high-pressure winter anticyclone over Siberia, Mongolia, and Tibet, which is normally centred southwest of Lake Baikal. Within the zone of the anticyclone there is relatively little strong air movement in protected basins and lowlands, but strong winds may affect the higher mountains and passes. The anticyclone is fed by subsiding upper air, by bursts of Arctic air flowing in from the north, and by the persistent westerly air drift that accompanies the gusty cyclonic low-pressure cells operating within the Northern Hemisphere cyclonic storm system. Drifts of cold, dry air move eastward and southward out of the continent, affecting eastern and southern Asia during the winter. Only a few of the winter cyclonic lows moving eastward out of Europe carry clear across Asia, but they do bring greater periodic change in weather in Western Siberia than is typical in Central Siberia. The zone of lowest temperature—the so-called cold pole—is found in the northeast, near Verhoyansk, where temperatures as low as −90° F (−68° C) are recorded. The outward drift of winter air creates a sharp temperature anomaly on Asia's eastern margin, where the climate is colder than the characteristic average for each given latitude. But episodic intrusions of oceanic air from the east and southeast moderate this anomaly, so that temperatures are not as severe here as in the centre of the anticyclone.

The polar front

The zone where the temperate and tropical air masses are in contact—called the polar front—shifts southward in winter. This movement is caused by a displacement, in the same direction, of the entire system of atmospheric circulation—a displacement resulting from the powerful climatic influence exerted by the chilled continent. The winter rainy season in the southern parts of the West Asian highlands, which is characteristic of the Mediterranean climate, is associated with this southerly movement of the polar front. In the more northerly areas of West and Middle Asia, the effect of cyclonic action is particularly strong in the spring, causing the maximum in annual precipitation to occur at this season. In summer, the polar front shifts northward, causing cyclonic rains in the mountains of Southern Siberia. In West, Middle, and Central Asia, a hot, dry, dusty, continental tropical wind blows at this time. Over the basin of the Indus River the heating creates a low-pressure area, known as the South Asian (or Iranian) low. The southern monsoon (a rain-bearing wind) advances along its southern edge, bringing copious rainfall to peninsular India, the southern Himalayas, and mainland Southeast Asia. Farther to the west the hot, dry air of North Africa and the katabatic (downward) current of air from Europe, blowing from the northwest, sweep in the direction of this low-pressure area. The aridity of the desert–tropical climate of Arabia and Pakistan is related to this phenomenon.

In eastern Asia the Pacific Ocean polar front creates atmospheric disturbances during the summer. From the warm sectors of cyclones moving westward through this region, the warm and moist summer monsoon blows toward the continent. Becoming chilled as it passes over cold ocean currents, this air brings fogs and drizzling rains. To the south of 38° N, where the warm Kuroshio (Japan) Current approaches the coast of Japan, the summer monsoon brings protracted rains and high humidity; together with high temperatures, this creates a hothouse atmosphere.

The summer period over China is a time of variable air movement out of the South Pacific. If that drift is strong and the summer continental low-pressure zone is marked, a strong summer monsoon may carry moisture well into Mongolia. If neither the drift nor the continental low is strong, the China summer monsoon may fail, falter over eastern China, or cause irregular weather patterns that may threaten China proper with crop failure.

Typhoons and monsoons

Tropical cyclones, or typhoons, occur along the East Asian weather fronts throughout the year but are most severe during the autumn months. These typhoons are accompanied by very strong winds and torrential rains so heavy that the maximum precipitation from the typhoons locally may exceed the total amounts received during the normal summer monsoons.

In winter the Pacific Ocean polar front is driven back to tropical latitudes by a steady drift of cold, dry Siberian air. On the East Asian islands the effect of the winter continental monsoon is tempered by the surrounding seas. In passing over them, it becomes warmed and saturated with moisture; then waters the northwestern slopes

of the island arcs. Occasionally, however, strong bursts of cold air carry cold spells as far south as Hong Kong and Manila.

In winter, continental tropical air prevails in subequatorial Asia; in summer it is replaced by equatorial ocean air. The winter season's dry and warm winds, directed toward the equatorial low-pressure axis, are analogous to trade winds but simultaneously act as the South Asian continental monsoon. The dry spring that follows changes abruptly and dramatically into the rainy summer with the onset of the monsoon. The summer monsoon brings enormous amounts of rain (up to about 25 inches in a month). Over the areas of Asia close to the Equator —southern Sri Lanka, Malaysia, and the Greater Sunda Islands—equatorial air prevails continuously, accompanied by even temperatures and abundant rainfall at all seasons. The Lesser Sunda Islands have a subequatorial monsoon climate; their wet and dry seasons are regulated by the calendar rhythm of the Southern Hemisphere, which is characterized by a wet summer from November to February and a dry winter from June to October.

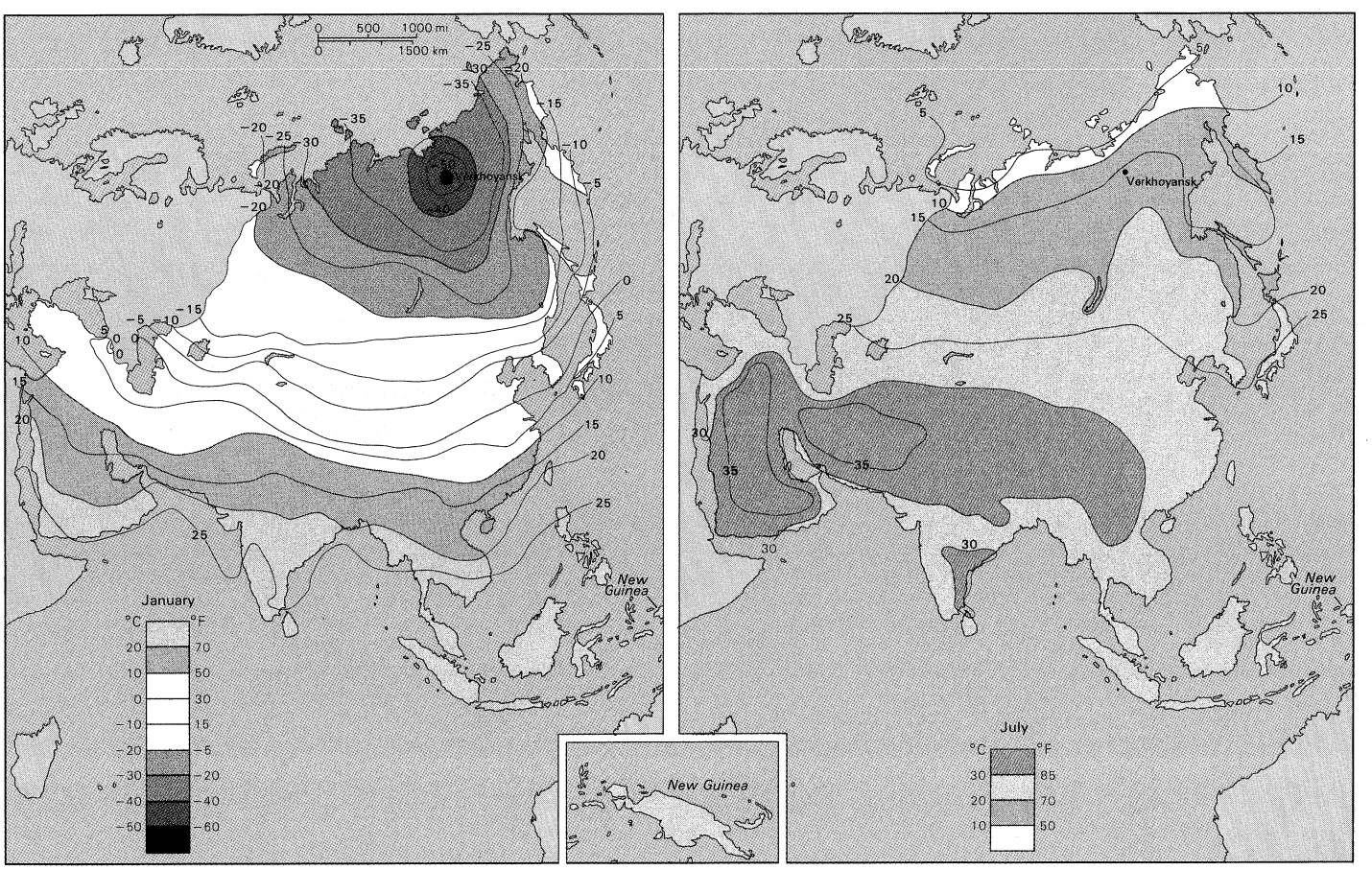
The influence of topography. Differences between the climatic conditions of the various regions of Asia are determined to a considerable degree by topography. Different altitudinal climatic zones are most clearly defined on the southern slopes of the Himalayas, where they vary from the subequatorial and tropical climates of the foothills, at the lowest levels, to the snowy climate of the peaks, at the highest altitudes. The degree of exposure also plays a large role—the different orientation of the opposite slopes of the ridges in relation to compass directions and to the prevailing winds. The sunny southern slopes differ from the shady northern ones, and windward slopes exposed to moist ocean winds differ from leeward slopes, which, lying in the wind (and rain) shadow, are necessarily drier. In addition to the physical isolation of the leeward slopes from the moisture-laden winds, the foehn effect is also found here. This occurs when a strong wind traverses a mountain range and is deflected downward as a warm, dry, gusty, erratic wind. Contrasts of climate resulting from exposure are manifested clearly in the Himalayas, the Elburz Mountains, Japan, Taiwan, the Philippines, the Tien Shan, the Transbaikalia, and many other places.

The foehn effect

The isolating barrier effect of the relief on the climate appears clearly in the West Asian highlands and in Central Asia. In these regions the surrounding mountains isolate the tablelands of the interior from the moisture-laden winds. The massiveness of the interior highlands is also a significant factor; it favours the formation of local anticyclones over them during the cold months of the year.

During the winter season some of the cyclonic storms that move eastward through the Mediterranean Basin are deflected south of the Tibetan Highlands, crossing northern India and southwestern China and then turning northeastward to return to the northern cyclic path. Such storms do not often bring winter rain, but they create short periods of cloudy, cool, or gusty weather and bring snow to the higher mountain ranges.

Temperature. The average January temperature over a considerable part of Siberia is below −4° F (−20° C), and in the Verkhoyansk region it reaches −58° F (−50° C). Along the coastal areas, the proximity of the Pacific Ocean moderates the temperatures to from 23° F to 5° F (−5° C to −15° C). The January isotherm (a line connecting points of equal temperature) of 32° F (0° C) passes through Samarkand, Peking, and the island of Honshu. An isotherm of 68° F (20° C) is traced along the Tropic of Cancer and one of 77° F (25° C) along the Equator. In July, when the average temperature is 86° F (30° C), the maximum temperatures are found in West Asia and in the Thar and Takla Makan deserts. The 68° F (20° C) isotherm moves as far as 55° to 60° N, but near the cool Pacific Ocean it bends to the south. Along the northern coasts of Asia the average temperature in July is below 50° F (10° C), which is typical for a tundra



Average temperatures for January and July in degrees Celsius for Asia.

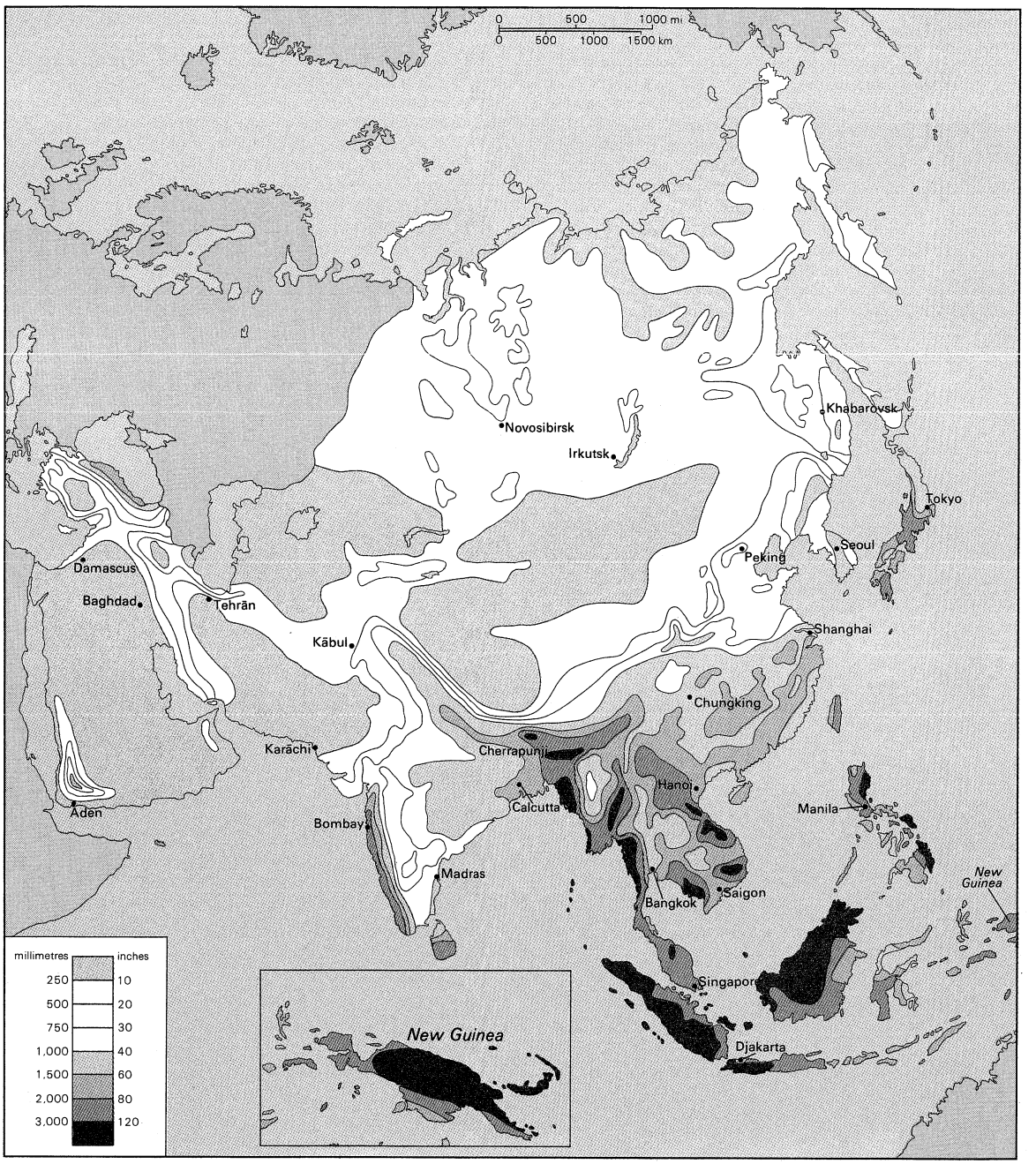
climate. The greatest amplitude in annual temperature range occurs near the "cold pole," which has surprisingly warm summers; the annual range may exceed 175° F (97° C).

Rainfall. Annual rainfall in the equatorial belt is approximately 80 inches; it is 80 to 120 inches and more (300 to 500 inches in places) on the windward maritime slopes in South and East Asia. In Cherrapunji (Meghalaya) 900 inches of rain fell in seven months in 1891. Precipitation is less than 40 inches on the lee slopes of the subequatorial regions. In the subtropical and temperate monsoon climates there is adequate rainfall, amounting to about 24 to 40 inches. Precipitation is less than ten inches in Eastern Siberia and averages six to eight inches (but may be less than four inches in some places) in the deserts of West, Middle, and Central Asia.

Climatic regions. The distribution pattern of rainfall throughout the year is varied. Relatively uniform moisture is characteristic of the Asian equatorial zone. Maximum summer precipitation and minimum winter precipitation are the rule in the subequatorial zones and in other regions with monsoon climates, as well as in those areas where there is summer movement of the fronts—the polar front in the mountains of Southern Siberia and the Arctic front in the sub-Arctic regions. Wet winters and dry summers are typical of the Mediterranean climatic region in West Asia, where precipitation is associated with the winter activity of the polar front. This polar front activity, accompanied by maximum precipitation, occurs in the spring in the interior parts of the West Asian highlands. Summer and winter precipitation merges in some parts of Asia. In the Kolkhida climate, the summer rains—brought by the northwesterly Atlantic air currents—merge with the cyclonic Mediterranean winter rains. In some areas of Japan and eastern China there is uniform precipitation when, in addition to the summer monsoon, the winter monsoon brings moisture.

As the aggregate result of these various meteorological patterns, the following types of climate may be distinguished in Asia: the tundra climate (associated with the cold, treeless plains of the Arctic lowlands of Asia); the cold, sharply continental climate of Eastern Siberia; the cold, moderately humid Western Siberian climate; the humid, subtropical Kolkhida climate; the desert climate

Types of climate



Average annual precipitation for Asia.

of the temperate zone; the Mediterranean subtropical climate of the western edge of West Asia; the subtropical desert climate; the mountain-steppe highland subtropical climate of West and Central Asia; the alpine desert climate; the climate of the Eastern Pamirs, Karakoram Mountains, and Tibetan Highlands; the climate of the tropical deserts; the temperate monsoon climate of the Soviet part of the Far East, and northern parts of Japan and East China; the subtropical monsoon climate of Southern Japan and of Southeastern China; the subequatorial monsoon climate of South Asia, eastern Java, and the Lesser Sunda Islands; and the equatorial climate of the Greater Sunda Islands.

Many climatic variants can be distinguished that are associated with such local topographical features as the degree of exposure of the slopes, the protective effect of the mountains, and altitudinal zonality, with temperatures dropping as the altitude increases. Low temperatures, however, are also found in low hollows where cold air stagnates or on coasts where air is chilled by cold ocean currents. The mountain climates, evidently, represent variants of those climates that are determined by latitude. All the various features of the types of climate mentioned exert a strong influence on other natural conditions, as well as on the landscape as a whole.

Urban climate. Distinctive variations of climatic characteristics result from the cultural and economic activities of human society. One example of this is provided by the microclimates associated with the cities and with large industrial complexes. The emission by the cities of quantities of dust and gases produces alterations of temperatures and changes in wind patterns. Such conditions are characteristic, for example, of Tokyo and the industrial region of northern Kyushu in Japan, of Calcutta and the industrial area of the northeastern part of peninsular India, and of the industrial regions of the Kuznetsk Basin in the Soviet Union.



The Ganges River winds through the fertile plain east of Delhi, in northern India.
Harrison Forman

DRAINAGE

Rivers. Asia is a land of great rivers. The Ob, the Irtysh, the Yenisey with the Angara, the Lena (with the waters of the Aldan and the Vilyuy), the Yana, the Indigirka, and the Kolyma rivers all flow into the Arctic Ocean. Among rivers draining into the Pacific Ocean are the Anadyr, the Amur (combined with the Sungari and the Ussuri), the Huang Ho, the Yangtze, the Hsi, the Song Hol, the Mekong, and the Chao Phraya. The Salween, the Irrawaddy, the Brahmaputra, the Ganges, the Godavari, the Krishna, and the Indus flow into the Indian Ocean, as also does the Shatt al-Arab, which is the confluence of the Tigris and Euphrates rivers. Only small mountain rivers flow from Asia into the Caspian, the Sea of Azov, the Black Sea, and the Mediterranean. The Amu Darya, the Syr Darya, the I-li, the Chu, the Tarim, the Gilmend, and the Tedzhen rivers empty into vast interior basins. Some of these rivers end in lakes, some end in dry deltas in the sands or salt marshes, and some flow into oases, where all the water is used to irrigate fields or else it evaporates.

Rivers of the Pacific

All the Siberian rivers freeze over in the winter, and some freeze to the bottom. In spring widespread flooding occurs. These rivers are important communication routes, being used by boats during the summer and as roads for sleighs in winter; they also teem with fish.

In the dry regions, where drainage is landlocked, the only large rivers are temporary ones fed by snow and glacier water in the mountains; they reach their peak water levels in summer. Rivers that are not fed by mountain runoff have little water; their levels vary sharply, and periodically or occasionally they dry up. The rivers of the monsoon climate regions reach their maximum volume in summer and are utilized for irrigating the rice fields. The Asian rivers in the vicinity of the Mediterranean that are not fed by mountain snows grow shallow in summer and sometimes even dry up. In the equatorial regions, however, the rivers are perennially full of water.

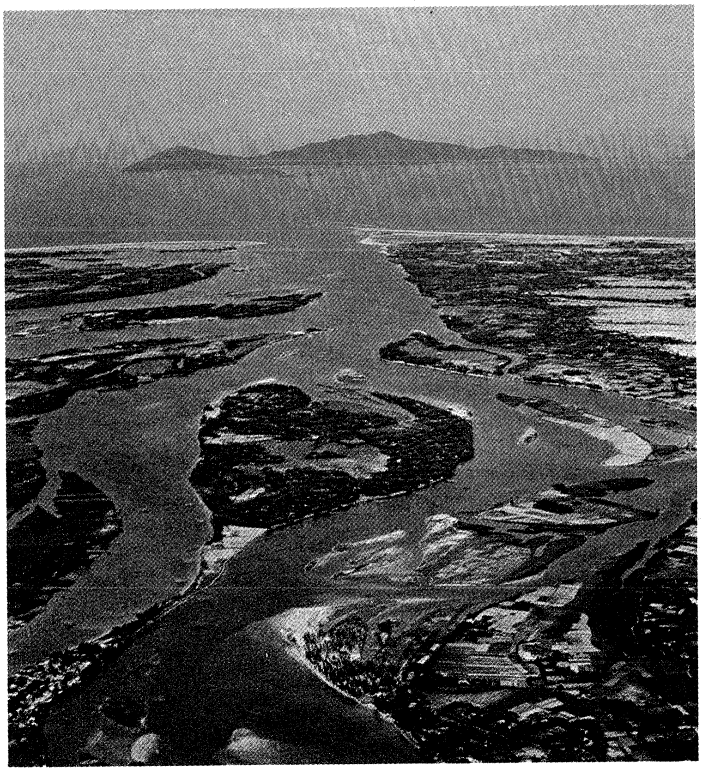
Lakes. The lakes of Asia are numerous, varying considerably in size and origin. The largest of them—the Caspian and Aral seas—are the remains of larger seas that existed here previously. Lakes Baikal, Issyk-Kul, and Khubsugul, the Dead Sea, and others lie in tectonic depressions. The basins of Lakes Van, Sevan, and Rezaĭyeh (Rezaĭe) are, furthermore, encircled by lava, and Lake Telets was gouged out by ancient glaciation. A number of lakes were formed as the result of landslides (Lake Sarez in the Pamirs); karst processes (the lakes of western Tavr, in Turkey); or the formation of lava dams (Tsin Bokhu in Northeast China, and several lakes in the Kuril Islands). In the volcanic regions of the eastern Asian islands, in the Philippines, and in the Malay Archipelago, lakes have formed in craters and calderas. The subArctic has a particularly large number of lakes; in addition to lakes formed as a result of permafrost and subsidence, there are also ancient glacial moraine lakes. Many lagunal lakes occur along low coastlines.

The lakes in the internal drainage basins—such as Koko Nor, Tuz, and others—are usually saline. Lake Balkhash has freshwater in the west and brackish water in the east. Lakes through which rivers flow are freshwater and regulate the flow of the rivers that issue from them or flow into them; examples of these are Lake Baikal, associated with the Angara River; Lake Khanka (the Sungacha and Ussuri rivers); Tung-t'ing Hu and P'o-yang Hu (the Yangtze River); and Tonle Sap (the Mekong). Large reservoirs have also been created by the construction of hydroelectric stations.

Subterranean water. In arid regions, subterranean water is often the only source of water supply. Large accumulations are known to exist in artesian basins and beneath the dipping plains at the foot of mountains; these are associated with the extensive oases of Middle Asia, Kashgaria, and many other regions.

SOILS

The soils of Asia are distinctly marked by the combined effects of climate, topography, hydrology, organic nature, and the economic activities of man. The horizontal zonality of the climate, the drainage conditions, the existing plant and animal life, and agriculture are each related to the considerable meridional extension of Asia, which is accompanied, of course, by a horizontal zonality of the soil cover, which is especially clearly defined in the plains of the continental sector.



A portion of the delta of the Mekong River as it flows through South Vietnam and empties into the South China Sea.
M. Gifford—De Wys Inc.

The Arctic zone. In the Arctic, where glacial and Arctic deserts predominate, the processes of soil building are manifested only in rudimentary form. The soils here are saturated and are low in humus. The sub-Arctic north of Asia is occupied by a timberless zone of tundra vegetation. Beneath the tundras, specifically tundra-type soils are formed, which are characterized by poor drainage (associated with the proximity of permafrost) and only a short period in which the decomposition of organic substances is possible. This results in the accumulation of undecomposed organic residues in the form of particles of peat. The poor drainage creates an oxygen-free medium in which the bluish substance known as gley is formed. Thus, peaty-gley soils are most characteristic of the tundra. There are widespread occurrence of movement by mud glaciers; heaving of the ground because of frost; settling or caving in of the ground from thawing; and the formation of stone rings around central areas of debris in bouldery regions.

The forest tundra. Farther south stretches the transitional belt of the forest tundra, where tundra and sparse forest alternate with regularity. Here tundra soils alternate with the soils of the taiga (the cold, swampy forested region to the south of the tundra, characterized by very low temperatures). The soils below the frozen taiga are called cryogenic (*i.e.*, having very low temperatures). In the mountainous regions the peaty-gley soils are replaced by mountain tundra and weakly developed, often embryonic soils of detritus and stony fragments.

The forest zone. The forest zone occupies the largest part of the temperate zone. Characteristic of soil formation in the forest zone is the leaching process. The forest leaves and needles that fall, together with dead remains of the sparse grass cover, are subjected to decomposition by organic acids in the litter of the forest floor; the duration of the summer season and the amount of precipitation are sufficient for complete decomposition of the soluble soil components; the soil solutions transport them and leach them into deeper soil horizons (layers). The undecomposed quartz grains remain in the upper horizon, which is therefore infertile; this layer resembles light-gray ashes, which is the reason soils of this type are called podzols ("under ashes"). The various subzones of the forest zone are subjected to different degrees of leaching. A dense, rusty-brown horizon of wash-down (deposition in an underlying layer of soil) underlies the podzolic portion of the soil profile (vertical section of the soil); its colour is related to the accumulation of iron and aluminum oxides. This layer, called orstein, or iron pan, is impervious to water and contributes to the self-swamping of the taiga forests. East of the Yenisey River, where the forest zone for its entire breadth is in the grip of permafrost, soil drainage (and consequently the leaching process) is made more difficult; the transfer of substances is complicated by freezing and thawing, and therefore the typical podzols are replaced by specific cryogenic taiga soils. Marshes and bog-type soils are widely distributed over a considerable part of the taiga subzones.

The leaching process

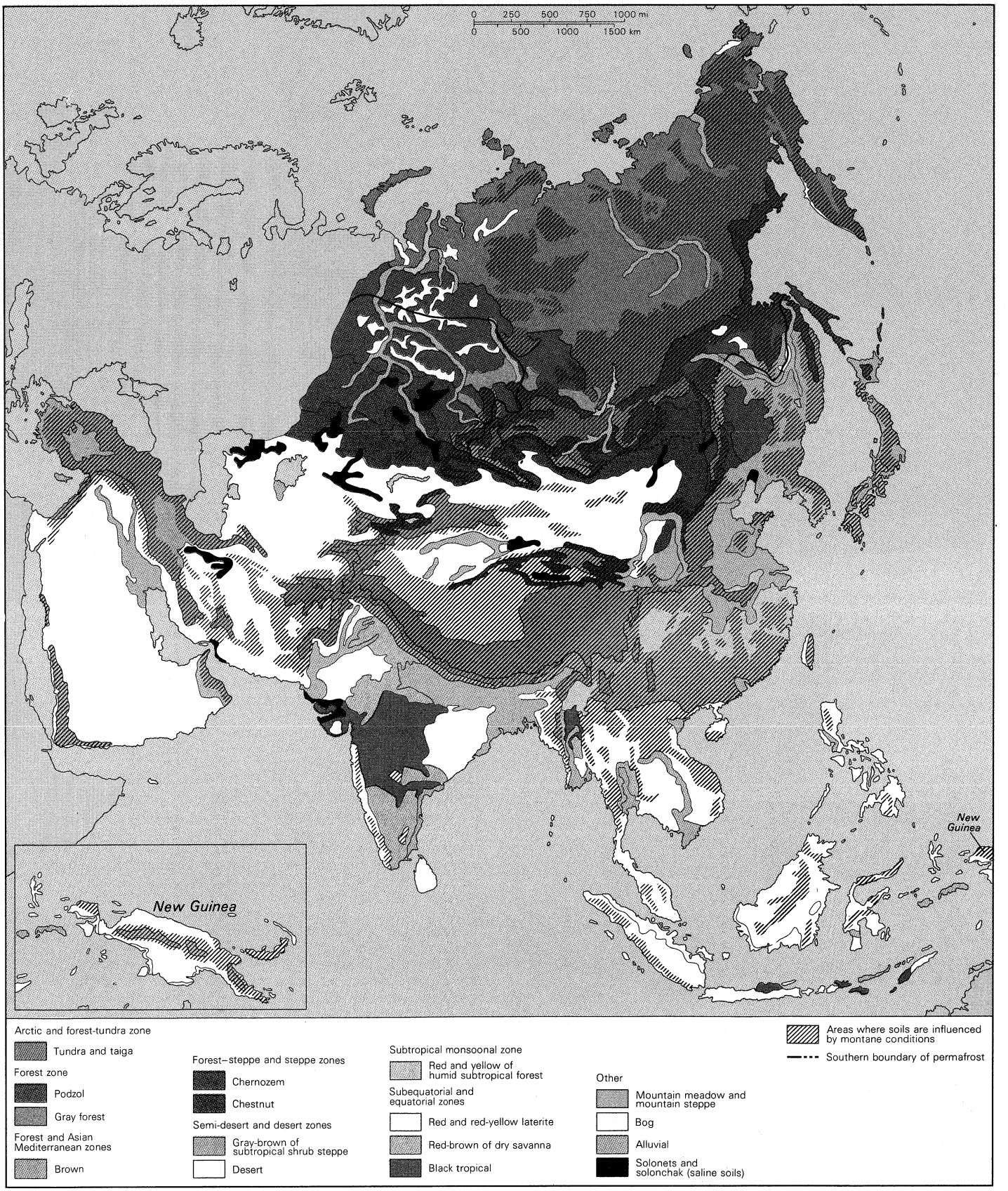
The deciduous forest subzones of Asia form two distinct areas. In Western Siberia there are small-leafed, primarily birch or aspen, forests on gray forest soils. They are more gray in colour than the podzols because of the greater amount of organic substances—such as tree leaves and a more abundant grass cover—feeding these soils. This explains their higher content of humus, as well as their greater fertility. The second section of the deciduous forest subzone has survived in the Far East, stretching from the Lesser Khingan Mountains in the north to the Japanese island of Honshu; in this subzone abundant warmth and moisture intensify chemical weathering, and iron oxides accumulate even in the surface soil horizons. In this manner brown forest soils, known as forest burozems, are formed.

The forest-steppe and steppe. Soil cover in the forest-steppe region is formed when the ratio of precipitation to evaporation is in equilibrium and as the leaching process of the wet season alternates with the upward flow of the soil solutions during the dry period. Under these conditions, with abundant organic material resulting from the dense vegetation, intensive accumulation of humus takes place in the soil, and dark-coloured soils are formed that are the most fertile in all of Asia; known as chernozems, they are the most fertile as well as the thickest of the forest-steppe and mixed grass subzones. Characteristic of the wooded-meadow plains of the Amur Basin (the "Amur prairiès") are meadow soils that are dark, semi-boggy, and often composed of blue gley. In the drier steppes, where vegetation is sparse, the amount of humus is reduced and the content of unleached mineral salts is increased; transport of the dissolved salts to the surface by the upward flow of soil solutions is also intensified. Associated with this process is a bleaching and salinization of the soil. The drier steppes thus form a transitional zone from the shallow southern chernozems to the chestnut soils. Broad expanses of the forest-steppe and steppe are under cultivation and serve as rich granaries for the cultivation of grain crops. Severe wind erosion occurs during the hot, dry seasons. In many areas impoverishment of the soil has also developed as the result of surface washout and gully erosion, despite preventive efforts.

Soils rich in humus

Semidesert and desert. Through inner Kazakhstan and Mongolia stretches a zone of semidesert, and through Middle Asia, Dzungaria, Takla Makan, and Inner Mongolia a belt of temperate zone deserts. A belt of subtropical deserts extends through the Levant, the Iranian highlands, and the southern edge of Middle Asia. Beneath the semideserts, with their mosaic of desert and arid-steppe vegetation, light-chestnut and light-brown semidesert soils form; these are low in humus but contain an abundance of strongly alkaline soil. Beneath the deserts, where the supply of organic substances, as well as the humus content, is extremely low, gray-brown soils form in the temperate zone; gray desert soils (sierozems) form in the arid subtropics. Here there is a great deal of saline soil, and agriculture is possible only with the use of irrigation, which is feasible in the infrequent oases, where specific cultivated types of sierozems have formed.

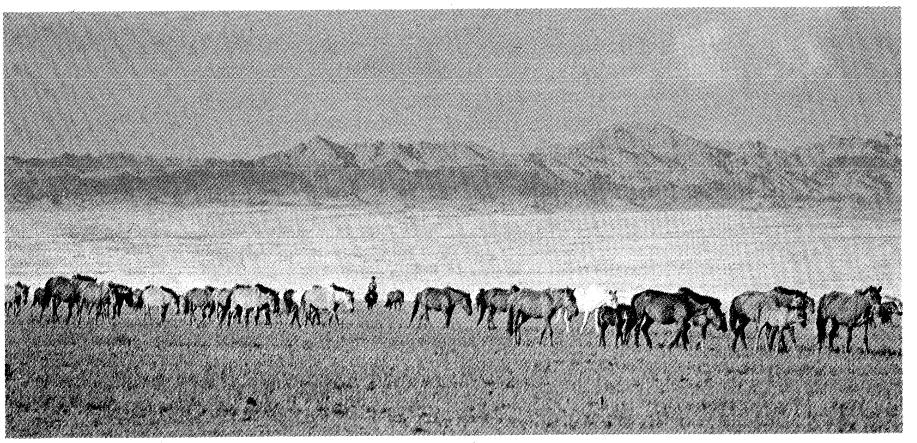
Only in western Asia is the tropical desert zone clearly defined. Here broad expanses are characterized by embryonic soils and desert crusts, as well as by blowing sands.



Soils of Asia.

The Asian Mediterranean. In the maritime areas of the Asiatic Mediterranean—Asia Minor and the Levant—xerophytic vegetation (vegetation structurally adapted to exist with very little water) of the Mediterranean scrub-woodland types, known as maquis (evergreen), shiblyak (an association of bushes and scrub characteristic of the Balkan Peninsula in Europe), and frigana (low-growing, prickly, and cushion-like bushes), is prevalent. The predominant soils under such vegetation are brown; they have accumulated iron as a result of the intense chemical weathering during the wet Mediterranean winter and of the upward flow of soil solutions dur-

Iron-bearing soils



Horses grazing on the steppe grass in the southern region of the Gobi. The Altai Mountains In the background form the southwestern boundary for the desert.
Paolo Koch—Rapho Guillumette

ing the dry summer. Frigana vegetation, characterized by thorn bushes, is widely represented in the West Asian semidesert highlands. Here soils have developed that are transitional between the brown soils and the sierozems.

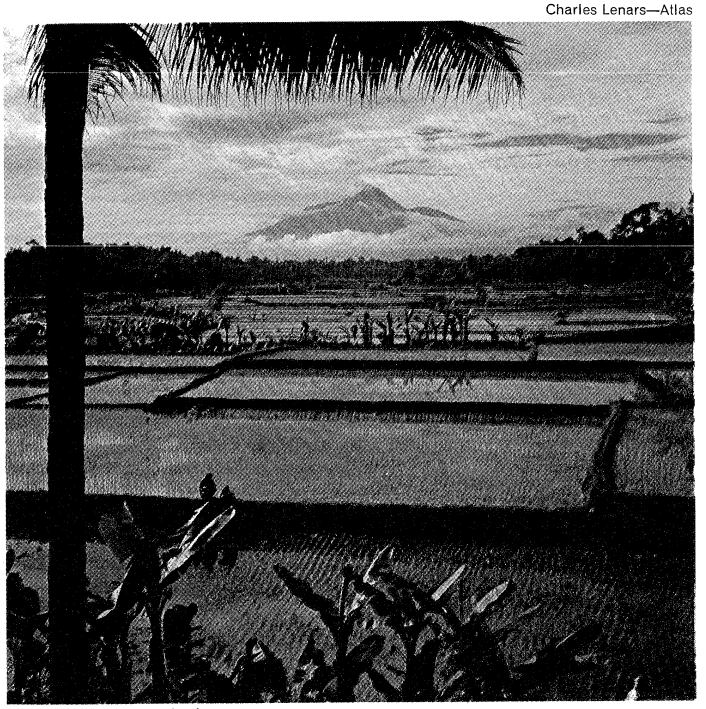
The subtropical monsoonal regions. Typical of the monsoon subtropics are the evergreen forests of the southern portion of the Korean Peninsula, of southwestern Japan, and of southeastern China. Intensive chemical weathering during the simultaneously warm and wet summer monsoon season results—as it also does in the more southerly torrid zones—in the decomposition and carrying away from the soil of many minerals, the accumulation of residual iron and aluminum oxides, and the consequent predominance of red and yellow soils as well as of podzolized soils. Agriculture, with the irrigation of rice fields, is especially widespread on the alluvial soils of the plains, which have been cultivated continuously by farmers for thousands of years. Terracing of the slopes is a widely applied practice.

The subequatorial and equatorial regions. The subequatorial zones of Asia are covered by savannas (grassy parklands) and dry-tropical deciduous forests, primarily situated in the rain shadow on the leeward slopes, and by wet-tropical evergreen forests on the rainy windward slopes facing the sea. Intensive leaching followed by evaporation is characteristic of these soils. Under the wet tropical forests, red-yellow lateritic (leached and hardened iron-bearing) soils predominate; beneath the savannas and dry tropical forests there are red lateritic soils that change, with increasing aridity, to red-brown and desert brown soils. Beneath the dry savannas of peninsular India there are unique black soils called regurs that are thought to be developed from basalt country rock.

In the equatorial zone (southern Malaysia and the Greater Sunda Islands), typical tropical rain forests have developed. In southwestern Sri Lanka and in Java they have been almost entirely replaced by an agricultural landscape in which mountain slopes and hills are covered with plantations of tea, coconut palms, and rubber trees. The soils are lateritic and are red-yellow or brick-red, with marginal degrees of laterization.

In the valleys of the subequatorial and equatorial zones, alluvial soils predominate; they have been developed by thousands of years of cultivation and irrigation of the rice fields. Artificial terracing of the slopes is practiced on a very large scale in the mountainous regions, both for purposes of irrigation and in order to prevent soil erosion.

Charles Lenars—Atlas



Tropical vegetation surrounding fertile rice paddies on the island of Java, Indonesia. The active volcano Gunung Merapi rises above the clouds in the background.

The mountains. In the mountains, zones of different soil types are found at different altitudes. As a rule they are skeletal, underdeveloped soils, clearly reflecting the differences in rock structure and origin and in the degree of exposure of the slopes. The boundaries of the vertical zones become higher from north to south; the number of zones increases. Mountain soils also correspond to the different vegetation zones occurring at different altitudes. Under the mountain forests of the northern portion of the temperate zone are mountain-podzolic soils. Mountain variants of the gray forest soils and taiga-cryogenic soils have also developed; above the tree line they are replaced by mountain-tundra soils. Mountain chernozem and mountain chestnut soils develop beneath the mountain steppes in the southern part of the temperate zone; brown mountain-forest soils are found in the wetter regions; beneath the alpine and subalpine meadows the soils are of the mountain-meadow type. Mountain red podzolized, yellow earth, and other lateritic soils are predominant in the wet areas of the lower latitudes; in the dry areas mountain brown and gray-brown soils, as well as mountain gray soils, occur. The alpine steppe and desert soils of Central Asia have a number of distinctive characteristics.

Underdeveloped mountain soils

The correlation of the vertical soil zones and of the landscape zones varies with the whole spectrum of vertical zonality. A zone of forest, followed higher up by meadows, with snow cover at the highest altitudes, is characteristic of the western maritime regions. On lower slopes in the western Caucasus, for example, broadleafed mountain forests occur on brown mountain-forest soils; above these are coniferous forests on mountain podzolic soils, followed by stunted trees, followed in turn

by subalpine and alpine meadows on mountain-meadow soils, while near the highest ridges perennial snow and glaciers are found. Associations of desert, steppe, meadowland, and snow zones are widespread in the interior of Asia and sometimes include mountain-forest zones. Thus, characteristic of the Tien Shan, for example, is the predominance of mountain-desert and semidesert landscapes, which occur in association with gray-brown and brown mountain soils in the foothills of the ranges, while higher up are mountain steppes associated with mountain chestnut soils and mountain chernozems. Under parts of the mountain forest-steppe and the mountain forests, the soils are podzolized. Above the stunted forest zone in the maritime sector, mountain-meadow soils occur beneath the meadows, but here, too, a distinctive snowy type of landscape occurs in the vicinity of the ridges.

Typical of the mountains of Eastern Siberia are the taiga-tundra spectra that occur in vertical zones; thus, mountain taiga on taiga-cryogenic soils is followed by a zone of dwarfed trees, followed by mountain tundra, and then finally by bald peaks.

In eastern Asia, the subalpine and alpine meadow zones with mountain-meadow soils usually disappear; instead, mountain-forest landscape extends as far up as the vicinity of the crests and is succeeded only by a zone of stunted trees. The spectra of the alpine regions of South Asia (the Himalayas) are distinguished by the most complex variety of vegetation and soil types. (Y.K.Y.)

III. Vegetation and animal life

VEGETATION

In Asia an immense range of vegetation is found, resulting from the continent's wide diversity of latitude, altitude, and climate. Natural conditions, however, are not entirely responsible for the associations of trees, plants, and grasses of Asia; natural landscapes have been transformed by 80 centuries of farming.

The geographic pattern of vegetation. *North Asia.* The natural landscape has been least affected by man in sparsely populated North Asia. Vast plains, continentality, and the nearness of the Arctic Ocean explain the

The tundra

presence here of a zone of tundra—cold, treeless plains with permanently frozen subsoil—similar to that found in the western Soviet Union and in Canada. In more flourishing parts the tundra has a discontinuous covering of lichens, mosses, sedges, rushes, some grasses, cushions of bilberries, and dwarf trees of willow and birch; in the far north, lichens grow on favourable hillsides. Thanks to the greater number of hours of daylight during the summer solstice in June, when the Arctic Circle receives the same amount of light energy as the tropics, the tundra at this season is covered with bright flowers. Nevertheless, climate conditions are extreme; in Severnaya Zemlya, along the Arctic coast, thawing begins in May and frosts begin in August, although in some years frosts may occur at night throughout the short summer. The soil never thaws below a depth of two or three feet; consequently, hollows are badly drained and turn into peat bogs. Windy conditions speed up evaporation, and the frozen soil cannot absorb water to compensate for this, so that surface drought often allows wind erosion and the transport of sediments deposited by annual riverine floods.

The tundra belt extends still farther south on higher ground. In the Arctic Urals tundra begins at about 3,000 feet, but at latitude 53° N it begins at 4,250 feet. Tundra extends over large areas of the Cherski, Verkkoyanski, and the Kamchatka mountains.

The taiga

The taiga zone—a belt of coniferous forest—begins south of the tundra, after a transitional zone of "wooded tundra" and forest galleries, found along streams between the tundra-covered watersheds. Taiga, although essentially coniferous, is mixed with hardy deciduous trees such as aspen and birch; there are sections of grass and shrub steppe in the drier zones. Larches account for 37 percent of the Siberian forest, which covers 2,700,000 square miles; pines cover 24 percent and spruce 4 percent. The geographic distribution of particular types of vegetation is determined chiefly by climate. Spruce, for example, unable to survive temperatures below −36° F (−38° C), is not found east of the Yenisey River. The taiga has a thin undergrowth of cranberries and bilberries, and there are numerous extensive peat bogs.

In Soviet Asia broadleaf deciduous forest does not extend eastward beyond the Yenisey River, where it gives way to the coniferous forests of Central Siberia, reappearing in Eastern Siberia near the Okhotsk Sea; here poplars, birches, and alders are numerous, as well as various conifers and larches. Forests around the Ussuri River include maples, ashes, walnut, elms, and lindens, in addition to species already mentioned. In the direction of China, as described below, the landscape becomes transitional.

South of the Siberian forest, the zone of prairie (continuous herbaceous cover) is not uninterrupted; forest frequently gives way to steppe (discontinuous cover).

Tibet, which is chiefly dry and cold, has a scattered vegetation of halophilic bushes (bushes flourishing in a salty environment) and Artemisia's tufts.

The Far East. In the Far East, the monsoon climate brings hot and rainy summers, giving rise to a great variety of temperate and tropical vegetation. China has the most varied vegetation of any country in the world, with about 15,000 species, excluding mushrooms and mosses. Far Eastern forests are fascinating to botanists because of the variety of their plant life; many trees have large, bright evergreen leaves, and there is a dense undergrowth with abundant creepers.

Japan has 68 percent of its area under forest, whereas China is almost entirely deforested, although sizable tracts remain untouched in the remote, rugged regions and many small areas have been reforested. The reason for this is that Japanese scenery was traditionally respected, and strict forestry regulations were severely enforced. The best examples of Far Eastern forest are found in Japan; for example, in the Kii Peninsula. Conifers are the principal species used for reforestation in Japan's policy of restoring its forests in order to meet the industrial need for wood.

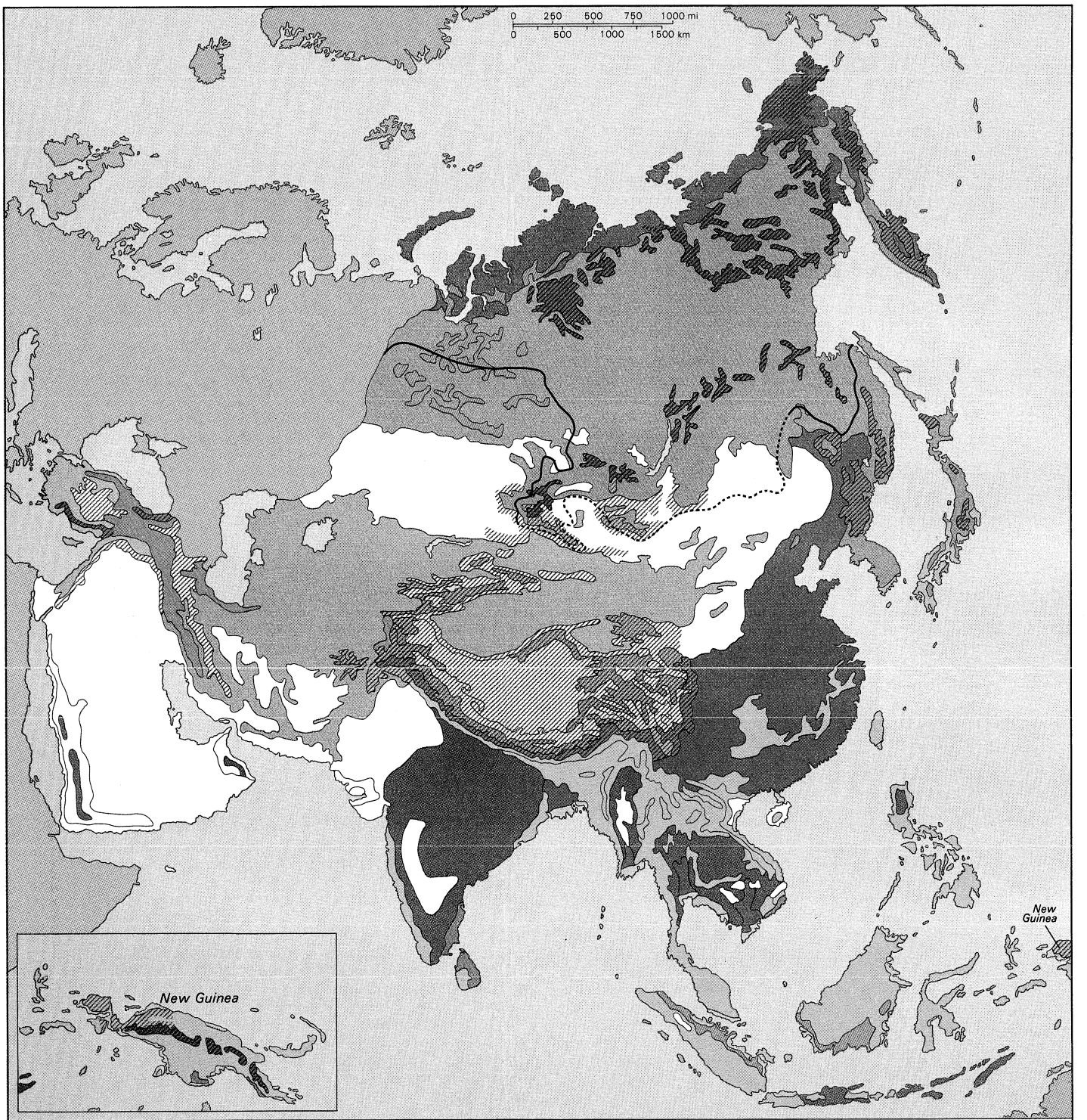
North of the Yangtze River, much of China was covered by primeval deciduous forest, most of which has been removed through farming. South of the Yangtze the "true" Chinese forest was prevalent before 1800. A wild growth of trees and shrubs survives, however, throughout the cultivated areas, and park-like tree growth and stands of bamboo are widespread. The "true" forest included 60 different genera of tall trees, including—among the temperate genera—oak and maple, linden, chestnut, hornbeam, and a species of hickory. Tropical genera included magnolia, the tulip tree, the camphor tree, the Spanish cedar, liquid ambar (a tropical tree of China), catalpa, and lianas (vines). A variety of conifers of both hemispheres was also to be found, and in the mountains of eastern Szechwan there grew a rare and ancient Chinese conifer, the metasequoia. Palm trees are found throughout South China and South Korea as well as in the southern half of Japan; many varieties of bamboo are also found in these regions.

The "true" Chinese forest

The Peking government is proceeding energetically with a program of reforestation. The new forests, however, consisting largely of pines, do not resemble the primeval Chinese forest.

South Asia. The wettest parts of peninsular India, such as the Western Ghāts, and of Southeast Asia have magnificent forests noteworthy for the variety of their plant life. The culturally controlled forests of Java and Sumatra alone include over 3,000 species of trees. The variety of tropical vegetation is accentuated by the diversity of influences, such as that of altitude upon climate. Temperate pines are found in Sumatra and in the Philippines, where eucalyptus also thrives; oaks occur in the mountains of New Guinea; and austral *Podocarpus* (evergreen trees with a pulpy fruit) in the eastern Himalayas. In the seasonal monsoonal climatic zone of central Indonesia, Thailand, Burma, and southern India, the teak forest thrives as an open park-like cover with little underbrush.

A notable feature of South Asian vegetation is the Dip-



Vegetation zones of Asia.



Taiga vegetation growing on the Siberian Lowland of south central Siberia.
Sovfoto

terocarpacea family (yielding aromatic oils and resins), which is here represented by more than 500 species. Mangrove thrives in muddy deltas along the South Asian coastline. In the southernmost areas, the bogs on the land-side edges of the mangrove swamps abound with the semi-aquatic nipa palm, the leaf fronds of which are widely used for thatching.

Hevea brasiliensis, the most successful rubber plant, is found in tropical Southeast Asia, where it was introduced from South America in the 1870s; it is particularly important in plantations in Malaysia and Indonesia.

Primeval evergreen rain forest remains in a few parts of South Asia. Secondary forest, which develops where temporary clearings have been made in the original forest, covers a much larger area. In drier tropical areas, secondary forest is deciduous; where the dry season is particularly long, park forest, with trees spaced at wide intervals, is found, as in the "sal" (East Indian hardwood) forests of India. Extensive fires in such areas have resulted in a herbaceous landscape, as in the cogonales (areas of coarse tall grasses, used for thatching) of the Philippines.

Mountain vegetation

In the higher mountains of Southeast Asia the cooler humid tropical climate gives rise to deciduous and coniferous temperate forest at altitudes of between about 4,250 feet and 10,000 feet. Above this level, low forests of plants, mostly shrubs of the heath family, are often found. Diverse types of trees grow in the mountain forests of the region. The Arakan Mountains of Burma, for example, are covered with a thick mantle of little bamboos. In the eastern Himalayas sal is intermingled with *Castanopsis* (a small genus of nut-bearing trees) and pines. Above these are found forests of shrubs and trees of the laurel family and, higher still, oaks and conifers; between about 10,000 feet and 13,000 feet, forests of firs occur. The central Himalayas present strikingly beautiful landscapes in the following upward succession: dry sal forest; pine forest; cedars, spruces, pines, and oaks; firs, birches, and tall rhododendrons; above 13,000 feet, bushes of rhododendrons, together with junipers; above 16,000 feet, perpetual snows.

West Asia. In western Asia naturally wild vegetation no longer occurs in clearly defined zones but is dispersed in small areas. The region is predominantly arid: desert-like depressions such as the Kyzylkum in the Kazakh and Uzbek regions of the Soviet Union and the Rubʿ al-Khali (Empty Quarter) of the Arabian Peninsula contrast with the moist, well-forested mountains that lie between them. Three climatic zones, however, characterize western Asia: a continental climate in the northern regions; a dry zone, except where northerly winds bring moisture to the mountains, to the south; and a Mediterranean climate along the western edges.

A few examples of the variety of vegetation associated with these climatic zones may be cited. In the valleys of the dead rivers of the Kara-Kum Desert grows a strange tree, the saxaul, which is oddly shaped, gnarled, and leafless; between the galleries of saxauls the desert is interspersed at very wide intervals with bushes and tufts of grass. A fringe of steppe covers the area between the Fertile Crescent (which sweeps in an arc from the Tigris–Euphrates Valley to the Mediterranean) and the north and west of the Syrian Desert. With more than 2,000 species of plants—more than in the whole of the Sahara—the borders of the latter desert are noteworthy for their floral variety. The moist northern slopes of the Pontic Mountains in northern Turkey are covered by magnificent forests of beeches and conifers, with an undergrowth of tall cherrylowels, hollies, and creepers. This type of forest is also found in Georgia and on the northern slopes of the Elburz Mountains in Iran. Along the Mediterranean border of Asia the vegetation is similar to that in other parts of the Mediterranean region: holm oak (an evergreen oak), Aleppo pine (characteristic of the city of Aleppo), cistus, mastic tree (which yields mastic, used as a chewing gum), and other species are found in landscapes of thick underbrush and open scrubland.

"Pontic" forest

Man and vegetation. *Vegetation in traditional civilization.* Asia's indigenous vegetation has provided many edible products, such as stone fruits, citrus fruits, bananas, mangoes, soybeans, and tea; building materials, such as wood, bamboo, and thatch; cotton and straw for clothing; bamboo, widely used in the making of utensils; and the bark of the paper mulberry, used for making bark cloth and paper. In addition, silkworms are fed upon mulberry leaves; lacquer is made from *Rhus vernicifera* (lacquer tree); and a multitude of other items are obtained from plants, including a styptic for stopping hemorrhage, an anti-asthmatic agent made from mahuang (a plant yielding ephedrine), and a fine fibre, extensively used in weaving, derived from ramie, a plant of the nettle family.

Similarly, the forested areas of Southeast Asia provide the sparse population with a wide variety of products: firewood; timber for construction; foodstuffs from a variety of trees, plants, and fungi, including sago, taro (a plant with an edible, starchy, tuberous rootstock), and mushroom; scented resins from sandalwood and eaglewood (an East Indian tree with soft resinous wood); and dyes from a variety of tubers. Probably the irrigated cultivation of taro and rice began in Southeast Asia.

Wheat is indigenous to the humid hill margins of western Asia; it still grows spontaneously in the hilly fringes of Asia Minor. Cherry, peach, and pistachio trees, as well as vines, were domesticated in the mountains of western Asia. The large acorn cups from the forests of *Quercus aegilops*—a species of oak, in western Asia Minor—are valued for the abundant tannin they contain.

Commercial forestry. China and the Indian subcontinent, with their enormous populations, are today poor in timber resources. The best resources of timber for building and for paper manufacture are in Siberia. Japan produces about 1,600,000,000 cubic feet of timber annually from its controlled forests. The tropical and equatorial forests of South Asia are difficult to exploit; because of the diversity of species found in these forests, commercially valuable trees are mixed with a majority that is of no economic value. Some deciduous forests, however, are of commercial interest since they are more homogeneous and consist of good quality trees. Included in this category are the forests of teak, sal, and ironwood. Such forests are important in the Burmese economy, timber being among the more important exports. The countries of Indochina, as well as Thailand, Malaysia, and the Philippines, are covered with extensive forests, but only a small fraction of their timber resources are commercially exploited. As in Burma, teak is the most valuable wood, but much of the remaining natural

Timber

timber is of low quality and not easily marketable. It is, nevertheless, increasingly being recognized that the forests of these countries form potentially valuable resources. Malaysia, in particular, is encouraging the development of forestry, and although overshadowed by rubber and tin, forestry is one of the most important Malaysian industries. In western Asia, Turkey has launched a program to increase production from its forests and thus reduce lumber imports. (P.Gu./Ed.)

ANIMAL LIFE

The Himalayas, stretching from east to west, form a barrier largely preventing the movement of animals southward or northward. Thus, Asia north of the Himalayas, with parts of western Asia and most of the Far East, belongs to the Palearctic (literally "ancient Arctic") zoogeographical region. Asia south of the Himalayas is called the Oriental, or Indian, Region. The boundary dividing these zones east and west of the Himalayas is not well marked, however, as there the mountain chains often have a north–south trend facilitating migration of animals between them.

The Palearctic Region. A distinction can be made between the animal life of the tundra in the north and that of the adjacent taiga farther south; the taiga in turn merges into the steppes, which have their own distinctive forms of animal life. The tundra subsoil is frozen throughout the year; hence, burrowing animals cannot live there, and, as the tundra is partly free from snow only during the short summer, conditions for life are poor. Most animals, including reindeer, Arctic hare, Arctic fox, and wolf, live here in summer only and migrate in autumn, but the lemmings (small rodents of circumpolar distribution) stay, feeding on the herbage buried beneath the snow. Hibernation is impossible, for the short summer does not allow the necessary accumulation of food reserve in the body.

During the summer, birds are numerous but they also desert the tundra in winter, except for such birds as the willow grouse and the ptarmigan, which live in tunnels in the snow, feeding on berries and leaves. Many species of waders, such as the gray plover, the sanderling, and several kinds of sandpipers, migrate to the tundra and breed there in the summers, feeding principally on the mosquitoes in the wet areas. Mosquitoes are also the staple food of passerine birds (true perching birds), such as the snow bunting and the Lapland bunting. Gyrfalcons (a subgenus of large Arctic falcons), rough-legged buzzards, and skuas (large, dark-coloured rapacious birds of northern seas) prey on these smaller birds and on lemmings. Several kinds of geese and ducks, Arctic tern, and three species of divers occupy the moist parts.

The taiga fauna

The taiga fauna is much richer than that of the tundra. The taiga is the haunt of brown bear, wolf, glutton (a kind of wolverine), otter, ermine, sable, lynx, elk, forest reindeer, hare, and several kinds of squirrel. Birds include various kinds of grouse and woodpecker, pine grosbeak, crossbill, siskin, redpoll, red-spotted bluethroat, rubythroat, redwing, fieldfare (a medium-sized thrush), nutcracker, Siberian jay, and many others. Wading birds include the terek sandpiper, which frequents marshes and pools.

The rivers of North Asia are inhabited by many common freshwater fishes and by several kinds of sturgeons, including the sterlet. Lake Baikal has a peculiar animal life, including many native species of sponges, worms, and crustaceans and a native species of seal.

The animal life of the steppes differs as much from that of the taiga as that of the tundra. It includes many burrowing rodents, such as jerboas, marmots, and piping hares, and, among larger animals, large numbers of antelope. The steppes were the original home of the northern cattle (*Bos taurus*), the horse, and probably the Bactrian (two-humped) camel; it is doubtful that any of these remain as truly wild animals. Typical birds are bustards, quails, sand grouse, and the red-legged hobby. Hoopoes and rollers are common locally, and bee eaters and the common sand martin nest along riverbanks. Waterfowl inhabit the reed beds of the great rivers, as do locusts, which migrate in almost unbelievably large numbers, devastating crops.

Wild sheep and goats live in the mountains and on the plateau to the north of the Himalayas. Tibet is the home of the wild yak, which is in great danger of extermination, although the domesticated yak survives.

The eastern part of the region, comprising Northeast and eastern China, has several peculiar kinds of deer. The Siberian tiger, originally native to southeastern Siberia, Manchuria, and Korea, has spread southward through eastern China into all of Southeast Asia and northern India. The giant panda inhabits the lower mountain margin of China bordering Tibet; the lesser panda is a Himalayan animal. Associated with the wastelands of the higher Himalayas is a legendary form of higher animal life—the yeti, or "Abominable Snowman." Some species of animals are peculiar to Japan, including a monkey related to the tailless Barbary ape of Gibraltar.

The large rivers of China have a rich fish life, among which *Psephurus gladius* (paddlefish) from the Yangtze and Huang Ho is of interest, as it is one of the two survivors of an otherwise extinct family, the other remnant of which is the paddlefish of North America. Another freshwater animal is the giant salamander, found in Japanese waters. Southeast Asia and southern China are the home of most of the carp family, from which the various forms of goldfish are derived.

The animal life of Asia Minor is much like that of other Mediterranean countries, but that of Israel, Syria, and Arabia also includes an African element, such as a species of coney and—in Lake Tiberias and the Dead Sea—fishes of the African genus *Tilapia* (the Nile perch). The donkey may have been domesticated in Southwest Asia, and the dromedary (one-humped) camel was originally native to the drier portions of Transcaspia.

The Oriental Region. The greater part of the Oriental Region is tropical. The northwestern part is dry and partly desert, so that animal life is chiefly confined to the forms related to those of the dry parts of the Ethiopian and Palearctic regions. Elsewhere, monkeys are common. Apes are found only in tropical rain forests—being represented by gibbons in Assam, Burma, the southeastern peninsula, and the Greater Sunda Islands—whereas the orangutan is restricted to Sumatra and Borneo, where it is in danger of extermination.

Large game animals

The Asiatic distribution of the African lion is now confined to the Gir Forest of the Kāthiāwār Peninsula in India, where it is protected, but a few specimens may still occur in southeast Iran. The tiger is now found from the Himalayas to Sumatra, Java, and Bali, but not in Borneo or Sri Lanka. Panthers range all over the region, except in Sumatra. Civets and mongooses are numerous. Among badgers, the ratel lives in the hilly districts of peninsular India and is even to be seen as far west as Israel. Jackals are plentiful in India; the striped hyena is confined to drier parts. Both are absent from the east.

Flying and ordinary squirrels are common in woodlands; the gaur (a large wild ox) is found in India and Burma, the banteng (the Malayan wild ox) in Burma and south to Borneo and Java, but not in Sumatra.

The most common antelope is the black buck, found in open brush-covered wild areas and cultivated plains all over India, except on the Malabār Coast; the nilgai, or blue bull, and the chousingha (a four-horned antelope of northern India) occupy hilly regions south of the Himalayas. Species of deer include musk deer in the pine zone of Kashmir, Nepal, and Sikkim; sambar deer practically over the whole region; and barking deer ranging northward into southernmost China.

Chevrotains (very small, hornless, deerlike ruminants) are typical, and wild pigs are also widely distributed. The Indian one-horned rhinoceros is protected and confined to Nepal and Assam; the Javanese rhinoceros is now restricted to Malaya, southern Sumatra, and western Java; the two-horned rhinoceros ranges from Burma to Sumatra. The Indian tapir lives in dense forests in southern Tenasserim, Malaya, and Sumatra. The Indian elephant is found throughout the region. Scaly anteaters, or pangolins—also found in Africa—are characteristic. The

tropical cattle (*Bos indicus*), known as Zeku or Brahman cattle and recognizable by its shoulder humps, was domesticated in India, as was the water buffalo, which is now distributed from Egypt to central China and the Philippines.

Birds

Game birds are important. The Indian peacock is to be seen throughout India, whereas another species of peacock (*Pavo muticus*) is restricted to Java. Numerous species of pheasants live in the forests of Burma, Thailand, Indochina, Malaya, Sumatra, and Borneo. Jungle fowl are unique to the Oriental Region and are the source of all domesticated chickens. Pigeons occur in great variety, but the number of species of parrots is small as compared with that of other tropical regions. Water and wood kingfishers are represented by many species. Hornbills show their greatest development in the Oriental Region. The Indian hoopoe is common in India but is only a migratory bird in the southeastern part of the region. Among cuckoos the brain-fever bird —an Asian hawk cuckoo that takes its name from the suggested effect of its repetitious cry—is well-known. Eagles, osprey, falcons, hawks, kites, and buzzards all occur; in the western part vultures are numerous and are found even in towns. The forests are inhabited by many species of woodpeckers. The barbets (loud-voiced tropical birds) are characteristic, the best known being the coppersmith bird. Bee eaters and rollers are common in India, but whereas the former can be found as far as the Malay Archipelago and beyond, rollers are absent in the southeast except in Celebes and beyond. The passerine birds are very numerous. The house crow, the Indian grackle, and the common mynah are familiar birds in India. Drongos (Old World passerines, usually black with hooked bills), flycatchers, bulbuls, tailorbirds, orioles, and many others are widely distributed, and broadbills are typical birds. Among the herons the white cattle egret is common all over the region, whereas spoonbills, cranes, and gulls are almost confined to the western part.

Of the crocodiles, the gavial (which has long slender jaws and a soft, inflatable nose tip) is restricted to the large rivers of northern India; a species of an allied genus is found in Sumatra and Borneo; and the mugger (the common freshwater crocodile) and the estuarine crocodile have a wider distribution. Freshwater turtles and land tortoises are well represented. Lizards are numerous, and flying lizards are also typical of the region. Chameleons are chiefly African, but one species is found in peninsular India and Sri Lanka. Snakes are numerous, among them the poisonous krait, cobra, and Russell's viper. Frogs and toads are abundant.

The freshwater fish life of the Oriental Region is rich. The carp and catfish families have many native genera and species. The labyrinth fishes (so named for a labyrinthine outpocketing of the gill chamber that permits them to take oxygen from air as well as water), to which the climbing perch and the gourami belong, are characteristic of the fish life of the region, as are spiny eels.

Insects, arachnoids (scorpions, spiders, and mites), mollusks, and other invertebrates inhabit this region in great numbers. Large bird-winged butterflies, allied to the well-represented swallowtails, are typical. Almost all known families of scorpions are present. Among land shells the absence of Helicidae (a family of land snails having lungs), common in the Palearctic Region, is noteworthy. Their place is taken by other forms, such as *Hemiplecta*, and by land mollusks having horny or shelly plates on their posterior dorsal surfaces. (L.F.deB.)

IV. Natural resources

MINERAL RESOURCES

Continental immensity and geological diversity explain the mineral wealth of Asia, which includes reserves of almost every important mineral. Abundant reserves of coal, oil, natural gas, and uranium, iron, bauxite, and other ores are either being exploited or await development; much wealth also remains to be surveyed. Difficulty of access, however, sometimes constitutes a barrier to exploitation.

Mineral fuels. *Coal.* Asia has enormous reserves of coal, amounting to almost 60 percent of the world total, but they are unevenly distributed. The largest reserves are found in China and in the Asian part of the Soviet Union; Taiwan, Japan, North Korea, South Korea, North Vietnam, Indonesia, and India have smaller but economically important reserves. Burma, Thailand, South Vietnam, Malaysia, and the Philippines have only insignificant amounts of poor coal. In Southwest Asia both Turkey and Afghanistan have small economic reserves.

Coalfields

Chinese coal reserves are estimated at over 1,100,000,000,000 tons and are chiefly high-grade coals. Every province has at least one coalfield, but the largest reserves are in Shansi and Shensi in the north. Szechwan, Shantung, and the Northeast (Fu-shun, in Liaoning Province) are old producing regions with good reserves, and a coal-mining region with large reserves has been developed in central Anhwei, north of the Yangtze River. Mines in Ningsia and Kansu supply northern industrial plants, but their reserves are not clearly known. The long-known reserves in western Hopeh are now being exploited.

In the Soviet Union the known coal reserves are tremendous, but the extent and quality of Siberian deposits are not yet fully known. The reserves exceed 1,700,000,000,000 tons, of which roughly 400,000,000,000 are relatively poor in grade; there are about 150 fields being worked, but as new economic developments occur the regional mining picture shifts somewhat according to quality of coal and cost of transport. The Moscow Basin brown coals and the higher quality coals of the Donets Basin of the eastern Ukraine continue to be important in the west, and the Vorkuta field west of the Urals and south of the Arctic coast will help supply the western zone. The Ural Mountains are not rich in coal, but there are scattered small fields of lower grade coals. The Karaganda fields in Kazakhstan in the southeast have huge deposits, but the coal has proven to be high in ash; and mining there is not now expanding since newer sources of better coals exist in Western Siberia. The new Ekibastuz field, north of the main Karaganda fields, is a new producer of high-quality coal.

Well over 90 percent of the known coal supplies of the Soviet Union lie in Siberia. The Kuznetsk Basin in Southwestern Siberia was the second ranking producer in the early 1970s, after the Donets Basin, and has large reserves. The Minusinsk Basin in the central region of Western Siberia, the Kansk region to the north along the Trans-Siberian Railway, the Cheremkhovo area west of Lake Baikal, and the Bureya Basin in the southeast are the major areas of production. Many smaller deposits are worked to supply local regions, such as the small and scattered fields north of Vladivostok, on Sakhalin Island, or in the hilly valleys of southeasternmost Turkistan.

Petroleum and natural gas. At least 60 percent of the world's known oil and gas reserves are also in Asia; the proportion may prove higher with the continued exploration of Siberia and the seas of southeastern Asia. Many of the island chains bordering eastern Asia have geological formations favouring petroleum accumulation, and oil fields are in production in Sumatra, Java, and Borneo. In the early 1970s, western Asia had the largest known oil reserves, located in Iran, Iraq, and the Persian Gulf area of Arabia. Other regions in Southwest Asia have only small amounts of oil, and known petroleum reserves on the Indian subcontinent are small.

Burma is the only oil-producing area on the mainland of Southeast Asia, although offshore waters may yield production after further exploration. The Philippines has not become a producing region, and the petroleum production of Japan is rather small. Korea appears to have slight prospect of production, but China is believed to have several modest oil-producing fields in Szechwan, Kansu, Sinkiang, and the Northeast. The Tsaidam Basin in northwestern Tsinghai Province may become a producing region, but information is scant. Some oil has been produced regularly from oil shales found in the Northeast, and natural gas is exploited in Szechwan.

In the Soviet Union, Siberia is expected to rival South-

The Soviet oil and gas fields

west Asia in the production of oil and natural gas. The old producing fields lay in the southern Volga Basin and in the margins of the Caucasus Mountains; the Volga Basin is the region in which many of the newer fields were in production in the early 1970s. The flanks of the Ural Mountains have a number of large oil fields and small gas fields. The northern Volga Basin, along the western flank of the Ural Mountains, contains the leading producing regions for oil. Major gas fields are located in the northeastern Ukraine south of Kharkov and in the Carpathian foothills of the western Ukraine. Uzen, on the Mangyshlak Peninsula on the eastern shore of the Caspian Sea, is a major gas-producing field that also yields oil. Another major field is that of Gazli in the Kyzyl Kum Desert south of the Aral Sea, and the rich gas field in the northern Ob River Basin at Berezovo indicates that the whole Ob Basin may yield natural gas. In the Lena River Basin, north of Yakutsk, there are large proven gas reserves.

Uranium. Reserves of uranium ore are found in Asia's ancient crystalline rocks. The Soviet Union, China, and India all have their own supplies of uranium. The Soviet Union, in particular, has rich ore fields in Kirgizia, between Osh and Tuya Muyun. Chinese uranium resources are probably located in northern Sinkiang and southern Hunan.

Metallic ores. *Iron.* All portions of Asia have deposits of iron ore, although not every political state has its own private supply. South Korea, South Vietnam, Taiwan, Sri Lanka, and several smaller countries in Southwest Asia appear to have only small iron-ore supplies. Japan has far less than needed by its large iron and steel industry and depends largely on imported supplies. The Philippines has much more ore than needed by its small industrial needs and is an ore exporter. Malaysia also exports considerable volume. Thailand, Burma, and Pakistan have fair amounts of relatively low-grade ores, and North Vietnam and Turkey have good ores in substantial volume. Indonesia and India both have large deposits of good iron ores that are reasonably distributed.

Although formerly regarded as deficient in iron ores, China contains huge quantities of varying grades of ores that are widely distributed and often located close to coal supplies. Regional centres of ore mining, smelting, and fabrication are located at An-shan, southern area of the Northeast; near Peking; in southern Anhwei, west of Shanghai; in central China, east of Wu-han; in southern Inner Mongolia, north of Pao-t'ou; in central western Kansu; and on Hainan Tao (Hainan Island), off the south coast. Large iron ore deposits also occur near Chungking in Szechwan. Iron ore in small local volumes is widely located in Kweichow and Yunnan in the Southwest.

The Soviet Union formerly depended largely on iron ore from the Krivoy Rog and Kerch basins of the southern Ukraine, but huge deposits of magnetic ore have been found near Belgorod. Iron ore has long been extracted from the Ural Mountains, and further deposits have been found at and near Magnitogorsk in the Southern Urals. There appears to be an unlimited supply of low-grade ore in the Kustanay Basin east of the Southern Urals in southwestern Siberia. A major iron ore range, at Kachkanar, west of the Northern Urals, contains low-grade ores. Large deposits of medium-grade iron ore have been found northwest of Lake Baikal, close to the Cherenkhovo coal deposits. Smaller deposits have been located in the Murmansk Peninsula and at several locations in Eastern Siberia.

Ferroalloy metals. Asian resources of nickel are not extensive. There is a notable Soviet nickel ore field at Norilsk (in Northern Siberia); Indonesia also possesses reserves. Those in Burma have not been fully surveyed.

Asian countries with reserves of chromium include Turkey, the Philippines, India, Iran, Pakistan, and Cyprus; reserves are also found in Soviet Asia. Chinese reserves are unknown.

Manganese is found in abundance. The Soviet Union has large reserves in Central Asia and in Siberia, and India also possesses large quantities. Chinese reserves are considerable, but those of Japan are limited.

China has exceptionally large reserves of tungsten, especially in southern China. Tungsten reserves of Soviet Asia are also important, as are those of molybdenum.

Copper, tin, and zinc

Nonferrous base metals. Asia is not richly endowed with copper ore. In Soviet Asia the principal fields are Almalyk (southeast of Tashkent), Dzhezkazgan (west of Karaganda), Kounrad (Lake Balkhash), and in the Kuznetsk Basin. Japan's widespread copper ore reserves are of medium importance, and the Philippines have limited reserves. The extent of China's reserves is not known, but deposits are located in Kansu, Hopei, Anhwei, and Hupeh. Turkey, Israel, India, and North and South Korea have small reserves.

Significant reserves of tin exist along a north–south axis running from southwestern China through the Malay Peninsula to Indonesia. Thailand, Burma, North Vietnam, Laos, and Yunnan Province in China also have deposits of tin. Soviet Asia has substantial reserves in Transbaikal and also in the Sikhote-Alin Range.

Soviet Asia's lead and zinc reserves—the largest in Asia—are located in the Kuznetsk Basin and in central and eastern Kazakhstan. China also has abundant reserves of zinc and lead ores, and North Korea has important lead resources.

Asia has enormous reserves of bauxite. In Soviet Asia, bauxite fields are located in Kazakhstan and in the Sayan Mountains. There are also large reserves in India, Indonesia, the Philippines, and Malaysia, as well as significant reserves in China.

Important quantities of mercury occur in south central China and in the Soviet regions of the Ukraine and Siberia. Magnesite is common in Asia. There are large reserves of antimony in central China; Turkey and Thailand also have substantial reserves.

Gold and silver

Precious metals. Many Asian countries have produced gold from alluvial stream deposits in past centuries, and some continue to do so. Small volumes of alluvial gold are produced in Burma, Cambodia, and Indonesia, and the headwaters of the Yangtze River in the Tibetan border region yield some gold. India formerly was a large producer of gold from lode mines, but the best ores appear to have been exhausted. Japan, North and South Korea, Taiwan, and the Philippines have significant gold-ore reserves, and Malaysia periodically produces gold from small lode mines.

The Soviet Union has produced gold from lode mines in the Central Ural Mountains for centuries, and in the 19th century there were several gold rushes to work alluvial stream deposits in Siberia on the Lena and Yenisey rivers. In the early 1970s Soviet gold production was rising, and lodes were worked in several Siberian locations, centring on the upper reaches of the Kolmya River and in the hill and mountain country surrounding Yakutsk. The lode opened up at Auezov in eastern Kazakhstan, south of Semipalatinsk, may allow the Soviet Union to threaten South Africa in rank as a gold producer.

Platinum is mined near Norilsk in the Central Siberian Plateau in Northern Siberia. Silver is not abundant.

Nonmetallic deposits. Reserves of asbestos are very localized; it is abundant in China, South Korea, and in Soviet Asia. Mica is abundant in Soviet Asia and is also found in important quantities in India. Asia has abundant reserves of rock salt; the hills and "glaciers" of salt in southern Iran are unexploitable, however, under present conditions. Reserves of sulfur and gypsum are abundant in Central and West Asia. Japan has large reserves of sulfur. In North Vietnam phosphates are obtainable from sedimentary fields and from apatite. The Soviet Union has large deposits of phosphates in the Mangyshlak Peninsula on the eastern shore of the Caspian Sea and other scattered deposits of lesser value. Diamonds are insignificant in India but are produced in east central Siberia.

WATER RESOURCES

Asia's water resources constitute a vast potential, both for generating hydroelectricity and for irrigation. In the arid parts of the continent, water is primarily useful for irrigation.

Siberian hydroelectric power

Siberian rivers have an excellent hydroelectric potential, for when dammed they provide low falls with an enormous volume of flow. Extreme cold and low winter water levels, however, hinder their exploitation. Thanks to abundant precipitation and great differences in water level, the Soviet Far East has an immense potential for hydroelectricity, although the remoteness of Eastern Siberia discourages industrialization.

Japan, a country of high man-made waterfalls but relatively small volumes of water flow, has already harnessed almost all its rivers that have a hydroelectric potential; this potential, however, is increased by heavy rains, particularly in summer.

The waterpower potential of northern China is extremely limited because the flow of the Huang Ho and other northern rivers is erratic, and all carry heavy volumes of silt. The hydroelectric potential of China south of the Tsinling Mountains, however, is great. The Yangtze River has a considerable waterpower potential, particularly near I-ch'ang, although this site would be expensive to develop.

The hydroelectric potential in the Indian subcontinent is subject to regional variations. The Western Ghāts, which slope down abruptly to the western maritime plains, permit high waterfalls; unfortunately, the rivulets that rise on the summit have an insignificant volume of winter flow. Rivers of the eastern slope of the Deccan, such as the Mahānadi and the Godāvari, lend themselves to the construction of dams with low falls and great volumes of flow, as also do the Himalayan rivers entering the Ganges Plain. The Himalayan ranges offer rich possibilities for the utilization of high waterfalls for generating hydroelectricity, but in winter their waters are very low, and most of the sites would be expensive to develop.

BIOLOGICAL RESOURCES

Widely varying climatic conditions, particularly in the distribution of rainfall, have produced terrains in Asia ranging from tundra and desert to forest and alluvial plain, each supporting its appropriate plant cover, on which, in turn, typical animals and birds subsist. The Arctic north of the continent and large areas of the central mountain massif—known as "the roof of the world"—are practically uninhabitable. In addition, even where there is water—and nowhere is water conservation pursued more carefully than in Asia—there are still many areas of undrained swamp. Much else is desert. By far the greater part of Asia remains uncultivated. Prime resources are the extraordinarily intensive agriculture made possible by irrigation of the alluvial soils of the great river deltas and courses; the forests, with commercially valuable species of trees; the flocks of sheep and goats supported by Asia's semi-arid deserts and grasslands; and the produce of the intensively fished surrounding seas of South and East Asia.

Timber resources. Much of Northern Siberia, south of the Arctic Circle, is covered by coniferous and mixed forest, which is commercially exploitable. The great deciduous forests of northeast India, Burma, Thailand, and Malaysia contain teak and other important hardwoods, as well as bamboo. Mangrove forests line the waters of the Ganges and Irrawaddy deltas and many small stretches of coast along the Malay Peninsula, Indonesia, and the Philippines. But in the Indian subcontinent lowland, forest has yielded place to cultivated land, as a result of population expansion; agriculture has similarly reduced the natural forest areas of China to insignificance, except in northern Manchuria. Japan, on the other hand, is relatively heavily forested in relation to its area and population, although much of the present cover is planted forest. More than half of the Philippines still carries heavy forest, but good commercial forests cover only one-third of the country. These forests produce valuable hardwoods and the soft "Philippine mahogany."

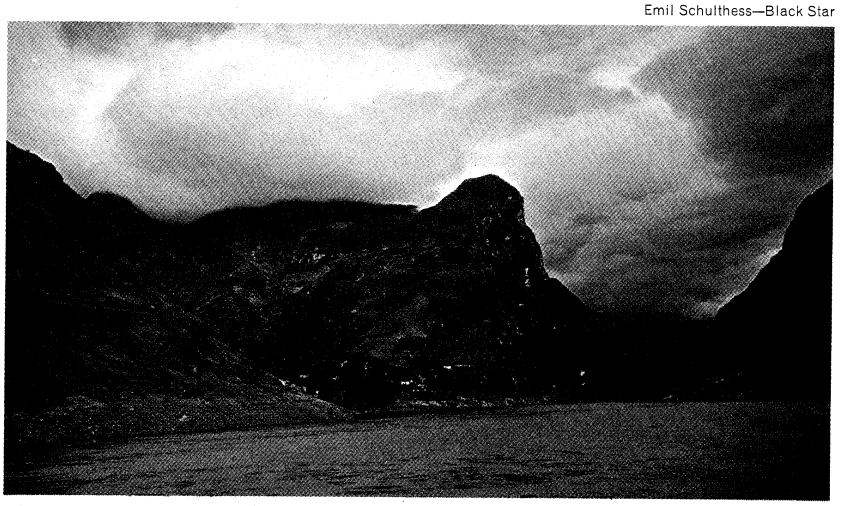
Resources for agriculture and animal husbandry. *Crops.* In Soviet Asia, the black-earth belt across Southern Siberia is cultivated with grain crops, of which wheat is the most important, as also are areas of the Soviet Central Asian republics. Grain crops, chiefly wheat, are cultivated in North China—where soybeans are also grown—and in Japan. Intensive use of water resources from wells, as well as from irrigated rivers, has enabled grain crops to be raised in Iraq, Iran, Pakistan, and northern India. The great staple of South Asia is rice. It is the chief food crop of Japan, South China, Taiwan, Southeast Asia, the Indonesian islands, the Philippines, Burma, Sri Lanka, and parts of India and Pakistan, and is found in Iran, southwestern Asia, and elsewhere.

Of plantation crops, rubber, from a Brazilian plant imported in the 19th century, is cultivated in Malaysia and Indonesia and also in India and Sri Lanka. Tea is grown on commercial plantations in the uplands of northern India and Sri Lanka for export and in China and Soviet Asia on small holdings for domestic consumption. Sugarcane is harvested in Java, the Philippines, India, and Central Asia; and tobacco is grown widely, notably in Turkey, Soviet Asia, China, and Indonesia. Citrus fruit is produced in the Mediterranean lands, in the Soviet Central Asian republics, and in China and Japan. Date palms are cultivated, particularly in Arabia. Licorice is grown in Turkey. Asia is also a producer of opium from the poppy.

Livestock. The uncultivated steppe lands and deserts of Central Asia and Mongolia support flocks of sheep and goats. Semi-nomadic pastoralism is the rule there, as it is in parts of Afghanistan, Pakistan, Iran, and Arabia. In Central Asia, the horse and the yak are the riding animal and the beast of burden, respectively; in Arabia, the camel is both. Cattle are raised in agricultural areas. Hides, wool, and other animal products are important

Semi-nomadic pastoralism

Emil Schulthess—Black Star



Yangtze Gorges at the Szechwan–Hupeh mountain border below Wan-hsien in Szechwan Province, China.

economically. Reindeer herds are kept in the northern tundra of Siberia, where they feed on mosses and shrubs. In Siberia, valuable furbearing animals have long been hunted. In India, Burma, and Thailand elephants still work as draft animals in the lumbering industry; particularly in Southeast Asia, the water buffalo is an important draft animal as well as a milk and butter producer. Angora goats are herded in Anatolian Turkey to provide the silky mohair for which they are noted. Silkworms are reared for silk in China, Japan, India, and Soviet Central Asia.

Game birds. North of the Himalayas, such game birds as ptarmigan, grouse, plover, and various kinds of waterfowl are found. South of the Himalayas, pigeons, pheasants, and other game birds are taken. Various kinds of hawk and falcon, trained to hunt, have their habitat in Arabia and other parts of Asia.

Seafood. In addition to fish and other sea creatures, various kinds of crab and shrimp are intensively fished off the coasts of China, Japan, and Southeast Asia. The sturgeon, prized for caviar, is fished commercially, particularly in the Caspian Sea and the rivers of Siberia.

(P.Gu.)

V. Human resources

A discussion of Asia and its peoples cannot entirely exclude other parts of the Old World when the origins of man, his ethnic divergence, and his migrational wanderings, or the evolution and historic development of his linguistic and culture systems are considered. The relatively modern division of the largest of the continents into Europe and Asia derives from arbitrary decisions made by peoples living in the western part of the peninsula of Europe; this division has only minor significance in relation to the historic patterns of human occupation of the continent. The ethnic and linguistic diversity of Asia is greater than that of any other continent, because it represents ethnic types and linguistic systems that have evolved in separated regional homelands, as well as repeated patterns of modification and intermixture, resulting from both peaceful and militant migrations. Ethnically and linguistically, some territories have become highly diversified mosaics in which there are mixed and overlapping elements.

EVOLUTION OF THE ETHNIC PATTERN

The three major racial stocks

Original racial stocks. The peoples of Asia include all the three major racial stocks of *Homo sapiens*—Congoid (Negroid), Caucasoid, and Mongoloid. The view is here taken that Congoid racial stocks derived from tropical origins in the zone extending from Africa eastward through the Indonesian Archipelago. It would, however, appear that to the east of Africa their original numbers were insufficient to ensure that they become the predominant element among the populations of South and Southeast Asia. The contrary belief that Congoid groups east of Africa originated in Africa leads to the conclusion that the migrating Congoids scattered out very thinly to the eastward and were able to contribute only to the skin pigmentation of South India and part of Southeast Asia. Whatever their origin, Congoid racial stocks at no time seem to have penetrated deeply into the Asian mainland. At the present time they constitute an element in ethnic stocks only in the southern peninsular fringes of Asia. The view here taken is that the Caucasoid racial stocks derive from a western Eurasian racial hearth, or region of origin, that also includes North Africa, and that Mongoloid racial stocks derive from an eastern Eurasian racial hearth. A very simplified view of the history of man in Eurasia, then, is that the western zone of the continent came to be populated primarily by the Caucasoid, or white, ethnic groups; that the eastern zone was populated primarily by the Mongoloid, or yellow, ethnic groups; and that the southern fringes of the continent were lightly occupied in early time by the Congoid, or black, ethnic groups. This pattern was apparently derived from the evolutionary beginnings of *Homo sapiens* possibly about 40,000 years ago, out of predecessor racial stocks, since the basic differentiation of the human species predates both modern man and the last glacial era. In considering modern man in Asia we cannot speak of distinct races in any definite way, as repeated migrational movements since the last glacial era have resulted in such intermixed racial stocks that the modern Asian cannot be racially defined. While it is possible to give a general description of an Indian, a Chinese, or a European, this can be done only in terms that connote ultimate regional origin in the broadest sense. In discussing Asian ethnology, therefore, we cannot refer to races but only to ethnic groups that speak particular languages and that have particular cultural characteristics. Thus, a Hunan Chinese from central China may be compared to a Czech from central Europe, or a Japanese from the southern island of Kyushu may be compared to an Iranian from southern Iran. All four belong to the single species but show marked differences in physique, language, and culture. It is, nevertheless, convenient to stereotype Chinese, Japanese, Czechs, and Iranians as members of "geographic races," meaning members of normally separated breeding populations.

Prehistoric migrations

Ancient migratory movements. The two primary prehistoric centres from which migrations over the continent took place were Southwest Asia and a region comprising the Mongolian plateaus and North China.

From prehistoric to historic times, possibly beginning as early as 30,000 years ago, movements from Southwest Asia continued toward Europe and into Central Asia; significant movements also took place into India. There were probably small divergent migrational movements in other directions that became swallowed up in later patterns of mixing. The Greeks were one of the late groups moving westward, about 2100 to 1900 BC, as were the Aryans who moved east to invade India from 1600 to 1500 BC. Mongoloid migrational movements have always been primarily toward Southeast Asia.

Important Mongoloid components, however, also moved westward through Central Asia toward the European peninsula. Such movements must have begun as early as 10,000 years ago, but they continued into the Christian Era as Mongols pushed Turkic peoples westward, setting off additional displacements of such peoples as the Finns and the Magyars. These westward Mongoloid movements also produced, over a period of time, much mixing of early Caucasoid and Mongoloid stocks in Central and West Asia. Northern Eurasia continued to be inhabited chiefly by thinly distributed residual elements of very early eastern Asian stocks, although some fairly late northward movements of Turkic peoples did take place.

There have been many small-stream movements away from the main trends, and these have often complicated the ethnic picture of any one region. At least one prehistoric Caucasoid movement penetrated East Asia and today is represented by the historic aboriginal population of Japan known as the Ainu. A countermovement out of India by a nomadic ethnic stock about AD 1000 contributed the Gypsy strain now so widespread in Europe.

Prehistoric countermovements along the China coast carried early Mongoloid migrants of Southeast Asia northward again into southern Korea and Japan, to leaven the later Mongoloid and Ainu stocks, from all three of which modern Japanese are derived. Similar northward drifts of early Mongoloid Indonesians account for a significant share of the ethnic ancestry of the population of the Philippines. Within the broad zone of Central Asia, prehistoric and historic movements have often retraced older migratory routes, creating overlapping and fragmented distributions of stocks that have yielded the many ethnic groups found there today. Secondary and tertiary intermixing of many of these regionally derived ethnic groupings has resulted in the emergence of regionally and physically distinct ethnic groups. Thus, the Uzbeks may originally have derived from a Mongoloid stock; some of them migrated westward to near the Volga River at an early date, then moved southward to become intermixed with Caucasoid-derived stocks. Uzbeks are now widely distributed in Central Asia and show considerable physical and linguistic variation.

Modern movements of peoples. Within historic time the aggressive expansion of particular ethnic groups has

The mixture of ethnic groups

either driven weaker groups away from their territory or has resulted in the newcomers' assuming control of the territory and reducing the older inhabitants to the status of ethnic minorities. Some of these weaker ethnic stocks eventually became so diluted by intermixture as virtually to lose their identity. In some instances a new and variant ethnic stock with a different dialect resulted from the mixing. Some areas are now given over to distinct enclaves occupied by several diverse ethnic stocks, each following its own way of life. Thus, in Southeast Asia, from the riverine and coastal lowlands to the higher mountain uplands, different ethnic stocks have been migrating southward to become resident in separate altitudinal layers, one above the other. Some of this migrational movement in Southeast Asia is as late as the 19th century, but it has been going on for thousands of years. Within what are now India and Pakistan the general trend, for many centuries, has been eastward and southward, producing very discontinuous patterns. Discontinuity also characterizes the ethnic patterns in Central and Southwest Asia.

Militant campaigns of Arabs spread Islām and Arab political structures out of Arabia westward into Africa and Spain and eastward through the Levant into Asia Minor, Transcaspia, and India. Beginning in the 7th century AD and lasting until the 16th century, these efforts spread Arab ethnic elements widely in Southwest Asia and northern Africa.

Within recent times, movements of Caucasoid European Russians eastward along the Central Asian routes of exploration, and the penetration of the oceanic fringes of South and East Asia by Caucasoid western Europeans, have carried Caucasoids to all parts of the Eurasian continent. This has resulted in an interbreeding that has produced many local and variant mixtures. Since the 17th century, intermarriage between Europeans and indigenous Asians has produced many mixtures, including the Anglo-Indians of India and the Burghers of Sri Lanka. Intermarriage between Chinese men and local women has produced many hybrid strains in Indonesia, Malaysia, Thailand, and the Philippines. The introduction of American white and black soldiers to East and Southeast Asia, during and after World War II, has further complicated the ethnic mosaic in China, Korea, Japan, Vietnam, and the Philippines. Such modern racial mixings are often viewed apart from the historic patterns of migration and racial mixing, but essentially they form a part of them.

POPULATION DISTRIBUTION AND REGIONAL ECOLOGY

Life-styles in the 18th century

The background. Around 1750 the distribution of ethnic groups was relatively easy to describe. The whole of northern Eurasia was rather lightly populated by diverse ethnic groups of Paleo-Asiatic, Tungusic, and Turkic peoples who engaged in hunting, collecting, fishing, or herding; some groups, such as the Samoyed, Yakut, and Chukchi, had somewhat distinctive single economies or had economies that were seasonally mixed. Central Asia, Tibet, and Mongolia formed a mixed zone dominated by nomadic pastoralism, but the lower plateaus and lowlands were sprinkled with agricultural oases in which towns and villages were occupied by sedentary crop growers. Population was relatively light; mountain regions were occupied only in summer, but there were locally dense populations centred on such large oases as Tashkent, Samarkand, Kashgar, and Urumchi, with smaller groupings around lesser sources of water. The Buriat Mongols and the Kirgiz were pastoral, whereas the Tadzhiks, Uighurs, and Uzbeks were sedentary oasis dwellers. Southwest Asia was then inhabited by Iranian, Arab, and Turkish peoples, with a scattering of minority ethnic stocks, practicing either traditional pastoralism or the agricultural economy of the oasis. Population was concentrated around cultivable areas, water resources, or grass pastures.

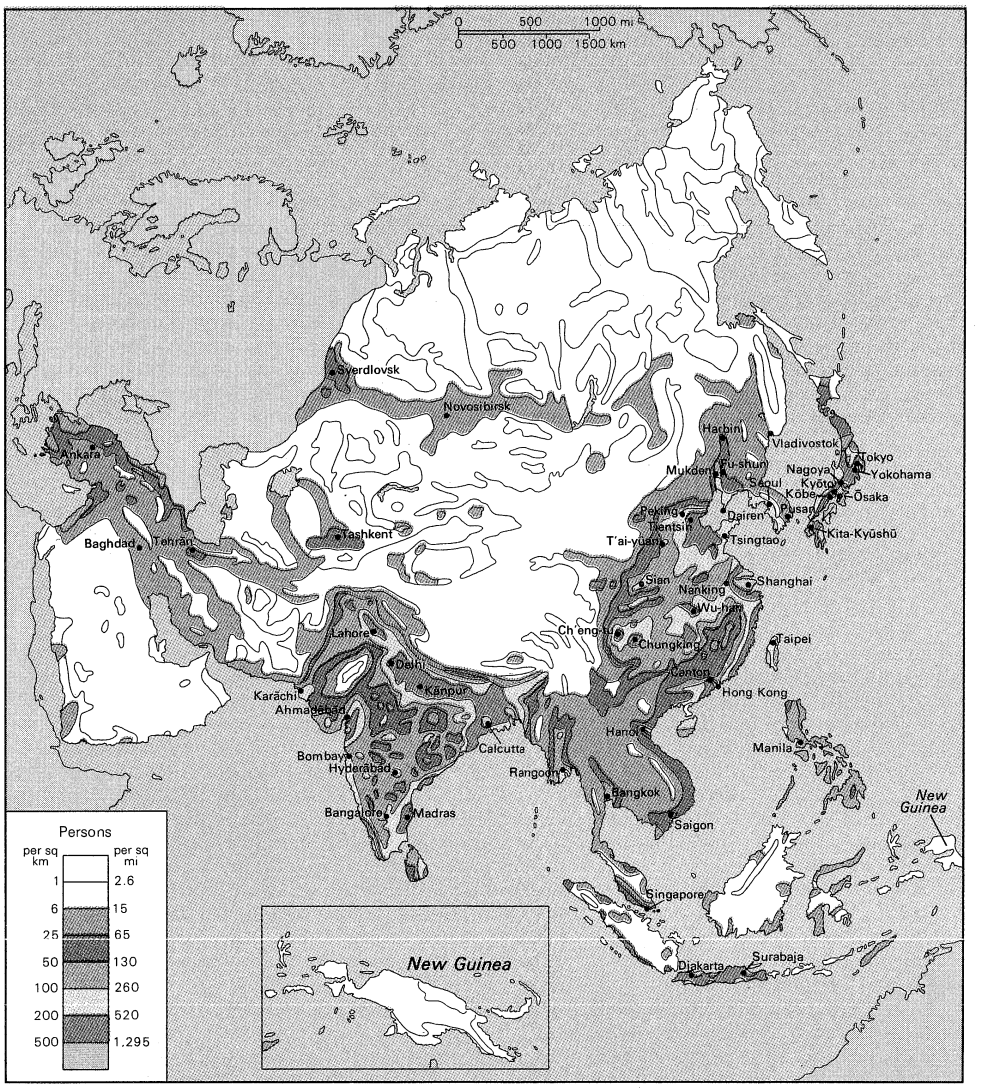
South and East Asia showed a more complex dual set of patterns. The largest components consisted of the highly civilized lowland populations, long settled on their land and engaged in sedentary agriculture and handicraft manufacturing. Market towns and cities were scattered over the countryside, and many small port towns dotted the seacoasts. Population density was heaviest in the best agricultural lowlands, which had also been occupied the longest, such as the North China Plain, southern Japan, coastal Vietnam, and the Ganges Valley.

Lesser components included the many diverse ethnic groups scattered in wet deltaic lowlands such as those of the Ganges, Irrawaddy, Chao Phraya, and Mekong rivers; the central plain of Luzon Island in the Philippines; and the north coast of Sumatra. Groups were also scattered throughout most of the hill and lower mountain country. Their living systems varied from simple hunting and collecting economies to more complex systems in which, apart from shifting cultivation (the shifting of cultivation to new land after a crop has been raised), hunting and collecting are also practiced. Generally these lesser areas had small populations scattered in hamlets or village settlements sustained by subsistence economies; only a little handicraft manufacturing took place, and trade was confined to minor products. The Nāga of northeast India, the upland Karen of Burma, and the Miao of Laos exemplified this life-style. Toward the end of the 18th century, European colonial efforts were beginning to shape the production systems of eastern Eurasia to conform to patterns of integrated world trade. The supplying of Europe with raw materials, which was to characterize the early 20th century, was also begun at this time.

The pattern of ethnic distribution. By the early 1970s, great changes had taken place in both the ethnic patterns and the associated life-styles in Asia. The divisions that prevailed in the 18th century were dissolving as the Soviet Union and China extended their economic and political control over Siberia and Central Asia, as the former colonial lands of South Asia established political independence, and as some of the component territories of the old Ottoman Empire were reshaped into the modern nations of Southwest Asia. Many of the hundreds of small ethnic groups were being absorbed into the populations of nation-states, many old languages were declining, and many formerly distinctive living systems were remaining in existence only as remnants or artificially preserved societies. The expansion of dominant ethnic groups was steadily restricting the territory available for older, simpler societies; and aspects of modern economic development were replacing the earlier systems. It is possible, still, to identify the region in which the Yukaghir formerly lived as a separate culture group in Eastern Siberia, but—for the few hundred Yukaghir who remain—political absorption, modernizing acculturation, and internal social decay have made the classic description of the group largely a historic one. Many former horse-riding, tent-dwelling, sheep-herding Kara Kirgiz today ride tractors on Soviet state grain farms, live in permanent villages, and speak Russian in public. Some men of the Chotanāgpur hill region of eastern India, who formerly engaged in hunting and practiced shifting cultivation, today work in the steel mills of Jamshedpur. The remnant Ainu of northern Japan today are gathered into "cultural villages" where their continued woodcarving and bear dances attract a flow of tourists from southern Japan.

Social changes of the 20th century

Contemporary developments. Population densities have everywhere increased, and planned programs of modernized agriculture, mineral exploitation, and industrialization are bringing cultural change. Some of the small ethnic groups are dying out, but larger groups have often accepted change and are again increasing in numbers. In South and East Asia, increasing lowland populations are pressing hard upon the available land as population densities exceed 2,000 persons per square mile. In Central Asia, both Chinese and Russian settlement programs are moving peoples from heavily populated regions into frontier zones, in order to develop both agricultural and industrial resources. In Southern Siberia the Soviet settlement program has spread a thick wedge of European Russians and assorted ethnic minorities eastward to the Pacific and northward along every river valley to the Arctic Ocean. As a result, the Paleosiberian ethnic remnants are being submerged and absorbed. Old trading posts, oasis towns, and the few old cities of South-



Population density of Asia.

Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.

ern Siberia and Soviet Central Asia are being developed into modern industrial centres; these are linked to modern transport systems by which raw materials and manufactured products flow back to the European regions. Most new cities are populated largely by European Russians, with the former Asian ethnic stocks remaining chiefly in the rural areas. The modernization of Southwestern Asia, through the renascence of Turkey and the impact of petroleum exploitation on the Arabian Peninsula, has altered many of the old patterns of ethnic groupings in these areas. A further alteration of the historic pattern came in 1948 with the creation of the State of Israel, to which over 1,000,000 Jews from Europe, North America, and the Middle East had migrated by the early 1970s.

Urbanism

Urbanism is becoming very marked in some parts of Asia, thus heightening regional contrasts in population densities. Japan, with about 70 percent of the population living in urban areas, is the most highly urbanized country in eastern Asia. Israel, in western Asia, is more than 80 percent urban. Elsewhere, the urban population varies from 20 to 35 percent of the total, although this percentage is rising as industrialization proceeds. Some cities are becoming among the largest in the world, and many of the newer cities resemble those of the West in terms of population, buildings, facilities, and congestion.

Ecological factors. Agriculture remains the mainstay of the Asian population, and more than 70 percent of Chinese, Indonesians, and Indians are engaged in agriculture and animal husbandry. Although marginal lands in many parts of southern and eastern Asia have been brought under cultivation, and many former pastoral ranges in Southwest and Central Asia are now irrigated, the broad ecologic factors touched upon above continue to give rise to zonal distinctions in population and economic activity. Parts of South and East Asia can support dense populations. Favoured localities in the southwest —for example, in Turkey and northern Iran—support large populations. In Southwest and Central Asia in general, however, agricultural productivity and population density vary markedly with the regional pattern of rainfall or the availability of water from humid highlands nearby. In the Soviet sector the older pastoral nomadism has been transformed into organized transhumance (seasonal migration of stock between lowlands and mountains); in consequence, the formerly nomadic families are now permanently resident in villages, and only herders accompany the flocks and herds. Northern Asia remains a semi-developed frontier region with short-season crop growing in favoured southern localities, even though breeding of newer varieties has extended agriculture northward. The Arctic fringe is being developed on the basis of mineral resource exploitation, but only in particular localities. Siberia remains lightly populated, with the population grouped around local centres.

The importance of water

The pattern of ethnic language distribution. A language map always tends to reflect past conditions because speech systems are constantly undergoing change.

In the past, language maps for the Eurasian continent have often shown eight languages—Turkic, Slavic, Tungusic, Chinese, Tibeto-Burman, Indo-Aryan, Iranian, and Mongol—almost blanketing the main portion of Asia and leaving other languages predominating only on peninsular appendages, island fringes, and in small pockets. Except for the large eastward expansion of the Slavic group, the map reflects distributional patterns that prevailed in the 18th century. At that time, by far the largest language groups, by number of speakers, were the Chinese and Indo-Aryan, but the Tungusic languages were probably used over wider areas. In recent times many of the small ethnic-group languages have, for practical purposes, been dying out, to be preserved only by professional linguists. The Tungusic group shows this decrease in usage, in spite of the Soviet practice of publishing books and newspapers in regional languages and the encouragement given to the preservation of the more important ethnic languages. Russian is now the dominant public language throughout the Soviet Union, and in the Asian territories it is spoken by large numbers of non-Slavic inhabitants. Similarly, Mandarin Chinese is expanding in China at the expense of local languages and dialects. In India, however, local languages are not losing ground, and language has become a territorial political issue. Meanwhile, the Indian Parliament continued, in the early 1970s, to debate in English. In Indonesia, which has many local languages and dialects, Bahasa Indonesia (the national language) has not yet spread throughout the Indonesian state. English remains the most commonly spoken single language in the Philippines, despite the adoption of a national Pilipino language.

The loss of small ethnic languages

Whereas many languages are dying out, as ethnic groups disappear or become merged into larger groups, some ethnic populations are increasing in numbers, thus increasing the relative importance of their languages. The large increase in the population of China now means that Mandarin Chinese is the world's leading language by number of speakers. The marked increases in population in Japan and the Indonesian island of Java mean that Japanese and Javanese rank much higher on the list of languages, by speakers, than they formerly did. Similarly, Western Hindi, spoken in northern India, is one of the larger languages by number of speakers. Though many of the Paleosiberian (Paleo-Asiatic) languages are dying out as their ethnic users decline in numbers, the Uzbeks and the Tadzhiks, for example, have adjusted to Russian control and are again increasing in numbers, forming significant ethnic components of the Soviet Union, with the result that their languages are being maintained. It appears that many ancient languages spoken in Asia have disappeared within the last millennium and that others have been greatly modified by linguistic change.

Regional situations have sometimes produced multiple and overlapping language patterns. Around some of the old Central Asian oases and in Southern Siberia, migrants from Russia and exiled ethnic groups have created ethnically mixed regional populations. A comparable pattern may be discerned in Chinese Central Asia. Such large cities as Manila, Singapore, and Bombay show complex linguistic patterns. As European Russians have moved into the new cities in Transcaspia and Western Siberia, Russian has become the language of the cities; the older languages have been confined chiefly to the countryside.

FORMS OF ETHNIC ADMINISTRATION

Imperial administration. The older forms of administration, by which political states controlled adjacent and frontier ethnic groups, generally used local native leaders, who normally were given honorific subordinate titles and made responsible for the orderly control of their territories. The Chinese Empire sometimes entered into treaty-like agreements with subordinate states on its periphery and either subsidized the non-Chinese states or exacted tribute as a token of subordinate or feudatory status. The British in India and Burma, the Dutch in Indonesia, and the French in Indochina developed systems of frontier agencies that employed resident officials to supervise local leaders, who exercised autonomy over what amounted to ethnolinguistic groups. The Thai maintained their control of Siam (now Thailand) as a buffer state between the French and British regions but in their northern area maintained the older form of control through native leaders. Beyond the reach of these larger imperial states simpler ethnic groups maintained their local sovereignty under the rule of chiefs, shamans, or clan leaders, sometimes forming limited confederations.

Multi-ethnic states. The development of modern forms of political administration among Asian states has produced some distinctive regional patterns. The Soviet Union was the first state to organize administrative districts on an ethnolinguistic basis. There are about 100 separate ethnic groups publicly recognized in the Soviet Union, as well as some minority groups never identified, and about 60 of these are represented by political administrative territories at major or minor levels. China under the Communist regime has adapted this system and has modified the Imperial political structure in regions containing ethnic or linguistic minorities—primarily in South and Southwest China, Northwest China, and Central Asia. In the Soviet Union such ethnic territorialism is relatively fixed and stable, but in China there continue to be changes in spatial arrangements of autonomous regions as various pressures are exerted, for not all minorities have yet been given internal territorial autonomy.

Ethnolinguistic political units

In India, with several hundred languages and many varieties of ethnic groupings, ethnolinguistic recognition is made only at the state level. Several of the political states of the Indian Union are now bounded by linguistic limits. Many minorities are not recognized at present, and the question of spatial ethnic and linguistic autonomy has given rise to considerable unrest within the Indian Union. The former northeastern Nāga tribal agency, however, has become a full state on the basis of its cultural unity. In Pakistan the tribal and frontier agencies formed during British Indian rule are still preserved; in these agencies, spatial autonomy derives from the ethnolinguistic situation. Burma has still not resolved the problems of integrating ethnic minorities into a modern political structure, and in the early 1970s several upland ethnic minority groups were expressing militant opposition to the forms of limited territorial autonomy offered by the government. Throughout the rest of Southeast Asia, except for Malaysia, ethnic minorities have received virtually no formal recognition, and each country is adopting different means of integrating its minorities.

Malaysia is a multi-ethnic state in which Malays total just less than half the total population; Chinese total just over one-third; and Indians, Pakistanis, and tribal groups almost equally split the remainder. The constitution makes no recognition of the plural ethnic composition; Malay is the official language; Islām is a state religion (although religious freedom is guaranteed); and the head of state must be a Malay. Quasi-legal political parties, however, represent ethnic groupings, and there are—in practice—many ways in which all ethnic elements are represented.

In Southwest Asia, minor populations exist in most political states without formal recognition of their status, the minority position deriving from ethnic, linguistic, or religious factors. Only in Saudi Arabia and Yemen (Aden) is there homogeneity in the three elements. Lebanon is ethnically and linguistically Arabic, but its population is almost equally divided between Christians and Muslims. Israel has a sizable Arab minority, and Iran is only half Persian in ethnic and linguistic terms. Most other states in Southwest Asia have comparable ethnic conditions.

VI. Political geography

HISTORICAL DEVELOPMENT

The Pre-European era. The first sophisticated organization of space, people, and cultural systems is customarily attributed to ancient Southwest Asia. In the earliest times there emerged the concept of a normal political state ruled by a god-king to whom all resources belonged and who held total power over the human population. Whereas some early states were city-states, the long-term

Ancient state systems

trend was toward the spatial state that included rural hinterlands. States rose and fell in the ancient Orient as the conceptual system spread both westward into the Mediterranean Basin, as well as into the peninsula of Europe, and eastward into India and China. Historically, Indian political systems dominated South and Southeast Asia, Chinese systems controlled East Asia, and variations on the original model continued in Southwest Asia, while steadily changing variations developed new patterns in the Mediterranean Basin and the European peninsula.

At the end of the 15th century, at the time of Vasco da Gama's voyage to India (which signalled the dawn of European influence in Asia), the situation on the continent was approximately as follows: East Asian systems were relatively stable; the political situation in Southeast Asia was in a state of flux, as several states were in decay, and Islāmic political missionaries from Arabia had not yet succeeded in consolidating their influence; South Asia was similarly in a state of flux as the Mughal Empire struggled to achieve spatial hegemony over the Indian subcontinent; Southwest Asian political systems were experiencing a period of readjustment as the Turkic peoples penetrated the region and began to assume control; Central Asia was experiencing the last phase of the expansion of the Mongols, as well as the spread of Islām; no formal political states had so far evolved in the Siberian zone.

The evolution of European contact. The evolution of Europe's political relationship with Asia may be conveniently divided into two phases. The earlier phase that characterized the modern political geography of Asia was one in which some of the states of the European peninsula were able to introduce political controls into the unstable parts of the remainder of the Eurasian continent. The second, and more recent, phase has been characterized by the withdrawal of these controls from the whole of the southern zone, despite the fact that varying economic and political links remain operative. In the northern zone, however, where at an earlier stage no political states were in existence, the current phase is marked by the integration of the territory into the Soviet Union, a state centred on the eastern part of the European peninsula.

South and East Asia. The earliest European contacts with parts of southern and eastern Asia aimed at trade in the exotic products of the East; Europeans used their seapower to establish control over the trade routes. The European countries fought each other for the monopoly of the Eastern trade and established trading posts ("factories") at various points extending from Persia to South China and to the East Indies. The Portuguese, who were the first to arrive in India (1498), Malaya (1511), and southern China (1514), began to trade with these areas and established the first European trading posts there. The Dutch and the British followed not long after and the French and the Danes a little later. By 1700 the coasts of southern and southeastern Asia were dotted with European-controlled trade ports. Gradually these ports became points of territorial expansion, as Europeans increasingly intervened in the hinterlands so as to extend their control over production and trade. Some areas, such as Burma, Thailand, and Vietnam, were of little trading interest to the Europeans, and others, such as China, Korea, and Japan, declined to deal with the Europeans freely. By 1700 Portuguese power was in decline and the primary division of trading territories resulted from wars between the Dutch, English, and French, the Eastern conflicts often mirroring those in Europe itself. As a result, the Indies became a Dutch preserve, India became a British zone, and Spain held the Philippines. Only in the latter had the Christianizing of Asians as well as the acquisition of territory been an initial objective.

Trade and politics

Political settlements in Europe after the conclusion of the Napoleonic Wars (1815) affected Southeast Asia. The Indies became totally Dutch, Britain acquired control over Malaya, and France gave up its claims to large territories in India in favour of Britain. The trading companies of The Netherlands and Great Britain ceased to operate in 1798 and 1858, respectively, and territorial political administration was assumed by the two governments. France returned to Southeast Asia somewhat later, taking over weak political states unable to resist encroachment, in order to establish power in Indochina. During the 19th century the United Kingdom took over Burma by stages, finally annexing it to India, and eventually assuming political control over a portion of western Borneo never effectively occupied by The Netherlands. The United States took over the Philippines from the Spanish in 1898. The German effort to gain control over the Chinese Shantung Peninsula was thwarted by other European powers, so that Germany had to be content with a long-term lease on the territory of the city of Tsingtao in southern Shantung, and a railroad-building concession. Germany did, however, become a political power in the Pacific islands in the late 19th century. During the last half of the 19th century, European countries pressured China into granting small holdings, called treaty ports, which were guaranteed by treaties, and these dotted the China coast and extended up the Yangtze Valley as far inland as Chungking in West China.

Thailand (then called Siam) remained free from political encroachment, and both Korea and Japan spurned all European effort to establish trade or to obtain political privilege. Late in the 19th century, however, Japan began its own political expansion, first in the form of establishing a sphere of influence, and then by conquest, taking over the Kuril Islands to the north and the Ryukyu Islands to the south, in 1875, and gained Korea and Taiwan in 1895. In 1905 Japan obtained southern Sakhalin and then gradually gained control of Manchuria, finally establishing the puppet state of Manchoukuo there in 1932. The expansion of the Turkish Ottoman Empire in Asia Minor left much of Arabia outside organized political control, and numerous small sheikhdoms continued to exist independently around oases in southern Arabia and along the Persian Gulf.

Japanese expansion

Central Asia. In AD 1400 the pastoral empire of the Golden Horde ruled all of Central Asia and much of eastern European Russia. During the 15th and 16th centuries Russian agricultural colonization spread eastward, and the exploitation of furs and forest timber products began; by the end of the 16th century, Russian explorers had crossed the Urals. In 1639 the first Russian explorer reached the Pacific Ocean at the Sea of Okhotsk, and in 1650 the Russians and the Chinese reached their first impasse over the control of trade and territory in the Amur River Valley. As in Siberia there were only small and fragmented ethnic territories lacking formal political organization, Russian political control was rapidly and easily achieved there. Expansion east of the Caspian Sea, however, proceeded more slowly, and it took until the end of the 19th century to bring the many Islāmic pastoral societies under control and to extend Russia's boundaries to the frontiers of Persia and Afghanistan in the south and to China in the Central Asian zone.

THE CONTEMPORARY PATTERN

The end of colonialism. During the late 19th and early 20th centuries the European powers were educating Asian peoples in methods of modern political administration and economic development. Political and cultural nationalisms gained in strength after 1900, focussing on traditional culture systems in the various regions. Political independence came to parts of southwestern Asia between 1920 and 1926, after the collapse of the Ottoman Empire, as a result of a deliberately planned separation of various ethnic groups. Conflicts and changes continued in Southwest Asia as the century progressed, heightened by the setting up of the State of Israel in 1948.

The Japanese defeat of Western military power in Southeast Asia in 1942 gave an enormous psychological stimulus to movements for political independence even during the Japanese occupation, and each colonial ruler was faced with demands for the end of colonial status as World War II ended.

The United States in 1935 had promised the Philippines

Independence of colonial states

independence in 1945; this promise was fulfilled in 1946, after World War II, and this first release from externally administered status in the East was the forerunner of other moves. Political independence came to India and Pakistan in 1947; Ceylon (now Sri Lanka) and Burma in 1948; Indonesia in 1949; Cambodia in 1954; nominally to Vietnam in 1954; also nominally to Laos in 1954; and to Malaya in 1957, with the Borneo colonies being added in 1963 to form the Federation of Malaysia. Singapore withdrew from Malaysia to become an independent state in 1965. By the early 1970s the British crown colony of Hong Kong, Portuguese Timor, and Portuguese Macau remained as the only European colonial holdings in Asia. Meanwhile, the Soviet Union was able to avoid the charge of imperialism in Central Asia and Siberia because territories there were given the status of nominally autonomous republics within the framework of a federal union.

Problems of Asian nationalism. Three major conflicts occurred after World War II in which Western powers and their allies, Asian or other, were drawn into conflict with Communist-supported Asian forces. The first of these was the French war in Indochina (1946 to 1954), which ended with the withdrawal of the French and the nominal independence of the component territories of the region. The second was the Korean War (1950 to 1953), in which North Korean troops, later supported by the Communist Chinese, invaded South Korea, whose government was supported by United Nations forces; the result was a perpetuation of the division of the country into North and South by a cease-fire agreement. The third conflict was the war in South Vietnam, in which the United States supported the South Vietnamese government against Communist-backed guerrilla forces; fighting later spread to Laos and Cambodia.

By the beginning of the 1970s, Asian nationalism had recovered internal political and cultural control from the Western powers in all the larger countries of Asia, except in the Asian part of the Soviet Union. Within the Soviet Union, Russian strength had been sufficient to prevent outright rebellion, although the flight of some pastoral ethnic groups had continued back and forth across the boundary between the Soviet Union and China. Political independence had been much more completely achieved than had economic independence or social and economic advancement. Internal stability was also sometimes threatened by regional or ethnic imbalances, as in East Pakistan (now known as Bangladesh). Japan emerged as the most ethnically unified nation-state in Asia, and its great economic development from 1945 to the early 1970s resulted in relative internal stability. Elsewhere in Asia, political and cultural conditions were everywhere in flux, as nationalistic groups, ethnic minorities, and political parties strove for the attainment of different objectives and as different nation-states attempted to pursue their national interests. Contrasts between traditional rural societies and modern urban industrial populations raise new kinds of problems. The establishment of the People's Republic of China in 1949 represented one of the major political developments of the century in Asia and is fraught with long-term consequences for South and Southeast Asia. In Southwest Asia the Arab–Israeli conflict threatens to create worldwide repercussions if the conflicting interests cannot be reconciled.

Problems after independence

Continuing European linguistic influences. Europe has imprinted many permanent marks upon Asian cultures.



Political divisions of Asia.

Perhaps this is most noticeable in the continued use of European languages. Russian, English, and French are the dominant languages used, but German, Spanish, Dutch, and Portuguese are also employed in particular regions. Thus, German ethnic minorities in Soviet Central Asia continue to use German; older Filipinos speak Spanish; Indonesians use Dutch; and Portuguese is employed in Macau, Timor, and former Portuguese India. Vietnam, Cambodia, and Laos utilize French as a lingua franca, and French is taught in many schools. Dutch was the lingua franca of the educated class in Indonesia and is still taught in some schools. Pilipino is the official language of the Philippines, while English serves as a lingua franca and is taught throughout the schools. In Malaysia, India, and Hong Kong, English has an official status, is taught in schools, is spoken widely among the educated classes, and is the language of parliamentary debate. In parts of Southwest Asia, French and English are commonly spoken languages and are often taught in schools. Within Soviet Central Asia and Siberia, Russian is becoming more widely used and is the only common language. English, the most widely used non-Asian language, is a lingua franca throughout East, South, and Southwest Asia. (Jo.E.S.)

VII. Resource development

Factors in resource development

The utilization of Asia's natural resources has depended, to a large extent, not only on the development of technology but also on political circumstances. Thus, until the end of World War II and the beginning of the process of decolonization in Asia, most Asian countries were not free to develop their own natural resources independently and without reference to the economic interest of a metropolitan power. Cultural attitudes also affect the utilization of resources. To give but one example, cattle, which are a source of immense wealth in many developed countries, are a drain on scarce resources in India, where cultural taboos prohibit the slaughter of cattle either for purposes of food or for the conservation of resources when the animals are no longer productive.

The value of natural resources also varies with the prevailing technology. For example, with the application of new technology to the production of cereals, the same area of land can give greatly increased production. The application of modern technology has also produced improvement in many other areas, such as in Japan for the production of silk or of cultured pearls. Technology may also make it possible to exploit mineral wealth that was previously unusable because of problems of accessibility or of juxtaposition of other minerals.

INDUSTRIES

Mining. Asia produces a variety of minerals. Many of these are mineral fuels, such as coal and petroleum. The largest Asian producer of coal is the People's Republic of China, followed by Soviet Asia, India, North Korea, and Japan. Smaller quantities of coal are produced in a number of other countries. The Arab countries of Southwest Asia are among the principal producers of petroleum in the world. The biggest producer among the Asian countries is Iran, which is responsible for more than one-fifth of the total Asian production. Indonesia comes next, followed by Soviet Asia, China, India, and Brunei; some other countries produce much smaller quantities. Efforts are also being made to undertake offshore prospecting on a systematic basis. The biggest producer of natural gas is Soviet Asia, followed by Pakistan, Iran, Japan, Afghanistan, and Taiwan.

The largest producers of iron ore and ores for ferroalloys are China, Soviet Asia, India, and North Korea. Together, these four account for some 90 percent of the total production of the region. The People's Republic of China and India are among the ten major world producers of manganese ore and between them account for some 85 percent of Asia's total output. Asia's biggest producer of chromite is the Philippines, followed by Turkey, India, Iran, and Japan. There is also some production of tungsten in China and both North and South Korea and nickel in Indonesia. Soviet Asia was expected to become a major producer of many of the ferroalloys during the 1970s.

Asia is one of the world's main producers of tin-in-concentrates (tin ore that has been partially processed to increase the concentration of tin), providing some 59 percent of the world's total production. Malaysia alone accounts for some 55 percent of Asia's production, and Thailand has replaced Indonesia as the second largest Asian producer with some 16 percent. There is also considerable production of copper ore in the Philippines, Japan, and China.

The bauxite produced in Asia represents only a small part of total world production, although production in Soviet Asia is expected to increase during the 1970s. Development of the eastern Siberian gold mines has given Soviet Asia a leading position in the world's production of gold. The region, however, accounts for a third of the world's production of sulfur, principally from Japan and China; and more than three-fifths of the world's production of graphite, from Korea (North and South) and China.

Heavy industry and engineering. Despite the fact that the continent has such a variety of mineral resources, metallurgical industries have not been well developed, except in Japan and in Soviet Asia during the 1970s. The major producers of steel in the region are Japan, China, and India; Soviet Asia may rival both China and India by 1980. Japan, China, and India are also the major steel consumers, although the consumption of steel is increasing in Soviet Asia, Pakistan, and the Philippines, and Hong Kong is one of the main consumers on a per capita basis. Japan, China, and India are also the region's leading producers of metallurgical coke. It is to be expected that other countries will begin to establish steel plants and that the production of steel will continue to grow in Japan, China, and India.

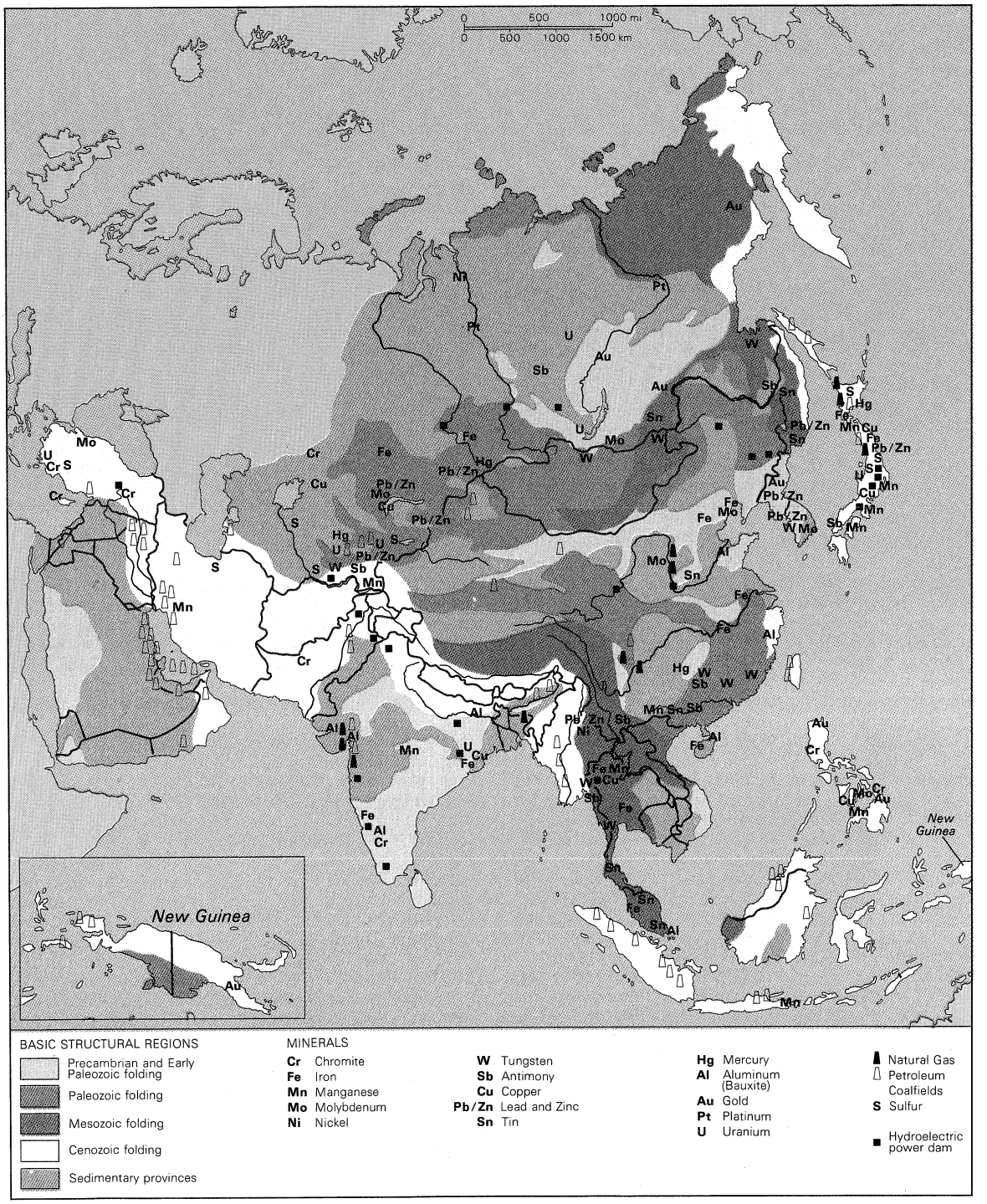
The production of aluminum is concentrated in four countries—Japan, Soviet Asia, India, and Taiwan. Japan accounts for over half of the total output of aluminum in the region. India is developing its aluminum production and has a relatively well-developed aluminum industry. There is also some production of copper, zinc, lead, and tin in Asia, with Japan leading in the production of zinc and lead and Malaysia in the production of tin. Japan is the leading consumer of tin, followed by India and China.

Asia's engineering industries

Japan produces every variety of engineering goods, from tankers and locomotives to miniaturized electronic equipment. In the postwar era, India has also gradually diversified its engineering industries and now produces heavy capital goods (machines and tools used in the production of other goods), various items of industrial machinery, prime movers (engines and other sources of motive power) and boilers, diesel engines, sewing machines, machine tools, agricultural machinery, and all kinds of electrical equipment. India also produces radio receivers, metal manufactures of various kinds, railway rolling stock, automobiles, bicycles, and precision instruments. The People's Republic of China also made considerable progress in the field of engineering industries during the 1950s and 1960s. Other Asian countries have concentrated on the production of durable consumer goods.

Chemical and petrochemical industries. The consumption of nitrogenous and phosphatic fertilizers has greatly increased in Asia and is certain to increase much more in the future as additional countries use the advanced techniques and improved seeds that have now become available. The major consumers of fertilizers, on a per acre basis of arable land, have been Japan, Taiwan, and South Korea. Because of their vast size, India and China, where the use of fertilizers is rapidly increasing, are, in absolute terms, among the major consumers. In production, Japan accounts for close to 55 percent of the output of nitrogenous fertilizers in Asia excluding the People's Republic of China. India has greatly increased its production, especially of ammonium sulfate, and has also experimented with fertilizers having a much higher nitrogen content, such as urea. Production of phosphatic fertilizers is also increasing.

Asia also produces and consumes basic chemicals, such



Basic structural regions and principal mineral and hydroelectric sites of Asia.

as caustic soda, soda ash, and sulfuric acid; Japan is the leading producer of these, followed by China and India.

The consumption of pulp and paper throughout the continent is growing steadily and is bound to increase with improved levels of living. The major consumers are Japan and India, and the major producers are China, Japan, and Indonesia.

Various surveys undertaken under the auspices of the United Nation's Economic Commission for Asia and the Far East (ECAFE) have shown that there is considerable scope for the manufacture of petrochemical products in Asia. At the Second Asian Conference on Industrialization (1970), a program for developing the petrochemical industry was initiated; this program included proposals to close the annual gap between prospective demand and capacity, estimated in terms of ethylene at about 500,000 tons. It was also recommended that more countries should manufacture machinery and equipment for the fabrication of plastics. An increase in the production of synthetic fibres, especially polyester, was also considered important.

The expanding textile industry

Manufacturing and textiles. The textile industries, particularly cotton, have expanded greatly in Asia since World War II. Japan and India have become the world's largest exporters of cotton textiles, and China, Taiwan, Pakistan, and Hong Kong have also entered the international market. The industry produces cotton yarn, cloth, and finished garments. There is also some production of wool (both yarn and woven fabrics) in the region. Here again, Japan, China, and India are the main producers and consumers; Japan is Asia's chief producer of woollen fabrics. Japan and India have also become major producers of woven rayon and acetate fabrics. Japan has also turned to noncellulose synthetic fibres, especially nylon,

acrylic, and polyester fibres. In the early 1970s Japan was second only to the United States in the production of these fibres and fabrics.

Industrial development in the region has made significant progress in relative terms, but it must be recognized that, in absolute terms, progress has been very limited; in relation to its size and vast population, the contribution of Asia to total world industrial output has been small. There is, however, a discernible trend toward a transition from light to heavy industry in many countries of the region, the development of Soviet Asia is dramatic, and it is likely that the continent will have an increasing share of world production in the final decades of the 20th century.

Timber, fisheries, and animal husbandry. Logs are exported from the Philippines, Malaysia, Indonesia, Thailand, and Burma to industrialized and timber-deficient countries, especially Japan. Thailand and Burma produce special varieties of timber such as teak. Thai teak is also exported to Europe.

Soviet Asia has an enormous forest area estimated at more than 500,000,000 hectares (1,250,000,000 acres). Present annual fellings in Soviet Asia are estimated at around 120,000,000 cubic metres (4,250,000,000 cubic feet), while natural losses of mature and overmature forests are estimated to be close to three times the present removals. Average yields of timber are 150 to 200 cubic metres per hectare (2,150 to 2,850 cubic feet per acre). The wood ranges from pine around the Bratsk area to a mixture of pine, larch, aspen, birch, and other species in the region south of Lake Baikal. Logging and transport operations are highly mechanized and have been facilitated by a road-building program.

Bamboos are an important component of wet evergreen, moist deciduous, and dry deciduous forests in the tropical parts of southeastern Asia, principally in Burma, Cambodia, Sri Lanka, India, Indonesia, Laos, Malaya, New Guinea, Pakistan, the Philippines, Thailand, and Vietnam. At higher altitudes and in temperate climates in Asia, as in Bhutan, China, Japan, and Nepal, many of the genera found in tropical parts are represented by different species, and other genera are common in China and Japan. In this connection, it is interesting to note that pure bamboo forests are common on slopes where temporary cultivation has been carried on in Burma, Bangladesh, and other parts of Asia.

Asia has a considerable potential for increased development of its fisheries. Japan has shown how far afield a well-organized fishing fleet can go in search of fish. In general, the problems of the fishing industry stem from lack of adequate capital and advanced technology, which tend to restrict fishing to coastal and offshore areas and make it difficult to extend the fishing to the deep seas. Because of a lack of refrigerated transport and storage, there are also the problems of preserving fish after the catch and of transporting the catches to centres of consumption. In some countries freshwater fish are also an important addition to the diet of the local people, and the raising of fish in culturally controlled ponds is important in southern China, Indonesia, and the Philippines. While the dairy industry is important in a few countries such as India, Pakistan, Soviet Asia, and Turkey, there is not much large-scale development of beef-cattle farming; Soviet Asia, however, is developing such patterns. Both China and Japan are discarding their traditional taboos against the use of milk products, and both countries have growing urban dairy industries. Only in areas where there are communities of overseas Chinese—in Singapore, for example—has an attempt been made to raise swine outside China, which probably leads the world in pork production. The poultry industry has made rapid strides during recent years, and the production of both eggs and broiling chicken has gathered considerable momentum. In the case of poultry, availability of feed is one of the major limiting factors for the further growth and development of the industry. Straw, obtained from the rice crop, is the primary fodder for livestock in southern Asia. Cattle feed is usually supplemented by concentrates, such as oil cake.

The raising of livestock

The raising of sheep and goats for meat and wool is especially important in Afghanistan, Pakistan, and southern Soviet Asia, and these animals are also raised in practically all the other countries of Asia. The sheep population of Southeast Asia is small.

In spite of the large number of cattle and sheep in the region, the hides and skins industry has not been adequately developed. Technological problems in connection with both the flaying of skins and their curing remain to be overcome.

Handicrafts. Traditional cottage industries and handicrafts continue to play an important role in the economies of all Asian countries. They not only constitute an important manufacturing activity in themselves but are almost the only available means of providing additional employment and of raising the level of living for both rural and urban populations. In view of the growing world market for the products of traditional Asian cottage industries and for Asian handicrafts, there is room for considerable expansion, especially in standardizing production and in marketing products in the advanced countries.

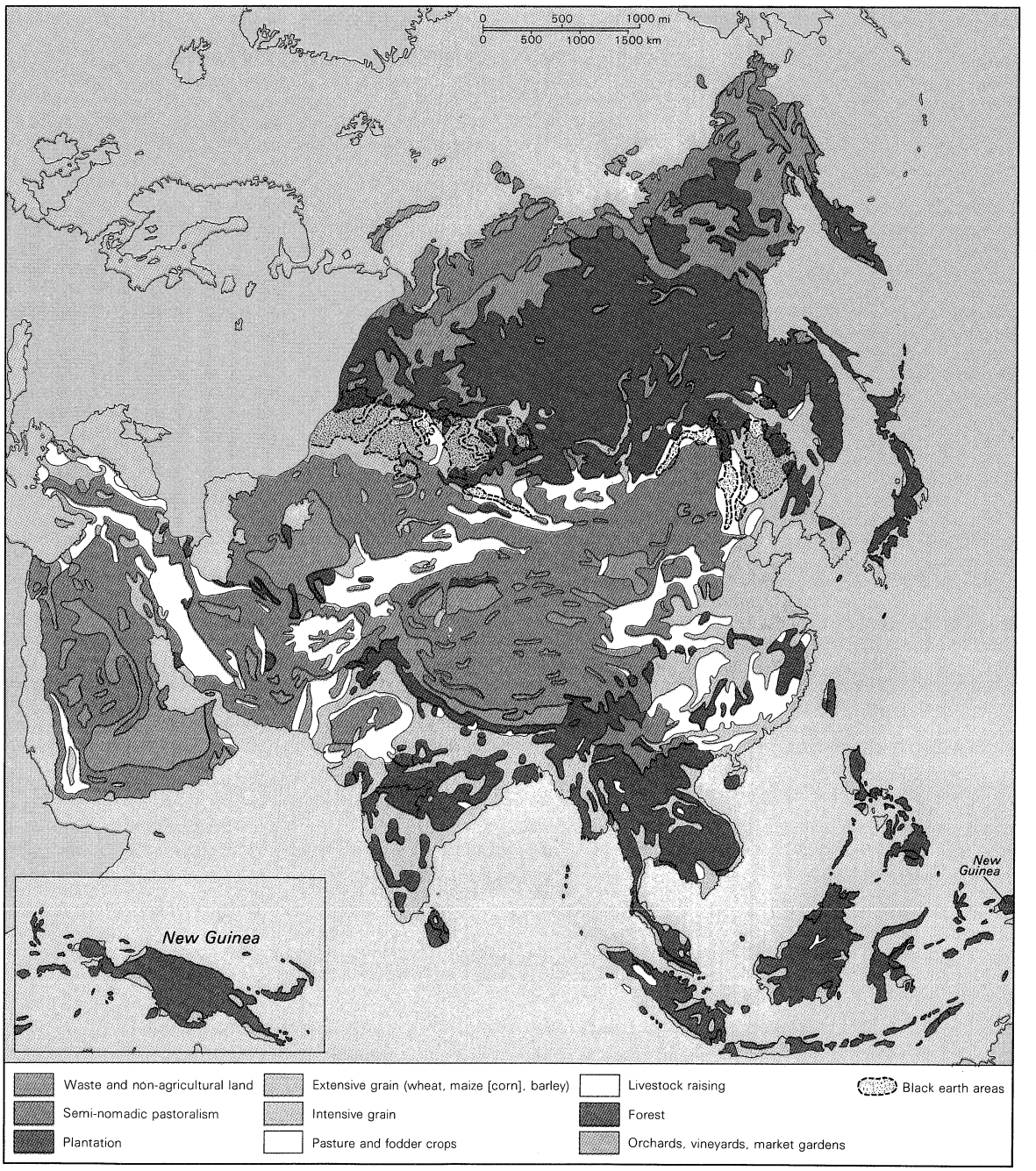
Other industries. Asian countries are at different stages in developing their pharmaceutical industries. The progress of the industry in Japan is comparable to that achieved in western Europe and the United States. Great progress has also been made in India and, in some respects, in Pakistan, but these two countries have still not reached the stage of being self-supporting in technology, raw materials, or equipment. China has begun to develop an industry based on a distinctive blending of Occidental and native pharmaceutical manufacturing. In most of the other countries, the pharmaceutical industry is only a processing industry based on basic drugs, imported in bulk, which are then marketed as capsules, tablets, and injectibles. It must also be remembered that in many Asian countries traditional medicinal products and treatments are still popular, especially in rural areas.

International tourism has developed to such an extent that by the early 1970s some 3,000,000 tourists annually were visiting Asian countries. The most visited places, in order of popularity in the year 1970–71, were Hong Kong, Japan, Thailand, Taiwan, Singapore, India, and Pakistan. Hong Kong and Singapore each have a big entrepôt trade and attract visitors primarily because they are duty-free ports. In the early 1970s, with the gradual lifting of the "bamboo curtain," it was anticipated that tourism in China would increase substantially as the decade progressed. There is considerable scope for further development of tourism in Asian countries, for many of them possess ancient monuments and natural attractions of great beauty. Unfortunately, many of these interesting places are not easily accessible, and the departments of tourism have often been unable either to organize or actively to promote tourism. In 1970 it was decided to establish the International Tourism Organization, as an intergovernmental agency. This institution is expected to provide greater technical assistance in the future to countries with specific plans to develop tourism.

The importance of tourism

POWER

The consumption and generation of electricity in Asia practically doubled during the 1960s. Even so, the per capita consumption of power in Asia is very small compared with the world average, amounting to less than 300 kilowatt-hours per capita, compared with a world average of between 1,000 and 3,000 kilowatt-hours in western Europe and more than 6,000 kilowatt-hours in North America. Japan is the biggest producer of power in Asia, and its generation capacity amounts to a little more than half the region's total. The annual electricity output by public utilities and similar bodies in the Asian countries belonging to the United Nations Economic Commission for Asia and the Far East (ECAFE) amounted to some 597,000,000,000 kilowatt-hours in 1970. Of this total, three-fourths was produced by thermal power and a quarter by hydroelectric power. Both types of production have progressed rapidly in Soviet Asia, which may rank close to Japan in total production by 1980.



Agricultural regions of Asia.

Thermal and hydroelectric power

Thermal power has, indeed, become the most important source of supply among the various types of electric power generated in the region. There is an increasing tendency to operate thermal power stations for meeting the regular demand and to build—or plan—hydroelectric power stations to serve during peak demand periods. In countries such as Afghanistan and Nepal, however, hydroelectric power generation is several times greater than the generation of thermal power. This is explained by the fact that, because these are landlocked countries, the imported fuel needed to supply thermal plants is expensive.

Nuclear power plants are being developed in a number of countries. Japan completed its first atomic power plant in 1967, with a capacity of 160 megawatts (a megawatt equals 1,000,000 watts), and is constructing more and more such plants, with a total capacity of some 2,400 megawatts. The capacity is expected to reach 6,000 megawatts by 1975. India commissioned its first nuclear power station at Tarapur in early 1969 with a capacity of 380 megawatts, and two more atomic power plants, each with a 400-megawatt capacity, are in the construction and planning stage. Pakistan has constructed a 137-megawatt nuclear power plant in Karachi. Taiwan is planning to have a 500-megawatt nuclear power station in operation by 1975. Other countries—the Philippines, South Korea, and Thailand—also have nuclear power plants either in the planning stage or under construction.

In the construction of steam thermal power stations, there is an increasing tendency toward a larger unit capacity operating at very high pressure and temperature. Japan has two small geothermal plants (which employ the heat of the Earth's interior) and is the only Asian country to use them. Small gas-turbine generating stations have also been installed in many countries. Pakistan has been using natural gas from the Sui gas field for both thermal and gas-turbine generation.

AGRICULTURE, IRRIGATION, AND LAND USE

High-yielding cereals

The most important development in Asian agriculture during recent years is the evolution of new high-yielding strains of cereals. This development is now being utilized in several Asian countries and is likely to have a marked

influence on the yield per acre of cereals during the 1970s and 1980s. Rice is, of course, the staple food crop among Asian countries. Asia produces some 60 percent of the total production of rice in the world. Except in Afghanistan, Soviet Asia, and Malaysia, rice occupies the greatest land area as a single crop. In Afghanistan and Siberia wheat is the dominant crop, while in Malaysia rubber occupies the greatest land area, with rice coming second. The total percentage of land under rice cultivation, as compared to total arable land, is highest in South Vietnam, Taiwan, and Thailand; it varies between 25 percent and 50 percent in most Asian countries.

In spite of the fact that the region is a major rice-producing area, several countries (among them Sri Lanka and Indonesia) are not self-sufficient in rice. Thailand, Burma, and Cambodia are exporters of rice, and South Vietnam was also an exporter until 1965. After rice, wheat is grown extensively in India, Pakistan, Iran, Turkey, Soviet Asia, northern China, Korea, and Japan. Barley and oats are also grown in these countries. Maize (corn) is grown in Soviet Asia, India, Pakistan, Afghanistan, Indonesia, and the Philippines. India, China, Pakistan, and Soviet Asia also grow sorghum and millet.

Asia produces several plantation crops, of which the most important are rubber, tea, coconuts, sugarcane, and pineapples. The major producers of rubber are Malaysia, Indonesia, India, Thailand, and Sri Lanka. India and Sri Lanka predominate in plantation tea production, and China, Taiwan, and Japan produce several types of tea on small holdings. Coconuts are an important crop in the Philippines, Indonesia, Sri Lanka, and India. India, the world's leader in sugarcane production, produces for domestic use only, whereas the Philippines, Indonesia, and Taiwan produce both for domestic consumption and export. The Philippines, Taiwan, and Malaysia produce pineapples, which are canned for export.

The continent produces a variety of tropical and subtropical fruit, mainly for domestic consumption. Transport facilities, where available, can be used only for limited distances. In view of the climatic conditions and the past lack of refrigerated transport, consumption is seasonal and is confined to areas close to centres of production. Among the main varieties of fruit produced are bananas, mangoes, apples, oranges, pineapples, several varieties of citrus fruit, papayas, and some specialities such as mangosteen (a dark reddish-brown fruit) and durian (a large oval fruit with a prickly rind, a soft pulp, and a peculiar odour). Taiwan, the Philippines, and Malaysia export bananas to Japan.

Except in a few countries, canning of surplus fruit has been developed only to a limited extent. In view of the tremendous potential for the production of fruit, there is vast scope for increased canning for export, whether of such fruits as mangoes and pineapples or of fruit juices from apples, oranges, or pineapples. The fruit canning industry has markedly increased in Taiwan.

The same factors affect the production of vegetables. Vegetables are produced mainly for local consumption, and only tubers can be transported over distances and stored for a period of time. In Taiwan, successful efforts have been made in the canning of mushrooms and asparagus, both of which are becoming leading exports.

Methods of irrigation

The traditional method of irrigation in Asia is by gravity water flow. The water from upstream storage reservoirs or diversion dams is carried through canals to field distributaries. The fields adjoin one another, and the water is able to flow from one field to the next; it may, however, take some time for the water to move across the terraced fields back to the canal system. The disadvantages of this system include the loss of water by evaporation and seepage and the possibility that the continuously flowing water will carry with it soil nutrients, fertilizers, and pesticides. In Japan and Taiwan, water is moved by small electric pumps, which operate continuously during the growing seasons.

In the early 1970s more attention was being given to the use of underground water by lift. The use of ordinary pumps as well as of deep-bore well turbine pumps is becoming more common, especially in India, Pakistan, and Iran. Such irrigation avoids some of the disadvantages of flow irrigation and allows for easier drainage.

Of the crops cultivated in the region, rice, sugarcane, and, in Soviet Asia, sugar beets need the most water. Cereals other than rice, legumes, and root crops can be grown even on rain-fed land. A survey of Asian agriculture undertaken under the auspices of the Asian Development Bank in 1968 quotes a figure of approximately 20 percent of the arable area under cereals as being served by irrigation. The availability of an assured source of water supply is an important element in the new technology, which also requires the use of fertilizer in conjunction with the improved cereal seeds that have been developed. Huge irrigation projects in Southern Siberia and southwestern Soviet Asia are rapidly altering traditional agricultural patterns.

VIII. Asian commerce

Several centuries before the Christian Era, Asian countries had commercial relations among themselves as well as with the countries of the West. In the earliest days, nomad tribes carried on this trade over considerable distances, using barter as the medium of exchange. Particularly important in such trade were fine textiles, silk, gold and other metals, various precious and semiprecious stones, and spices and aromatic products. There was a considerable expansion of trade during the Greek era around the 4th century BC, by which time various land routes had been well established connecting Greece, via Asia Minor, with the northwestern part of the Indian subcontinent. A further development of land and sea routes, especially to South India, occurred during Roman times. This East–West trade flourished in the first four centuries AD but was subject to considerable vicissitudes in later centuries. During this period there was also a great expansion of trade to Southeast Asia and to China through what are now Malaysia and Cambodia.

Ancient trade patterns

After Spain, in the 15th century, became interested in discovering a direct sea route to Asia—an interest that led to the discovery of the Western Hemisphere—the era of the great circumnavigators arrived in the 16th century. Portugal was one of the first countries to dream of establishing a monopoly over the lucrative spice trade with Asia. The Dutch and the British started similar enterprises at the turn of the 17th century, each country establishing its own East India company. The British began by centring their activities on the Indian subcontinent and extended their interest in due course to Burma, Ceylon (now Sri Lanka), and present-day Malaysia. The Dutch first concentrated on Ceylon but later expanded into and concentrated on Southeast Asia. The French were able to establish only minor footholds on the Indian subcontinent, but their 19th-century penetration of Indochina was more successful. Spain was content with the control it had established over the Philippines. In the course of time, the trading companies developed into colonial empires.

The export trade of Asia was formed by these historical circumstances. In the days before the East India companies, the products of Asia were those demanding the exploitation of human skills, such as silk and textiles, and such precious commodities as spices and aromatic products, which depend especially upon climatic and soil conditions and careful human labour.

The colonial trading system

With the development of the East India companies and the further assumption of colonial rule, a new pattern of trade emerged. Generally speaking, the colonial countries became the exporters of raw materials and imported the finished products from their metropolitan rulers. During this period, tea and tobacco also entered into international trade, and jute became a monopoly product of the Indian subcontinent. Until the end of the 19th century, Japan had trading relations mainly with Korea and China and generally remained somewhat aloof from world trade. China was the first country of Asia to have any significant trade relations with the Western Hemisphere; these were especially with the United States.

The latter half of the 19th century and the early part of the 20th constituted the heyday of colonial rule. By the

first decade of the 20th century, Japan had emerged as a major military and naval power and gradually developed into an important trading partner with the rest of the world. The era that followed was that of the colonies' struggle for political independence, which reached its climax immediately after World War II. Fifteen years after World War II the great British, French, and Dutch empires had virtually ceased to exist in Asia.

As a result of the achievement of independence by the former colonies, the emergence of the People's Republic of China, and the development of Japan into one of the major producing and trading countries of the world, the old established patterns of trade have changed considerably. Most of the newly independent countries are planning their economic and social development and have tried—with varying degrees of success—to diversify their production and to develop new industries, even if they cannot wholly overcome the handicaps of the colonial past.

TRANSPORTATION

Reference has already been made to the main transport systems that linked Asia and the Western world. Until the 19th century, the land, or caravan, routes, supplemented by oceangoing vessels, were predominant. In the latter half of the 19th century there was a major shift to seagoing vessels. Rail transport has begun to play a progressively more important role, mainly in the internal movement of passengers within individual states and in the transport of heavier goods over longer distances. Concurrently, there has been a considerable development of ports and harbours, linked to their hinterlands by rail and road. Air transport is also proving to be not only the speediest but often the cheapest means of transport, especially for costly items of relatively small weight and bulk. Air transport plays a particularly important role in landlocked countries—such as Afghanistan, Nepal, and Laos—and in the opening up of relatively inaccessible areas.

Means of internal transport

Within Asian countries, animal transport still remains the main means for the local transport of goods from one village to another or from the villages to a central marketplace. The animals, used for plowing during the cropping season, are also used for transport of goods at other times. The diesel truck, however, is now rapidly replacing draft animals for internal traffic, and there is a simultaneous development of roads and highways in most countries. In the 1960s and early 1970s, an attempt was being made, with international backing and cooperation, to establish an Asian highway network. When this project is completed, there will be a direct road link from Istanbul, in Turkey, to Singapore, and to Saigon, in South Vietnam.

Inland navigation is important in certain countries; a good river and canal system is capable of carrying goods and passengers at small cost over considerable distances. Among the countries having a well-developed inland water transport system are Bangladesh, Burma, Thailand, the former Indochina countries, and China. There are also great riverine ports such as Calcutta, Rangoon (Burma), Bangkok (Thailand), and Saigon; oceangoing ships can navigate the Mekong River to inland ports such as Phnom Penh (Cambodia) and can sail up the Yangtze River to Wu-han (China). Ultimately, it might be possible to connect even Laos with the sea by an extension of inland navigation facilities on the Mekong. The Yangtze, Sungari, and Hsi rivers of China provide a wide network of routes for motorized barges, supplementing traditional water transport.

For the movement of petroleum products, there has been some development of pipelines, especially in West Asia and western and southwestern Soviet Asia. Pipelines have considerable advantages, such as economy and speed, but they also have the disadvantage of being subject to political vicissitudes when they cross international boundaries. Meanwhile, there has been a considerable development of ever larger oil tankers, which, through economies of scale, can compete with the most efficient pipelines and can also effect point to point delivery.

INTERNAL TRADE

In view of the division of labour that existed between the colonial countries and the metropolitan powers in colonial days, it is not surprising that the economies of the newly independent countries of Asia are more competitive than complementary. In the case of certain countries, such as the Philippines, Sri Lanka, India, Afghanistan, Iran, and Pakistan, their intraregional exports in recent years amounted only to 4 to 15 percent of their total exports. On the import side, Iran, Afghanistan, Taiwan, South Korea, the Philippines, and India imported less than 10 percent of their requirements from other developing Asian countries. Japan, the most developed of Asian countries, exported about 35 percent of its products to developing ECAFE countries and received from them 23 percent of its imports.

Competitive economies

Asia is the biggest producer of rice in the world; rice is indeed one of the most important commodities of intraregional trade, and it is the most important export item of such countries as Burma, Thailand, and Cambodia.

There has been an effort on the part of Asian countries to improve their trading position by joining in commodity agreements. Malaysia, for example, is a member of the International Tin and Rubber Agreement. Other Asian countries are members of the International Sugar Agreement. Sri Lanka, India, Indonesia, and Pakistan have set up an International Tea Committee. The Asian Coconut Community was established in 1968, with Sri Lanka, India, Indonesia, Malaysia, the Philippines, Singapore, and Thailand participating. Most recently, there has been an effort to set up an Asian Pepper Board. Participation in these commodity agreements is not designed so much to promote intraregional trade as to help stabilize prices of primary products produced in Asia that enter into world trade.

The People's Republic of China was one of the main trading partners of the Socialist bloc of countries in Europe, especially the Soviet Union, in the early days after the revolution. China is now increasingly seeking trade contacts with other Asian countries. In many cases trade agreements between China and other Asian countries are in the form of barter agreements. China's most important trading partners in Asia have been countries such as Burma, Cambodia, Sri Lanka, India, Indonesia, Japan, Malaysia, Pakistan, and Singapore.

In 1958 a special effort was made to promote intraregional trade by the establishment, under the auspices of ECAFE, of intraregional trade talks. In the early 1970s, efforts were being made to establish an Asian Clearing Union, as a first step toward more ambitious regional and sub-regional monetary co-operation.

There has been little effort at trade integration on a regional or subregional basis in Asia. In this respect Asia lags behind Latin America and, to some extent, Africa. Trade between India and Pakistan, which could be of great mutual benefit, is virtually nonexistent because of the state of political relations between the two countries.

Many Asian countries are engaged in the process of diversifying their internal production. Their goal is frequently self-sufficiency rather than specialization and division of labour among a number of countries. As Asian countries increasingly become industrialized, however, and as they make greater use of their individual natural advantages, trade among the countries of the region could become more complementary.

EXTERNAL TRADE

The external trade of Asian countries has been considerably affected by the political ties that existed in colonial times, when important trade connections were established and when a certain trade pattern developed that proved convenient for both trading partners to continue. This convenience extended even to payments arrangements; the former British colonies were part of the sterling area, while the former French colonies found the franc a convenient medium of exchange. In addition, a system of Commonwealth preferences existed within the sterling area.

Several factors tended to interfere with this pattern in the 1960s. First, the European Economic Community (EEC) was formed, with its new system of preferences. Second, Japan was rapidly becoming a major producer of both consumer and capital goods and a major market for the commodities exported from other Asian countries. Third, the People's Republic of China began to take increasing interest in trade outside the Socialist bloc.

Major export items

The main items exported from the developing countries of Asia were rubber, tea, crude petroleum and petroleum products, rice, sugar, copra, coconut oil and palm oil, cotton and cotton fabrics, jute and jute fabrics, tin-in-concentrates, tobacco, wood products, iron ore, wool, and hides and skins. Spices, which were such an important part of the trade with the West in earlier centuries, now form a very small part of the total exports.

Since the major exports of the developing countries of Asia are still primary products, both their volume and price depend upon external demand, which fluctuates according to the level of industrial activity in the advanced countries. As a result, there are fluctuations both in the volume and price of the primary products exported. Further, there has been a tendency on the part of the advanced countries to replace some of these primary products in part by synthetic products, such as synthetic rubber and nylon and polyester fabrics. The developing countries are faced, therefore, with the problem that, even when they are able to increase their volume of exports, there is no corresponding increase in the value of their exports. In addition, the price of the consumer and capital goods exported by the advanced countries to the developing countries has been steadily rising. The establishment of the United Nations Conference on Trade and Development (UNCTAD), with a permanent secretariat in Geneva, is a response to the recognized need for an international agency to deal with some of these major problems of the developing countries and to help promote their trade with the advanced countries.

The main imports

The main imports of the developing countries have been machinery and transport equipment (including trucks, automobiles, and tractors); other manufactured goods; chemicals, including fertilizers; food, beverages, and tobacco, especially cigarettes; mineral fuels; and oils and fats. There has been a decline in British exports to the sterling bloc countries of Asia in the postwar period. The major trading partners of France in Asia have been the countries of Indochina, while The Netherlands trades mainly with Indonesia. West Germany has developed considerable trade with Asia in the postwar period.

Since 1959, the Soviet Union has increased its trade with Asian countries, especially Burma, Ceylon (now Sri Lanka), India, and Indonesia, which have been pursuing a neutralist foreign policy and have accepted economic aid from the Soviet Union as well.

The United States has been the biggest provider of economic aid to many Asian countries in the postwar period; there has also been a concurrent increase in trade between the United States and Asian countries. The dominant Asian trading partner of the United States is Japan; Hong Kong is next, followed by Taiwan and South Korea, after which come India, the Philippines, Indonesia, Pakistan, and Thailand.

In its efforts to trade with countries outside the Communist bloc, the People's Republic of China has increased its trade with countries such as the United Kingdom, Australia, Canada, and France; it has also entered into a series of trade agreements with Japan. Trade between China and the United States came to a virtual standstill in 1950, when the United States imposed an embargo on hundreds of commodities; while the ban continues to apply to traffic in certain strategic goods, in 1971 it was lifted for other commodities. As a result, it is to be expected that, in the course of the 1970s, the United States will become a significant trading partner of China. China has also been trying to extend its trade relations with the newly independent countries in Africa.

During this century, petroleum has become an important part of the trade of countries of West Asia with the rest of the world. Iran and the Arab countries and sheikhdoms are among the chief beneficiaries of this trade. Originally, the Western powers had considerable economic superiority, which they used to negotiate terms that were less advantageous to the producing countries. The producing countries, in their turn, began to work closely together to protect their common interest; recent negotiations, resulting in much better terms for the producing countries, have shown the strength of the Organization of the Petroleum Exporting Countries (formed in 1960).

COMMERCIAL PROSPECTS

Asia, with three-fifths of the world's population, has today a disproportionately small share of the world's trade. The trade between the People's Republic of China and the rest of the world is just beginning to grow and—with the lifting of the United States embargo on trade in most commodities—China is likely to have an increasing share of Asian trade. Japan has embarked on a period of growth that may be expected to continue. India is rapidly becoming a major industrial producer. Meanwhile, thanks to the Green Revolution (*i.e.*, the spectacular increase in agricultural yields due to advances in technology), the production of cereals—especially rice in Asia—almost doubled during the latter part of the 1960s. While much of this increased production will be needed to feed the millions of hungry mouths in Asia itself and may reduce the interregional trade in rice, it will also make it possible for Asian countries to make greater efforts toward industrialization and diversification.

The Green Revolution

The trade of Asia with the rest of the world, in percentage terms, is bound to increase in the final quarter of the 20th century. Similarly, trade among Asian countries is also bound to grow. It is likely, however, that there will be a marked change in the pattern of trade, with Asia exporting more finished and semifinished goods as well as raw materials, and also competing with the West in exporting certain types of manufactured products. Even so, there will still be a range of sophisticated products, such as intercontinental aircraft and high-speed computers, which most Asian countries will still find it necessary to import.

(C.V.N.)

IX. Demographic patterns

PRESENT POPULATION PATTERNS

Size of population. Asia, covering about a third of the total land area of the world, had a population of about 2,164,000,000, or just over half the world total, in the early 1970s. Asia includes the two countries that in 1971 had the largest populations in the world—the People's Republic of China, with an estimated population of about 732,000,000, and India, with a population of about 550,000,000; these two countries alone had populations that together were estimated to comprise more than a third of the world's people.

Age and sex composition. The age structure of Asia's population, particularly in the developing countries, is predominantly young. Roughly 40 percent of the population is under 15 years of age; about 55 percent is between 15 and 59 years old; and about 5 percent is 60 years old or older. One consequence of this is that the number of dependents—particularly children—is disproportionately large in relation to the number of gainfully employed adults. Another is that, in view of the present high birth rate, the age structure favours large additions in the future to the massive population already in existence.

In nearly all countries of the world, more male than female babies are born. In advanced industrialized countries, where maternal mortality is low and where baby girls receive as much care as baby boys, the male death rates are higher than female death rates at every stage of life; the numerical excess of males at birth is, in consequence, gradually reduced until females outnumber males in the older age groups.

Some Asian countries, particularly India and Sri Lanka, as well as Pakistan and a few predominantly Muslim countries, have a high sex ratio—*i.e.*, the number of males per thousand females—in the sense that males

outnumber females in all age groups, even though in a few categories of the population females predominate at birth. This unusual sex ratio has led to some controversy as to the reasons for it. In some countries, current social attitudes are held to be responsible for the differential mortality rates of the sexes after birth. Early marriages —if not quite child marriages—increase the initial balance in favour of males; the reason for this is the relatively high mortality rate of mothers in childbirth.

Density and distribution. Although it is difficult to generalize about a continent as vast as Asia, certain common features and trends can be mentioned. At the outset it may be pointed out that Japan, which is modernized and Westernized, and Israel, which is virtually a European enclave on Asian soil, are the two best known exceptions to many generalizations.

Regions of high population densities

While Asia and Europe are the two most densely populated continents, Asia is slightly more overcrowded than Europe. The density in Asia is 127 persons per square mile, while that in Europe is 114. The density in North and South America is 28 and 35, respectively, while the comparable figures for Africa and Oceania are 30 and 5. In the United States the density per square mile is about 57, but it is 1,089 in Taiwan and 728 in Japan. The densities of India and Pakistan are about 434 and 318, respectively, while the newly created nation of Bangladesh (formerly East Pakistan) has a population density of 1,361. Kerala state in India has 1,420 persons per square mile, and West Bengal has about 1,310. The island of Java in Indonesia has about 1,365 persons per square mile. The density of China, however, is only about 200 per square mile.

Migration to towns and cities. The distribution of Asia's population has traditionally followed a pattern of dense settlement in river valleys, where soil was fertile because of perennial irrigation, and where double-cropping (the harvesting of two crops a year) sustained large numbers. Today this traditional pattern is still in existence, despite the emergence of another pattern, represented by the attraction of large and congested industrialized cities such as Tokyo, Calcutta, Bombay, Shanghai, and Tientsin—cities to which the underemployed rural population often migrates. The depressed rural economy, the periodic failure of the monsoon, and the near-famine conditions that result have contributed to the continual drift of population to the towns and cities; often, however, the shortage of available jobs results in many unskilled village farmers, who come to the city in search of work, ending up in slums rather than as operators of factory machines.

While haphazard, unplanned urbanization is proceeding in many Asian countries, the population in general remains predominantly (about 70 percent) rural, except in Japan, Taiwan, Israel, and the Philippines.

FUTURE TRENDS

The two variables that determine future trends in the size of population are the birth rate and death rate. The factors affecting the behaviour of these two variables, including birth control, may therefore be examined before indicating possible trends in the size of population.

Birth rates. The world average birth rate during the last half of the 1960s was estimated at about 34 per thousand persons. Africa has the highest birth rate, with 46 per thousand. Asia comes second, with a rate of about 38 per thousand. Within Asia, the birth rates range from 51 per thousand in Pakistan, 49 in Jordan, and 42 in Malaysia to 39 in India, 32 in China, 29 in Ceylon (now Sri Lanka), 27 in Lebanon, and 27 in Soviet Asia. During the 1960s the birth rate was 42 per thousand in West Asia, compared with 33 in East Asia.

Several Asian countries, aware of demographic trends and their adverse effect on economic growth and social progress, have embarked on official birth control programs, which have met with varying degrees of success. Japan's program has perhaps been the most effective. In operation since World War II, it includes well-publicized family-planning services, legalized abortion, and the provision of all forms of contraceptive devices. Programs in China, South Korea, Taiwan, India, and Sri Lanka offer family-planning services, birth-control clinics, vasectomies, and contraceptives (including intra-uterine devices). The Soviet Union has an ambivalent population control policy, and birth rates in Soviet Asia continue to be relatively high. The Southeast and Southwest Asian countries lag behind in formal programs, but public consciousness and basic planning were growing in the early 1970s. It is not yet clear whether or not methods of birth control will be successful in the short term, but the long term effect is expected to show important decreases in the birth rate.

Death rates. A falling death rate and its corollary, a rising life expectancy, are due in a large measure to man's increasing control over disease. The death rate in northwestern Europe and the Soviet Union is less than 10 per thousand persons, whereas the rates in Asia are relatively high—about 20 to 35 per thousand in Nepal and Burma, 20 in Malaysia, about 17 in China, 14 in India, about 12 in Soviet Asia, about 8 in Sri Lanka, and 7 in Japan and Thailand. Medical and public health facilities have a long way to go in much of Southwest Asia, India, Pakistan, and Indonesia before they will equal European standards, but as they improve, the Asian rates will drop still further.

Infant mortality rate

The infant mortality rate—a sensitive index of the standard of living and of the cultural milieu—is also declining in many countries. Compared with an infant mortality rate of between 20 to 35 per thousand live births in northwestern Europe, Canada, and the United States, the rates in Asian countries have dropped, particularly since 1940, to 139 per thousand in India and 125 in Indonesia, 50 in Sri Lanka, 50 in China, 12 in Japan, 21 in Taiwan, and 23 in Israel. The rates in Burma and the Philippines are about 150 and 67, respectively. The present rates offer room for considerable reduction, but it must be remembered that in some parts of Asia they were between 150 and 200 as late as 1940.

Population growth. With the population of most Asian countries increasing at a rate of between 2.5 and 3.5 percent per year, Asians are increasing at a faster rate than Europeans; this was not so during the 18th and 19th centuries. The increase has been caused by the belated health revolution, bringing a dramatic decline in the death rate and particularly in the infant mortality rate. Since, moreover, Asia's population is a young one, the potential for future growth is great.

A declining death rate and a high and slowly declining, if not stationary, birth rate mean that the survival rate is high, yielding huge net annual additions to the population. The population of India, for instance, increased by more than 78,000,000 during the decade from 1951 to 1961 and added some 108,000,000 between 1961 and 1971.

Population projections

At the current rate of increase, according to one estimate, Asia will add another 1,000,000,000 in the decade from 1970 to 1980. By AD 2000, according to one projection, Asia may have more than 60 percent of the total world population, which by then would have doubled to about 7,000,000,000. If Asia reaches 4,000,000,000 to 4,500,000,000 by AD 2000, it is difficult to see how it can afford this enormous increase in the light of its present depressed standards of living and its current rate of economic development; indeed, it is highly questionable whether so many people could exist on the basis of projected available resources. The projections of astounding growth, however, do not account for present and future programs of population control. It may be said that the course of future events in Asia depends primarily on the response made to the challenge posed by the problem of overpopulation. (S.Ch.)

BIBLIOGRAPHY. The works cited below represent only a fraction of the extant literature on Asia, the purpose being to provide a highly selective reading list that bears on some of the topics dealt with in the article itself. Additional references to further literature on a large number of special topics may be found in these works. For a detailed description of the land and people of Asia, see GEORGE B. CRESSEY, *Asia's Lands and Peoples: A Geography of One-Third of the Earth and Two-*

thirds of Its People, 3rd ed. (1963); FRANK M. LeBAR, GERALD C. HICKEY, and JOHN K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964); БОРИС ФЕДОРОВИЧ ДОБРЫНИН *et al.*, *Зарубежная Азия; физическая география* (1956); and *Общий обзор* in *Советский Союз* (1972). Works treating the influence of geographical conditions on political, social, cultural, and economic conditions include: J.E. SPENCER and WILLIAM L. THOMAS, *Asia, East by South: A Cultural Geography*, 2nd ed. (1971); and C.A. FISHER, *South-east Asia: A Social, Economic and Political Geography*, 2nd ed. (1966). OWEN LATTIMORE (ed.), *Silks, Spices and Empire: Asia Seen Through the Eyes of Its Discoverers* (1968), is a readable survey of early travel accounts. For more specialized treatment of specific aspects of some Asian countries, the following works may be consulted: GEORGE B. CRESSEY, *Land of the 500 Million* (1955); KENNETH S. LATOURETTE, *The Chinese: Their History and Culture*, 3rd ed. rev. (1946); ARTHUR L. BASHAM, *The Wonder That Was India: A Study of the History and Culture of the Indian Sub-Continent Before the Coming of the Muslims*, rev. ed. (1963); JAWAHARLAL NEHRU, *The Discovery of India*, 4th ed. (1956); RUTH BENEDICT, *The Chrysanthemum and the Sword: Patterns of Japanese Culture* (1946, reprinted 1967); and GEORGE B. SANSOM, *Japan: A Short Cultural History*, rev. ed. (1943, reprinted 1962). For Eastern religion and philosophy, see KENNETH P. LANDON, *Southeast Asia: Crossroad of Religions* (1949, reprinted 1969); SUKUMAR DUTT, *Buddhism in East Asia: An Outline of Buddhism in the History and Culture of the Peoples of East Asia* (1966); S. RADHAKRISHNAN, *Eastern Religions and Western Thought* (1959) and *The Hindu View of Life* (1957); and JOSEPH M. KITAGAWA, *Religions of the East*, enl. ed. (1968). Works dealing with the varied impact of major European nations upon the peoples of Asia include: K.M. PANIKKAR, *Asia and Western Dominance: A Survey of the Vasco da Gama Epoch of Asian History, 1498–1945* (1953); EDWIN O. REISCHAUER, JOHN K. FAIRBANK, and ALBERT M. CRAIG, *East Asia: The Modern Transformation* (1965); GEORGE B. SANSOM, *The Western World and Japan* (1950); and NORMAN JACOBS, *The Origin of Modern Capitalism in Eastern Asia* (1958). The rise and development of nationalist movements in Asia is dealt with in JAN M. ROMEIN, *Das Jahrhundert asiens: Geschichte des modern asiatischen Nationalismus* (1958; Eng. trans., *The Asian Century: A History of Modern Nationalism in Asia*, 1962); PHILIP WARREN THAYER and WILLIAM T. PHILLIPS (eds.), *Nationalism and Progress in Free Asia* (1956); WILLIAM MacMAHON BALL, *Nationalism and Communism in East Asia*, 2nd ed. rev. (1956); and WILLIAM L. HOLLAND (ed.), *Asian Nationalism and the West* (1953).

Geological history: JOHN W. GREGORY, *The Structure of Asia* (1929); KURT LEUCHS, *Geologie von Asien*, vol. 1, 2 pt. (1935–37); SSU-KUANG LI, *The Geology of China* (1939); JAPAN, GEOLOGICAL SURVEY, *Geology and Mineral Resources of Japan*, 2nd ed. (1960); MASAO MINATO, MASAO GORAI, and MITSUO HUNAHASHI (eds.), *The Geologic Developments of the Japanese Islands* (1965); RAYMOND FURON, *Introduction à la géologie et à l'hydrogeologie de la Turquie* (1953); JOVAN STOCKLIN, "Structural History and Tectonics of Iran: A Review," *Bull. Am. Assoc. Petrol. Geol.*, 52:1229–1258 (1968); AUGUSTO GANSSER, *Geology of Himalayas* (1964); P.V. RAO, "Geology and Mineral Resources of India," *Int. Geol. Congr.*, 22nd session, New Delhi (1964); REINOUT VAN BEMMELEN, *The Geology of Indonesia*, 2 vol. (1949); F.A. VENINGMEINESZ, "Indonesian Archipelago: A Geophysical Study," *Bull. Geol. Soc. Am.*, 65:143–164 (1954); B.M. GOZON, "Geology of the Philippine Islands," *Petrol. Engr.*, 33:B64–66, 69 (1961).

Physical geography: General works include: PIERRE GOUROU, *L'Asie* (1953); R.R. RAWSON and W.G. EAST, *Asia* (1966); RAOUL BLANCHARD, *Asie occidentale*, and FERNAND GRENARD, *Haute Asie* (1929); J. SION, *Asie de moussouns*, 2 vol. (1928–29); and R.R. RAWSON, *The Monsoon Lands of Asia* (1964). Representative photographs of the Asian landscape are presented in MARTIN HURLIMANN, *Asien: Bilder seiner Landschaften* (1956; Eng. trans., *Asia: 289 Pictures in Photogravure*, 1957). The geomorphology of Asia is treated in FRITZ MACHATSCHEK, *Das Relief der Erde*, 2 vol. (1955).

Flora and fauna: L.S. BERG, *Natural Regions of the U.S.S.R.* (1950; Eng. trans. from the 2nd Russian ed., 1938); H.G. CHAMPION, *A Preliminary Survey of the Forest Types of India and Burma* (1936); PHILIP J. DARLINGTON, *Zoogeography* (1957); *Flore générale de l'Indochine*, 7 vol. (1905–52); GUSTAV FOCHLER-HAUKE, "Das Waldkleid un die Pflanzenbezirke Süd-Chinas," *Mitt. Geogr. Ges. Wien*, 78:158–178 (1935); *Forestry in Japan*, issued by the Tokyo Forestry Agency (1964); BUNZO HAYATA, "General Aspects of the Flora of Japan . . .," in *Scientific Japan, Past and Present*, pp. 77–104 (1926); P. LEGRIS, *La Végétation de l'Inde* (1963); LIOU HO, *Lauracées de Chine et d'Indochine* (1934); PAUL MAURAND, *L'Indochine forestière* (1943); F.J. ORMELING, *The Timor Problem* (1955); JULES VIDAL, *La Végétation du Laos* (1956); E.H. WALKER, "The Plants of China and Their Usefulness to Man," *A. Rep. Smithson. Instn.*, pp. 325–361 (1943); WANG CHI-WU, *The Forests of China* (1961); R.O. WHYTE, "The Phytogeographical Zones of Palestine," *Geogrl. Rev.*, 40:600–614 (1950).

Natural resources: E.A. ACKERMAN, *Japan's Natural Resources and Their Relation to Japan's Economic Future* (1953); VIOLET CONOLLY, *Beyond the Urals: Economic Developments in Soviet Asia* (1967); J.A. HODGKINS, *Soviet Power: Energy Resources, Production and Potentials* (1961); M.S. KRISHNAN, *Geology of India and Burma*, 5th ed. (1968); P.E. LYDOLPH and THEODORE SHABAD, "The Oil and Gas Industries in the U.S.S.R.," *Ann. Ass. Am. Geogr.*, 50:461–486 (1960); A.A. MINC, "Geographische Probleme der Ausnutzung der natürlichen Ressourcen in der UdSSR," *Petermanns Geog. Mitt.*, 114:21–28 (1970); "Natural Resources in Malaysia and Singapore," in B.C. STONE (ed.), *Proceedings of the 2nd Symposium* (Kuala Lumpur, 1969); D.B. SHIMKIN, *Minerals: A Key to Soviet Power* (1953) and *The Soviet Mineral-Fuels Industries, 1928–1958: A Statistical Survey* (1963); WANG KUNG-PING, "Mineral Resources of China, with Special Reference to the Nonferrous Metals, *Geogrl. Rev.*, 34:621–635 (1944); R.O. WHYTE, *Grasslands of the Monsoon* (1968) and *Land, Livestock, and Human Nutrition in India* (1968).

Human resources and political geography: W.G. EAST and O.H.K. SPATE, *The Changing Map of Asia*, 4th ed. (1961), is a general political geography; and ALISTAIR LAMB, *Asian Frontiers: Studies in a Continuing Problem* (1968), concentrates on the problem of political frontiers. GUY WINT (ed.), *Asia Handbook* (1969), is a current events summary and detailed fact book dealing with the continent. C.G.F. SIMKIN, *The Traditional Trade of Asia* (1968), discusses the history of the growth of trade between the Occident and the Asian regional zone, often with considerable geographic materials included. W.B. FISHER, *The Middle East: A Physical, Social, and Regional Geography*, 5th ed. (1964), provides the most complete coverage of southwestern Asia; and GEORGE B. CRESSEY, *Crossroads: Land and Life in Southwest Asia* (1960), is somewhat more thematic in its coverage. J.E. SPENCER and W.L. THOMAS, *Asia, East by South: A Cultural Geography*, 2nd ed. (1971), is a full-scale geography dealing with the region from Pakistan through Japan; whereas C.A. FISHER, *South-East Asia: A Social, Economic, and Political Geography*, 2nd ed. (1966); and DONALD W. FRYER, *Emerging Southeast Asia: A Study in Growth and Stagnation* (1970), deal in more detail with the mainland of Southeast Asia and Indonesia–Philippines. RICHARD A. BUTWELL, *Southeast Asia Today—and Tomorrow: Problems of Political Development*, 2nd ed. (1969), concentrates on the problems of political nationalism in the former colonial countries. P.E. LYDOLPH, *Geography of the U.S.S.R.*, 2nd ed. (1970), is a full-scale geography of the whole of the Soviet lands, including Central Asia and Siberia. W.H. PARKER, *An Historical Geography of Russia* (1968), concentrates on the historical geography and expansion of the Soviet political state. EDWARD ALLWORTH (ed.), *Central Asia: A Century of Russian Rule* (1967), is a political-historical study of Russian control of Central Asia; and LAWRENCE KRADER, *Peoples of Central Asia* (1963), is primarily concerned with the historic ethnic groupings of Central Asia.

Resource development: For the continent as a whole, the basic, most authoritative, and up-to-date sources of economic data and information are the publications of the United Nations Economic Commission for Asia and the Far East (ECAFE) and certain other intergovernmental organizations. The *Economic Survey of Asia and the Far East*, prepared and published annually by ECAFE, is the most comprehensive account of the economic situation in the continent as a whole. A multi-disciplinary analysis of the problems of underdevelopment, development, and planning for development in Asian countries will be found in GUNNAR MYRDAL, *Asian Drama: An Inquiry into the Poverty of Nations*, 3 vol. (1968); the countries covered include Burma, Sri Lanka, Cambodia, Laos, Indonesia, Malaysia, Thailand, and Pakistan, but most of the work is devoted to India. The following studies in ECAFE's "Mineral Resources Development Series" may be particularly mentioned in respect to mining and industries: *Mining Developments in Asia and the Far East* (annual); *Mining Developments in Asia and the Far East: A Twenty-Year Review, 1945–1965* (1967); *Lignite Resources of Asia and the Far East: Their Exploration, Exploitation and Utilization* (1956); *Copper, Lead and Zinc Ore Resources of Asia and the Far East* (1960); *Bauxite Ore Resources and Aluminum Industry of Asia and the Far East* (1962); *Tin Ore Resources*

of Asia and Australia (1964); *Mineral Raw Material Resources for the Fertilizer Industry in Asia and the Far East* (1967). Data on mineral resources of China may be found in K.P. WANG, "The Mineral Resource Base of Communist China," in U.S. CONGRESS, JOINT ECONOMIC COMMITTEE, *An Economic Profile of Mainland China*, vol. 1 (1967). ECAFE's studies relating to industries include *Asian Industrial Development News* (annual); *Industrial Development in Asia and the Far East* (1965); *Development Prospects of Basic Chemical and Allied Industries in Asia and the Far East* (1963); and *Industrial Developments in Asia and the Far East: Selected Documents Presented to the Asian Conference in Industrialization*, 4 vol. (1966). Works dealing with industries in China include CHOH-MING LI (ed.), *Industrial Development in Communist China* (1964); FREDERICK M. CONE, *Chinese Industrial Growth: Brief Studies of Selected Investment Areas* (1968); BARRY M. RICHMAN, *Industrial Society in Communist China* (1969); and YUAN-LI WU, *The Steel Industry in Communist China* (1965). The role of small industries in the development effort is discussed in "Modernization of Small Industries in Asia," *Economic Bulletin for Asia and the Far East*, 11:24–40 (1960); and CARL RISKIN, "Small Industry and the Chinese Model of Development," *China Quarterly*, 46:245–273 (1971).

ECAFE's publications relating to the development of power resources include: *Electric Power in Asia and the Far East* (annual); *Proceedings of the Regional Seminar on Energy Resources and Electric Power Development* (1962); and *The Role and Application of Electric Power in the Industrialization of Asia and the Far East* (1965). Projections of future energy requirements are made in the development plans of many countries, but the NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Demand for Energy in India, 1960–1975* (1960), may be cited as an early example of non-official estimates. Material relating to the power industry in China includes: YUAN-LI WU, *Economic Development and the Use of Energy Resources in Communist China* (1963); JOHN ASHTON, "Development of Electric Energy Resources in Communist China," in U.S. CONGRESS, JOINT ECONOMIC COMMITTEE, *An Economic Profile of Mainland China*, vol. 1 (1967); and ROBERT CARRIN, *Power Industry in Communist China* (1969). Possibilities and the potential of nuclear power development are discussed in the *Reports* (1959–60) of the Preliminary Assistance Missions sent to several Asian countries by the International Atomic Energy Agency. A list of civilian power reactors in operation or under construction with all pertinent technical data may be found in the INTERNATIONAL ATOMIC ENERGY AGENCY, *Power and Research Reactors in Member States* (published twice a year).

Agriculture: Two among the several publications of the Food and Agriculture Organization of the United Nations (FAO) are: *The State of Food and Agriculture* and *FAO Rice Report* (both published annually). The ASIAN DEVELOPMENT BANK, *Asian Agricultural Survey*, 2 vol. (1968), is another recent comprehensive survey covering most of the continent. A succinct account of Asian agriculture may be found in COLIN CLARK, "Agriculture in Asia," *Pacific Community*, 4: 283–294 (1970); and the ways to improve its contribution to the economy are discussed in "Strategies for Agricultural Growth," *Economic Survey of Asia and the Far East, 1969*, pt. 1A (1970). The agricultural situation in China is described in J.L. BUCK, O.L. DAWSON, and YUAN-LI WU, *Food and Agriculture in Communist China* (1966). The growth in cereals output and its economic and social impact may be found in LESTER R. BROWN, "The Agricultural Revolution in India," *Foreign Affairs*, 46:688–698 (1968); SAM-CHUNG HSIEH, "New Outlook for Asian Agriculture," *International Development Review*, 10:6–9 (1968); UNITED STATES DEPARTMENT OF AGRICULTURE, *Taiwan's Agricultural Development: Its Relevance for Developing Countries Today* (1968); NORMAN E. BORLAUG, "The Green Revolution: For Bread or Peace?" *Bull. Atom. Scient.* (1971).

Commerce: (Transport): The status and problems of transport in Asian countries and its role in their economic development is exhaustively dealt with in "Transport Development," *Economic Bulletin for Asia and the Far East*, vol. 11, no. 3 (1960). The Asian Highway Project, designed to connect the capitals and seaports of Asian countries, is described by M.S. AHMAD, "The Asian Highway," *ibid.*, 19:45–48 (1968). Despite its status as a major mode of transport in the rural areas of most Asian countries, not much literature is available on animal transport; the INDIAN PLANNING COMMISSION, *Role of Bullock Carts and Trucks in Rural Transport: Case Studies* (1963), is most valuable. WILFRED OWEN, *Distance and Development: Transport and Communications in India* (1968), is an excellent treatment of the subject for India, as are VICTOR D. LIPPIT, "Development of Transportation in Communist China," *China Quarterly*, 27:101–119 (1966); and YUAN-LI WU, *The Spatial Economy of Communist China: A Study on Industrial Location and Transportation* (1967). *(Trade)*: A historical perspective of Asian trade from its remote beginnings may be found in C.G.F. SIMKIN, *The Traditional Trade of Asia* (1968). J.C. VAN LEUR, *Indonesian Trade and Society: Essays in Asian Social and Economic History* (1955), includes a sociological interpretation of early Asian trade. B.G. GHATE, *Asia's Trade* (1948), is a descriptive account of the situation up to the early 1940s; and ALFRED K. HO, *The Far East in World Trade: Developments and Growth Since 1945* (1967), brings the general survey up to 1960. Purely descriptive literature for recent periods is scanty, and what is cited hereafter is more analytical. Among several studies prepared by ECAFE, the following may be particularly mentioned: "Asia's Trade with Western Europe, with Special Reference to the Common Market," *Economic Survey of Asia and the Far East, 1962*, pt. 1 (1963); "Foreign Trade of ECAFE Primary Producing Countries," *Economic Survey of Asia and the Far East, 1959*, pt. 2 (1960); "Trade Between Developing ECAFE Countries and Centrally-Planned Economies," *Economic Bulletin for Asia and the Far East*, 15:16–51 (1964); and "Intra-Regional Trade As Growth Strategy," *Economic Survey of Asia and the Far East, 1969*, pt. 1B (1970). SEIJI NAYA, "The Commodity Pattern and Export Performance of Developing Asian Countries to the Developed Areas," *Economic Development and Cultural Change*, 15: 420–437 (1967), is comprehensive and analytical. Material relating to the foreign trade of China includes: PAULINE LEWIN, *The Foreign Trade of Communist China: Its Impact on the Free World* (1964); and ALEXANDER ECKSTEIN, *Communist China's Economic Growth and Foreign Trade: Implications for U.S. Policy* (1966).

Demographic patterns: GEORG BORGSTROM, *The Hungry Planet: The Modern World at the Edge of Famine* (1965); S. CHANDRASEKHAR, *India's Population: Facts, Problem and Policy* (1967), *Hungry People and Empty Lands*, 3rd ed. (1954), *Population and Planned Parenthood in India*, 2nd ed. (1961), *Communist China Today*, 4th rev. ed. (1964), *Asia's Population Problems* (1967), and *Infant Mortality, Population Growth and Family Planning in India* (1972); C.D. COWAN (ed.), *The Economic Development of Southeast Asia* (1964); KINGSLEY DAVIS, *The Population of India and Pakistan* (1951); PHILIP M. HAUSER (ed.), *Urbanization in Asia and the Far East* (1957); PING-TI HO, *Studies on the Population of China, 1368–1953* (1959); JACQUES M. MAY, *The Ecology of Malnutrition in the Far and Near East* (1961); POLITICAL AND ECONOMIC PLANNING (PEP), *World Population and Resources* (1955); JOHN ROBBINS, *Too Many Asians* (1959); EDGAR SNOW, *The Other Side of the River: Red China Today* (1962); JOSEPH E. SPENCER, *Asia, East by South: A Cultural Geography* (1954); MORRIS B. ULLMAN, *Cities of Mainland China: 1953 and 1958* (1960); United Nations, DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, *Report on the World Social Situation* (1957–), and *The Future Growth of World Population* (1958); and C.K. YANG, *A Chinese Village in Early Communist Transition* (1959).

(C.V.N./P.N.K./Y.K.Y./P.Gu./L.F.deB./Jo.E.S.)

Asia Minor, Religions of

The religions of Asia Minor consist of the beliefs and practices of the ancient peoples and civilizations of the Anatolian Peninsula (modern Turkey), Soviet Armenia, Syria, and north Mesopotamia, including the Hattians, Hittites, Hurrians, Assyrian colonists, Urartians, Phrygians, and Luwians.

SOURCES OF KNOWLEDGE

Until comparatively recent times, the pre-Christian religions of Asia Minor were known only through the works of classical writers, supplemented by coins and the monuments reported by travellers. For the Greeks and Romans, Asia Minor was above all the home of the religion of Cybele, the Great Mother of the cult centred in Phrygian Pessinus. A monument such as the colossal, but much weathered, figure of a Hittite goddess carved high up on the slopes of Mt. Sipylus was of necessity ascribed by the 2nd-century-AD Greek traveller and geographer Pausanias to the Mother of the Gods (a title of Cybele), since no other ancient Anatolian goddess was known to him.

In the 19th century a series of inscriptions renewed interest in the ancient civilization of Asia Minor, and in the 20th century systematic excavations provided his-

torians and archaeologists with vital new facts. The discovery of the royal archives of the Hittites at Boğazköy (ancient Hattusas) in 1907 made available for the first time a mass of indigenous literary evidence for an Anatolian civilization, a civilization belonging to the 2nd millennium BC, before the arrival of the Phrygians. Because of the discovery of these clay tablets, the religion of the Hittites necessarily predominates in any account of the religions of Asia Minor. Later Hittite history has been further clarified by the decipherment of the Hittite hieroglyphic inscriptions on monuments dating for the most part from the early centuries of the 1st millennium BC, after the downfall of Hattusas. For the same period, the cuneiform inscriptions of the kingdom of Urartu in the region of Lake Van contain some information on the religion of that area, though they are mostly concerned with other matters.

The Hittite records at Boğazköy give abundant evidence for a state religious cult. Religious ideals, for instance, are revealed in the prayers offered by various members of royalty and all important state matters, including royal decrees and treaties, were referred to certain deities. The clay tablets testify to a unification of religion and the state, for the state and monarchy were placed under the protection of national deities at the ancient capital of Hattusas.

The clay tablets of Assyrian commercial colonists found at Kültepe, Alişar, and Boğazköy belong to the period immediately preceding the rise of the kingdom of Hattusas, but they contain little information bearing on the life of the indigenous population. For all earlier periods, scholars are dependent on the inarticulate data of archaeology—isolated finds, the interpretation of which leaves a large element of uncertainty.

PREHISTORIC PERIODS

The earliest evidence of religious beliefs has come to light at the mound of Çatal Hüyük, to the south of modern Konya. Here in four seasons of excavations (1961–65), James Mellaart discovered remains of a Neolithic village of mud-brick houses, many of which could be identified as shrines. They are dated by radiocarbon to *c.* 6500–5800 BC (calculated with a half-life of 5,730). Huge figures of goddesses in the posture of giving birth, leopards, and the heads of bulls and rams are modelled in high relief on the walls of some of these shrines. Others contain frescoes showing elaborate scenes such as the hunting of deer and aurochs, or vultures devouring headless human corpses. A series of stone and terracotta statuettes found in these shrines represent a female figure, sometimes accompanied by leopards and, from the earlier levels of excavation, a male either bearded and seated on a bull or youthful and riding a leopard. The main deity of these Neolithic people was evidently a goddess, a mistress of animals, with whom were associated both a son and a consort. Her character is vividly shown by a schist plaque carved to represent two scenes, a sacred marriage and a mother with child. The dead appear to have been excarnated in a mortuary outside the village by exposure to vultures, as shown in the painting, before being buried under the platforms in the houses.

Cults of goddesses

At Hacilar, near Lake Burdur, a somewhat later culture was unearthed by the same excavator, and here again were found statuettes of goddesses associated with felines; but, as in the later levels at Çatal Hüyük, the son or consort is absent.

Entirely different and far removed in time and place are the discoveries at Alaca Hüyük and Horoztepe in northern Anatolia. Here, dating from the latter half of the 3rd millennium BC (*c.* 2400–2200), were found royal tombs richly furnished with artifacts in bronze and precious metals. Beside the heads of skeletons lay female figurines; one such figure found in a grave at Horoztepe represents a mother nursing her child. Many of the objects found in these graves must have had ritual significance. At Horoztepe a bronze sistrum, or rattle, was found. But the outstanding feature of the graves at both sites is the occurrence of bronze standards, which may have been carried on poles. They are open-work objects of circular or occasionally rhomboid form and are adorned with figures of animals (bulls, stags, and, in one instance, felines), birds, flowers, and swastikas and other geometrical patterns. Other standards, consisting of simple statuettes of stags or bulls, also occur.

The archaeological finds of central Anatolia follow immediately after the period of these royal tombs from the Pontic region. Kültepe, near Kayseri, became in the 19th century BC the centre of the Assyrian trading outposts (*kārum*) already mentioned; but from the mound itself, from a level just prior to the foundation of the Assyrian colonies, have come a series of remarkable statuettes. The majority of these are abstract, disk-shaped idols without limbs; many of them have two, three, or even four heads, and others bear on their chests small male figures in relief, in one case accompanied by a lion. There can be little doubt that here again is a representation of a divine family—a mother goddess with consort and child or children. From a level at Boğazköy contemporary with Kültepe comes a limestone mold of a "mistress of animals," a nude goddess standing on a pair of felines and holding aloft an animal in either hand. Molds for a pair of figures, a bearded god and a goddess—the god carries various weapons or emblems, the goddess in most instances holds a baby—have been found at several sites at a somewhat later level.

Though the Old Assyrian tablets are concerned exclusively with commercial matters, the seal impressions that they bear contain a new and elaborate system of religious symbolism (iconography) that later reached its maturity under the Hittites. Here a whole pantheon of deities, some recognizably Mesopotamian, others native Anatolian, are distinguished by such features as dress, attendant animals, weapons, actions, and attitudes. Among them are several weather gods, all associated with a bull, but distinguished in various ways; the weather is depicted in the form of rain falling above the god. A bull alone, carrying an enigmatic pyramid upon its back, sometimes surmounted by a bird, is a particularly common motif and probably symbolizes a weather god. Other deities are a war god holding various weapons, a hunting god holding a bird or hare, a god in a horse-drawn chariot, another in a wagon drawn by boars, a goddess enthroned and surrounded by animals, a nude goddess, and several composite beings. On many seals the deity—and especially the bull with the pyramid—are shown receiving ritual offerings.

RELIGIONS OF THE HITTITES, HATTIANS, AND HURRIANS

An interval of only a few decades separates the end of the Assyrian colony period from the earliest records of the kingdom of Hatti, and for the next five centuries (*c.* 1700–1200 BC) the history of Asia Minor is well documented. The texts reveal a country inhabited by a number of distinct peoples. The Hittites in the centre, the Luwians in the south and west, and the Palaians in the north were speakers of related Indo-European languages. In the southeast were the Hurrians, comparatively late arrivals from the region of Lake Urmia. The Hattians, whose language appears to have become extinct, were most probably the earliest inhabitants of the kingdom of Hatti itself.

Each of these nations had its own pantheon, and individual cult centres had their own names for deities. The result is a bewildering number of divine names, and even when a deity is denoted not by a name but by a logogram (sign or signs standing for a word) to indicate weather god, sun god, moon god, etc., it seems that the deity of each city was regarded by the Hittite theologians as a distinct personality. There are even special weather gods, such as the weather god of the lightning, the weather god of the clouds, the weather god of the rain, the weather god of the palace, the weather god of the royal person, the weather god of the sceptre, and the weather god of the army, each again conceived as a separate personality. These were probably only manifestations or

Types of gods

aspects of a single deity, and this is reflected to some extent in the iconography, the pattern of religious symbolism, in which, as in the preceding period, there is a well-defined and limited number of divine types. Shrines are distributed widely throughout the country.

The pantheon. The most widely worshipped deity of Hittite Anatolia was clearly the weather god, as befits a country dependent on rain for its fertility; and under the title "weather god of Hatti" he became the chief deity of the official pantheon, a great figure who bestowed kingship, brought victory in war, and probably represented the nation in its dealings with foreign powers. Thus the treaty with Egypt is said to be "for the purpose of making eternal the relations which the sun-god [of Egypt] and the weather-god [of Hatti] have established for the Land of Egypt and the Land of Hatti." His name in Luwian, and probably also in Hittite, was Tarhun (Tarhund); in Hattic he was called Taru, and in Hurrian, Teshub. He is associated with the sacred bull and appears on monuments either attended by a pair of divine bulls or driving over mountains in a chariot drawn by bulls. In the cult itself Tarhun may even have been represented by a bull. Often, deities were represented by a symbol on clubs and other weapons. An example is the rock carving of a sword deity in Yazılıkaya (Inscribed Rock) near Boğazköy. A human head tops the hilt, which is carved in the form of four crouching lions.

As Tarhun's spouse, the great goddess of the city of Arinna was exalted as patroness of the state. (Arinna has not been located, but it was situated somewhere in the heartland of the Hittite kingdom, within a day's journey of the capital.) Her name in Hattic was Wurusemu, but the Hittites worshipped her under the epithet Arinnitti. She is always called a sun goddess, and sun disks appear as emblems in her cult, but there are indications that she may originally have had chthonic, or underworld, characteristics. As "sun goddess of the earth" she might be identified with Lelwani, the ruler of the netherworld. The king and queen were her high priest and priestess.

The weather god of another city, Nerik, was regarded as the son of this supreme pair, and they had daughters named Mezzulla and Hulla and a granddaughter, Zintuhi. Telipinu was another son of the weather god and had similar attributes. He was a central figure in the Hittite myths.

Solar and lunar deities

There was also a male sun god, distinct from the sun goddess of Arinna, a special form of whom was the "sun god in the water," probably the sun as reflected in the waters of a lake. His name in Hittite was Istanu, borrowed from the Hattic Estan (Luwian Tiwat, Hurrian Shimegi). There was also a moon god (Hittite and Luwian Arma, Hurrian Kushuh), but he plays little part in the texts. In the iconography, the sun god was represented in the robes of the king, whose title was "My Sun"; the moon god was shown as a winged figure with a crescent on his helmet, sometimes standing on a lion. According to official theology there also existed a sun god or goddess of the underworld. In this place resided the sun on its journey from west to east during the night.

The god who is known from the Kültepe seals as the god of hunting appears frequently on Hittite monuments; he holds a bird and a hare, as at Kültepe, and he stands on a stag as his sacred animal. From descriptions of the statues it appears that this is the deity denoted in the texts by the logogram KAL, perhaps to be read Tuwata, later Ruwata, Runda. The war god also appears, though his Hittite name is concealed behind the logographic name ZABABA, the name of the Mesopotamian war god. His Hattic name was Wurunkatti, his Hurrian counterpart Astabi. His Hattic name meant "King of the land."

The Hittite goddess of love and war is similarly disguised under the logogram of the Babylonian ISHTAR; she was evidently much revered and was the special protectress of Hattusilis III. Her Hurrian name was Shaushka. As a warrior goddess she was represented as a winged figure standing on a lion with a peculiar robe gathered at the knees and accompanied by doves and two female attendants.

There was a mother goddess, Hannahanna "the grandmother," closely associated with birth, creation, and destiny, but the theologians appear to have regarded her as a minor deity.

It is impossible to enumerate the lesser deities, many of whom are mere names to scholars. Among them were deities of many mountains, rivers, and springs, and the spirits of past kings and queens who had "become gods" at death. Demons are conspicuous by their absence; sickness and misfortune were ascribed either to sorcery or to divine retribution.

During the later years of the Hittite kingdom, the state cult came under strong Hurrian influence. The sun goddess of Arinna and the weather god of Nerik were identified with the Hurrian queen of the gods, Hebat, and her son, Sharruma; and at a holy place near the capital (now named Yazilikaya), where a rocky outcrop forming a natural open chamber was adorned with a series of 64 bas-reliefs that represented the national pantheon, every identifiable deity bears a Hurrian name, written in Hittite hieroglyphs. The central group is recognizable as the family of the sun goddess, but she is named Hepatu, her son Sharruma. They both stand on felines, she, perhaps, on a lion or lioness, and he on a panther. The Hittites had here already begun a process of assimilation.

Gods and men. The gods were imagined to have their own lives, though also needing the service of their worshippers, who in turn were dependent on the gods for their well-being. They lived in their temples, where they had to be fed, clothed, washed, and entertained. Part of their time, however, might be spent in heaven or in roaming the sea or the mountains. They might withdraw in anger and so cause life on earth to wither and cease. One of the most characteristic rituals of the Hittites was the invocation by which a god who had absented himself was induced to return and attend to his duties by a combination of prayer and magic.

The relation between man and god resembled that between servant and master. "If a servant has committed an offence and confesses his guilt before his master, his master may do with him whatever he pleases; but because he has confessed his guilt . . . his master's spirit is appeased and he will not call that servant to account." Confession and expiation form the main theme of the extant royal prayers.

Divination. Divination, through which the cause of divine displeasure was ascertained, was of three kinds: augury (divination by flight of birds), haruspicy (divination by examining the entrails of sacred animals), and dice throwing, arts said to be practiced respectively by the "bird-watcher," the seer, and the "old woman." The omens, as interpreted by these experts, were either favourable or unfavourable, and would give a yes or no answer according to the sense of the question put to them. In this way, by a lengthy process of elimination, it was possible to determine the precise offence that required expiation. Divination was a science inherited by the Hittites from the Babylonian seers. Signs of the peoples' fate were thought to be sent by the gods, manifested in unusual occurrences. Haruspicy, as noted, was one of the most popular practices of divination. The liver and viscera of the sacrificial victim were examined, and according to their configuration it was decided whether the omen was favourable or unfavourable. Records of these practices have survived in large numbers, but they are also among the worst written of the Hittite tablets.

Sacrifices and festivals

The cult. The proper conduct for temple personnel was laid down in a tablet of instructions that gives some insight into the organization of a temple. Divine vengeance is threatened against those who misappropriate food or drink brought for sacrifice, who admit unclean animals or unauthorized persons into the temples, who purloin vessels or implements belonging to the god, who

fail to celebrate festivals at the proper time, and who desert their posts to spend the night with their wives.

Many extant texts consist of descriptions of festivals in which the king or queen is the chief officiant. These festivals were numerous, but their names are largely unintelligible. Many of them were seasonal. The preliminary details, such as the robing of the king and his entry into the temple, accompanied by various dignitaries and by musicians playing their instruments, differed little from one festival to another. Owing to the very large number of fragmentary texts, it has not yet been possible to discern special characteristics of the festivals. They invariably culminated in libations and frequently in a cultic meal. One such festival lasted 38 days and involved celebrations in a dozen different cities.

Burial customs. The tablets from Boğazköy have yielded much information about the burial practices of the Hittites. One tablet tells of a burial ritual for a king or queen that lasted 13 days and in which the body was cremated. In the usual Hittite fashion, the body was initially burned and the fire extinguished with potable liquids. The bones were then dipped in oil or fat and wrapped in cloth. A feast followed their placement on a stool in a stone chamber. Although cremations were practiced to a great extent, burial of the body in an earthen grave was not uncommon. In 1952 Kurt Bittel excavated a site near Yazılıkaya close to a natural rock outcrop. The site contained 72 burials, 50 of which were cremations. There was no indication to show that cremations were exclusively the right of the elite.

Mythology. In Anatolia itself myth seems to have remained on a rather primitive level and is mainly to be found embedded in magical or ritual texts. Writings of this type constituted a large portion of Hittite literature and indicate the prevalance of both black and white magic. Myths were consequently associated with magical rituals aimed at curing diseases, ensuring good fortune, dispersing evil spirits, and the like.

A particularly well-attested type of myth occurs in connection with the invocation of an absent god and tells how the god once disappeared and caused a blight on earth, how he was sought and found, and eventually returned to restore life and vigour. In one such myth the weather god withdraws in anger and the search is conducted by the sun god (whose messenger is an eagle), the father of the weather god, his grandfather, and his grandmother Hannahanna. In another, it is Telipinu who is angry, and the gods who search are the sun god, the weather god, and Hannahanna, the grandfather being omitted. In both these versions, the missing god is found by a bee sent forth by Hannahanna. In another similar story, the sun god and Telipinu are both missing, not from anger, but because they have been seized by "Torpor," which has paralyzed nature. In yet another version, the weather god of Nerik is said to have gone down to the netherworld through a hole in the ground, apparently the hole from which the river Marassantiya (modern Kızıl Irmak) gushed forth, which suggests that this weather god may really have been a god of the underground waters.

Another myth, the "Slaying of the Dragon," connected with the Hattian city Nerik, was apparently recited at a great annual Spring festival called Purulli. It tells how the weather god fought the dragon and was at first defeated, but subsequently, by means of a ruse (of which there are two quite distinct versions), succeeded in getting the better of him and killing him. The ritual associated with this tale has not been identified but its primitive character establishes it as folklore.

Other mythological tales of Hittite and Hattian deities existed, but they are too fragmentarily preserved to give any connected story.

The elaborate epic of the struggle against Ullikummi, and the *Theogony*, though written in Hittite, are Hurrian in origin and refer to Hurrian and even Mesopotamian deities. The *Theogony* tells of the struggle for kingship among the gods. Alalu, after holding the kingship for nine years, was defeated by Anu (the Babylonian sky god) and went down to the netherworld. Anu in his turn, after nine years, gave way to Kumarbi, a Hurrian god, and went up to heaven. Eventually the weather god Teshub was born, and though the god KAL apparently reigned for a period, and the end of the tale is lost, it is certain that Teshub was the final victor, for there are many allusions to the "former gods" who were banished to the netherworld by him. The conception itself derives from Babylonia.

The "Song of Ullikummi" tells of a plot by Kumarbi to depose Teshub from his supremacy by begetting a monstrous stone as champion. Ullikummi, the stone monster, grows in the sea, which reaches his waist, while his head touches the sky; he stands on the shoulder of Upelluri, an Atlas figure who carries heaven and earth. Teshub is warned of the danger and goes out to battle in his chariot drawn by bulls, but he fails and appeals for help to Ea (Babylonian god of wisdom). The latter orders the "former gods" to produce the ancient tool by which heaven and earth had once been cut apart (the only surviving hint of a Hittite creation myth), and with this he severs Ullikummi from the giant and so destroys his power. Again the end is lost, but it is certain that the final victory went to Teshub.

RELIGIONS OF SUCCESSOR STATES

When Hattusas fell, *c.* 1180 BC, the Luwians moved eastward and southward into Cappadocia, Cilicia, and North Syria. Here they formed a number of small successor kingdoms. Shortly afterward the Phrygians crossed the Bosporus from Thrace and occupied the centre of the Anatolian plateau, cutting off in the extreme southwest a remnant of the Luwian people, who became known as the Lycians and maintained their reverence for the Luwian gods Tarhun, Runda, Arma, and Santa into classical times.

The East Luwians, whose rulers used the Hittite hieroglyphic script to record their deeds, worshipped these same deities; but their chief goddess was Kubaba, who hardly appears in the archives of Hattusas except as the local goddess of Carchemish in Syria. Her prominence was due to political factors, for Carchemish was then the leading Hittite city.

The traditional Hittite iconography survived, but was gradually permeated by Aramaic and Assyrian influences. Orthostats (stone slabs set at the base of a wall) from Malatya on the Euphrates show Tarhun in his bull-drawn chariot receiving libations from a king dressed in his traditional robes, and there is a relief showing his battle with the dragon. At Carchemish was found a representation of the winged moon god with the sun god, both standing on a single lion. Kubaba on a stela appears enthroned, the throne resting on a lion. Runda (the Hittite Tuwata or KAL) is regularly symbolized by a stag's head or antler.

Urartu. In the far east of Anatolia, the Hurrian nation formed around Lake Van a new kingdom, which rose to considerable power, *c.* 900–600 BC. With few exceptions, the cuneiform inscriptions of this kingdom of Urartu are historical and reveal nothing of its religion, except the names of deities. The national god was Haldi, and he is associated with a weather god, Tesheba, a sun goddess, Shiwini (compare Hurrian Teshub and Shimegi), and a goddess, Bagbartu (or Bagmashtu). Haldi is represented standing on a lion, Tesheba on a bull, Shiwini as a goddess holding a winged sun disk above her head. The cult was practiced not only in temples (one of which is shown in detail on an Assyrian relief) but also in front of rock-hewn niches in the form of gates through which the deity was probably believed to manifest himself.

The Phrygians. Little would be known of the religion of the Phrygians but for the fact that in 204 BC the Roman Senate, on the instructions of the priests, who had consulted the Sibylline books, had the sacred black stone of the Phrygian mother goddess, Cybele, or Cybebe, transported from Pessinus, together with her priests, and installed in a temple on the Palatine. As a

result, there is much information about the cult and its mythology, though it must be remembered that during 200 years of Persian rule Anatolia had been exposed to many alien influences from the east, which may have affected this cult.

The high priest of Cybele was given the name of Attis, and—at least in later times—she was attended by a band of fanatical devotees called *galli*, whose orgiastic dancing, at the climax of which they castrated themselves in their ecstasy, was notorious.

The cult myth of these rites told how Cybele (known at Pessinus as Agdistis, from Mt. Agdos in the vicinity) loved a beautiful youth named Attis. According to the earliest version, Attis was killed, as was the Syrian Adonis, by a boar. All later versions, however, refer to wild revelry and castration. Agdistis is a bisexual monster who is trapped by Bacchus and castrated; Attis is betrothed to a daughter of Midas (or Gallus, the king of Pessinus); the wedding guests are driven mad by Cybele, and first Gallus, then Attis, castrates himself, the latter as he lies beneath a pine tree; in one version Attis is turned into a pine tree. A poem by Catullus describes how a young Greek wanderer named Attis was caught up in the revels and sacrificed his virility, only to be prostrated later with remorse. The "Phrygian rites" introduced into Rome by Claudius included the ceremonious felling of a pine tree to represent the dead youth and its transport in procession to the temple. Still later, the taurobolium (sacrifice of a bull) and the belief in the resurrection of Attis were added to the cult.

How much of this myth belonged to the original cult of the Phrygian mother goddess is questionable. Herodotus, in describing the celebration of the rites by the legendary Scythian sage Anacharsis, mentions only that he did so in a grove, that he carried a timbrel (a small hand drum or tambourine), and fastened images about his person. There is no suggestion of orgiastic rites.

In Asia Minor itself, the cult of Cybele is marked by carved rock facades with niches or by rock-hewn thrones, on which the statue would be set; in front of these, the rites were celebrated in the open air. Cybele was a goddess of the mountains, out of which she was believed to manifest herself to her devotees. Representations of the goddess show her in her niche, sometimes flanked by lions, draped in a long garment and wearing a high polos (cylindrical crown or headdress) or with bared breasts and flanked by musicians. Her name and her association with the lion cannot be separated from the Hittite Kubaba, whose cult had spread from Carchemish to the borders of Phrygia; but the process by which this matronly figure was transformed into the Mountain Mother of the Phrygians can only be surmised.

The goddess Ma of Comana, despite her name (Mother), was regarded at least by the Romans as a deity distinct from Cybele and identified with the war goddess Bellona. Her relationship to the ancient Hittite-Hurrian goddess Hebat of Kummanni (=Comana) remains obscure, for there is no evidence that the latter was a goddess of war.

The god Men, who appears on numerous monuments of the Hellenistic period, was an equestrian moon god, later identified with Attis and with the Thracian Sabazius. He is basically the Persian moon god Mao, as (Artemis) Anaitis is the Persian Anahita.

SUMMARY

Asia Minor shows a remarkable continuity in its worship. From the Neolithic Age, for 6,000 years, the population venerated a divine pair, mother goddess and weather god, the former in association with the lion, the latter with the bull; a divine son, associated with the panther; and a god of hunting whose symbolic animal was the stag. To the ancients, for whom the essence of a thing lay in its name, this continuity was less obvious than it is today. The many names under which the deities were known at different times and places appear to us of less significance, in a religious sense, than the constancy of the types.

BIBLIOGRAPHY. A. GOETZE, *Kleinasien*, in I. VON MULLER, *Handbuch der Altertumswissenschaft*, 2nd ed., vol. 3, pt. 1, sect. 3 (1957), a classic work covering all periods, and translations of Hittite texts in J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament*, 3rd ed. (1969); E. AKURGAL and M. HIRMER, *The Art of the Hittites* (1962), an excellent presentation of Hittite and pre-Hittite art and iconography; M. VIEYRA, *Hittite Art, 2300–750 B.C.* (1955); SETON LLOYD, *Early Highland Peoples of Anatolia* (1967), a popular but excellent account of all periods, with many illustrations; O.R. GURNEY, *The Hittites*, rev. ed. (1961), a general description of Hittite civilization, and "Hittite Kingship," in S.H. HOOKE (ed.), *Myth, Ritual and Kingship* (1958); H.G. GUTERBOCK, "Hittite Religion," in V. FERM (ed.), *Forgotten Religions* (1950), and "Hittite Mythology," in S.N. KRAMER (ed.), *Mythologies of the Ancient World* (1961); J. MELLAART, *Catal Hüyük* (1967), an illustrated account of the evidence from Çatal Hüyük and Hacilar; N. OZGUC, *The Anatolian Group of Cylinder Seal Impressions from Kültepe* (1965), an important publication of new discoveries; F. CUMONT, *The Oriental Religions in Roman Paganism* (1911), a classic; E.N. LANE, "A Re-Study of the God Men," in *Berytus*, vol. 17 (1968), a valuable summary of recent work on Men and Cybele; R.D. BARNETT, "Phrygia and the Peoples of Anatolia in the Iron Age," *Cambridge Ancient History*, rev. ed., vol. 2, ch. 30 (1967).

(O.R.G.)

Asian Peoples and Cultures

Asia, a geographical term defining a major continent, is characterized by too little or too much: too little precipitation or too much; too many people or too few; too much central government control, or not enough to foster grass roots development; overindulgence or undernourishment; Playboy clubs coupled with poverty; Nobel Prize winners and massive illiteracy.

Asian diversity

Asia contains most physical varieties of modern man: Mongoloids, Caucasoids, Australoids, blacks, Negritos, and mixtures of all these.

Linguistically, Asia also embraces most major non-African language families: Indo-European, Uralic-Altaic, Sino-Tibetan, Austro-Asiatic (including Munda, or Kolarian), Dravidian, Tai-Kadai, Malayo-Polynesian.

Buddhism, Confucianism, Hinduism, Islām, Judaism, and Christianity have competed ideologically and politically. Shamanism, a belief in demons and ancestral spirits reached by the medium of diviners called shamans, and also localized nature worship of various types overlie the major religions, partly because widespread illiteracy among the people denies them access to religious literature.

Geography and ecology vary as much as the people, ranging from the highest mountains on Earth to the almost sea-level deltas of South and Southeast Asia, and from the Central Asian steppes and deserts to the north, where forest steppelands blend into taiga.

ASIAN CULTURAL TYPES

Common characteristics. All Asian nations have moved toward some sort of industrialization or economic diversification, but most peoples live at agricultural subsistence levels in peasant and tribal milieus. Such societies have attributes and attitudes 180° from those of the pluralistic, industrialized, highly specialized, multi-institutional societies of the West. Five major factors deserve consideration:

1. Illiteracy and linguistic diversity. Most Asians cannot read and write, and literacy constitutes an important tool in any people's cultural tool kit. Even if all Asians could read and write, linguistic diversity would still foster divisiveness. Usually, language and ethnic group tend to coincide, and regional linguistic affiliations are major cohesive factors.

2. Basic food production. Most Asians spend most of their time engaged in basic food production, either agriculture or herding, or combinations of the two. An interesting correlation appears to exist between the percentage of illiterates and the percentage of agriculturists and herders in Asian societies.

3. Lack of mobility. Emphasis on the group rather than on the individual results in a relative lack of social, eco-

nomic, and political mobility in Asian societies, the most intensive manifestation being the Hindu caste system.

Rural orientation

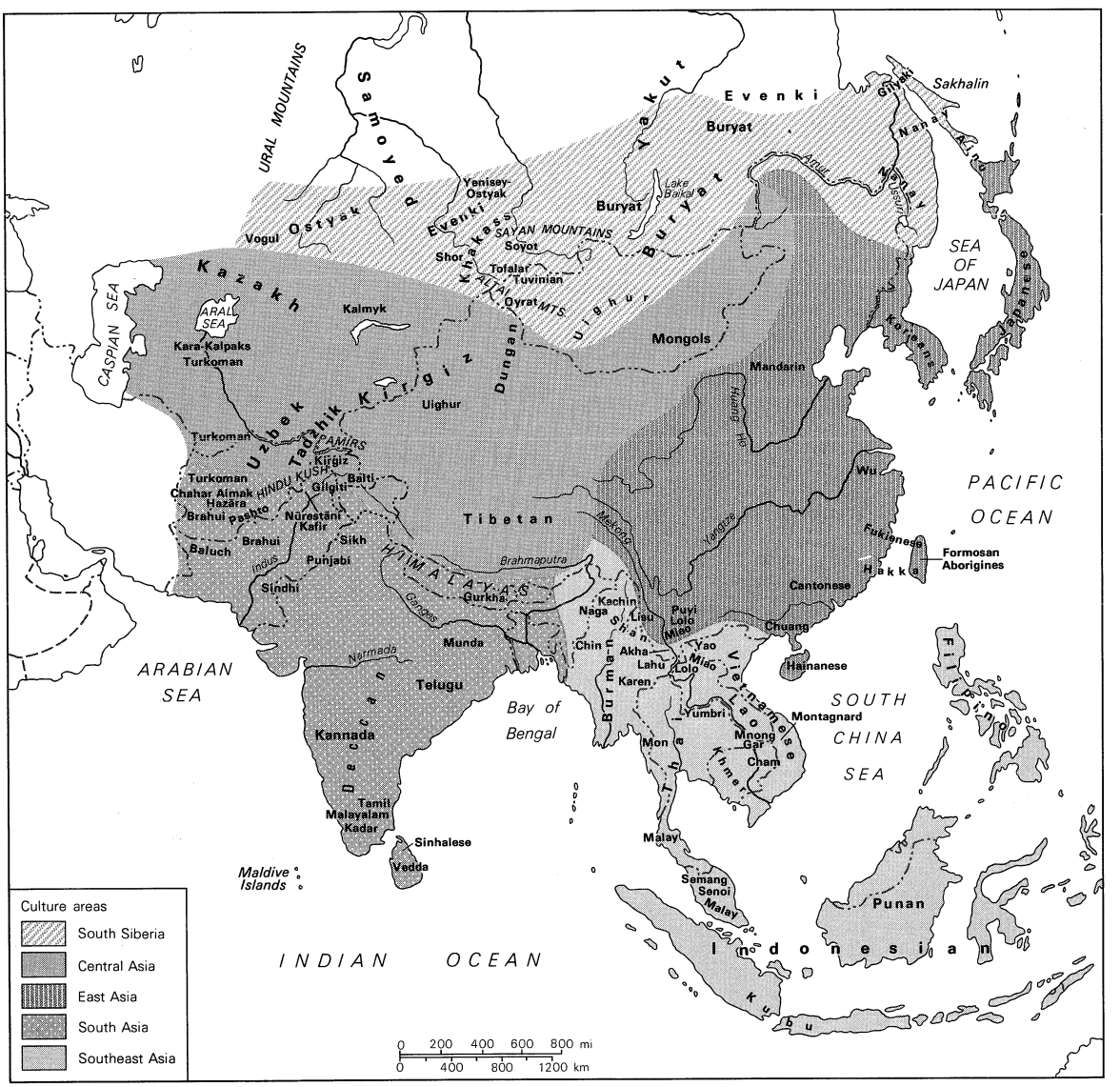
Endogamous marriage patterns dominate—that is, persons marry strictly within their group or class. Even exogamous peoples—peoples requiring marriage outside the group—usually select mates within a tightly defined set of alternatives. Economically, men and women usually follow the occupations of their parents; politically, people are born into statuses of leaders or followers (only on rare occasions have peasants become kings and vice versa). Lack of mobility helps hold people to the land and perpetuates existing leadership patterns, both essential in the survival economic patterns of Asia—in contrast to the surplus economic situation in the West. When the individual (or his family group) moves to the city, he often remains rural oriented, even though urban based. Even nomads, in spite of physical migration, lack fundamental mobility, for they move along the same routes from year to year.

4. Absence of adolescence. In Western societies, adolescence is primarily a time of learning, away from the family, in order to prepare the individual for entrance into an ultraspecialized, multi-institutional situation. In Asian society, however, the individual usually lives in a generalized, uni-institutional society: the extended family and other structured, hierarchal tribal units. Socialization of the Asian child takes place inside the family, and the child becomes an adult almost overnight, with a full range of social, economic, and political responsibilities. While the father–mother generation (the present) forms the economic unit, the grandparents (the past) teach the children (the future) about their history, customs, and values. Economic roles are usually learned from parents and older brothers and sisters. The gradual introduction of formal educational systems based on literacy has affected these patterns in varying degrees, however, and young urban literates now often reject the idea of having respect for age.

5. Kinship in place of government. Few central governments in Asia can replace the network of reciprocal social, economic, and political rights and obligations that are currently vested in kin units. In the past, outsiders have come to the rural area to extract taxes, rents, slaves, conscripts for armies, and women for harems. Peasants continue to resist penetration into their areas, often using passive resistance; tribesmen and nomads usually resist violently.

The concept of ecological time also forms an important pattern, in which people adjust to the annual cycle of nature unencumbered by the artificial time units established to control and limit Western economic and noneconomic activities.

All the attributes sketchily discussed above tend to perpetuate an "inward-looking" society in which men and women are born into a "set of answers." In the pluralistic, ultraspecialized, developed world of the West, men and women are born into a "set of questions," or an "outward-looking" society.



Distribution of Asian peoples and cultures.

General settlement patterns. Among other factors, ecology, technology, and historical accident influence settlement patterns and crystallize Asian peoples into two general categories, sedentary and nonsedentary. Interactional patterns include linked networks between sedentary villages, towns, and cities and their relationships with nomads, semi-nomads, and semi-sedentary groups. Numbers of peoples in Asian Siberia still support themselves by hunting, trapping, fishing, and logging, travelling with the season. Fishing villages abound in all Asian areas having access to such large bodies of water as oceans, lakes, rivers, and inlets. Pockets of hunters and gatherers still roam the hills and forests of South and Southeast Asia.

Sedentary patterns

Two sedentary patterns exist: the linear type is common wherever villages and towns string out along major rivers that usually serve as highways; the nuclear type is more common in drier and mountainous areas and grew in response to needs for water and defense, with several villages clustering about a town and several clusters of village–town nuclei hovering around cities.

Villages are usually self-sufficient subsistence units but obtain many commodities and all luxuries from either town bazaars, travelling peddlers, or nomadic traders. (Some villages, though, will have small bazaars catering to local, daily needs.)

Towns generally occur wherever several major trails intersect, usually near a large river, oasis, or other major source of water. For the villages, the town usually constitutes the commercial, administrative, and communications centre. Often villagers bring raw materials, local craft products, and agricultural produce to town bazaars by donkey, horse, camel, or human back, or by boat if on a waterway. Goods are shipped from the town by truck, camel, oxcart, or donkey or, again, by boat if waterways form a major part of the transportation network.

The town bazaar (often held only one or two days a week) also acts as a disseminator of news, but cheap transistor radios have brought about a revolution in communications. Most villages have at least one, and even nomads move from pastureland to pastureland with battery-powered radios blasting away.

Full-time specialists in the towns (relatively few exist in villages), such as ironworkers, carpenters, weavers, dyers, tailors, shoemakers, masons, general storekeepers, butchers, bakers, bicycle repairmen, automotive repairmen, and caravansary owners, generally live above their shops (or near them) in the bazaar; in other words, a unity of occupation and residence is the pattern for most Asian towns. Also in the towns are the headquarters of lower grade civil servants and quasi-military police, representing the state. Often absentee landowners live in the town.

Where main commercial routes have met and permitted easy access to the outside world, cities have sprung up. The city, therefore, is a major commercial, administrative, and communications centre, linking the interior with the outside world. Large guildlike groups of specialists live in separate sections of the bazaar in the "old city." Individual artisans—and, more recently, factories—produce items for export and for sale in the town bazaars. In addition, more and more groups migrate seasonally to towns and cities to take advantage of new job opportunities brought about by post-World War II development processes. Some remain in the cities to make Asia's urban problems annually grow more acute.

To handle the growing administrative problems, "new cities" spring up alongside the "old," often separated by a river. Government officials, the growing entrepreneurial class, foreign diplomats, and development technicians usually live and work in the new city, while the bazaar and the bulk of the population remain in the old. Elected (or appointed) parliaments sit in the capitals and sometimes the provincial centres, and courts administer justice based on myriad legal codes. Most Asian peoples, however, prefer to have as little contact with government officials as possible.

Sedentary farmers are agriculturists who live permanently in the same village and leave only when forced to or in order to accept a better farming opportunity. Semi-sedentary farmers are either agriculturists who own enough livestock to be moved in the summer to highland pastures or agriculturists who move to the fields at harvest time and live in portable huts.

Nomadism

Nomads are herdsmen who move as a group from summer to winter pasturages and back again. Steppe and desert nomads often move great horizontal distances; nomads near mountains usually move vertically to alpine grasslands in the summer and plains grasslands in the winter. Few nomads move constantly—only seasonally.

Semi-nomads are herdsmen who practice some agriculture. A sizable part (more than 50 percent) of a semi-nomadic group will move with the livestock to summer pasturages, while the remainder tend crops in the winter headquarters. Their movements, primarily vertical, usually cover much shorter distances than those of true nomads. (Seasonal nomadism or semi-nomadism, it may be noted, is also called transhumance.)

Throughout history, competition for land between sedentary and nonsedentary peoples has caused nomads, voluntarily or involuntarily, to evolve into semi-nomads or semi-sedentary peoples but seldom fully sedentary unless under duress. Most nomads and semi-nomads have many functional, symbiotic relationships with villagers along their routes of migration. Livestock furnishes meat, dairy products, and wool, which the nomads often trade for grains, vegetables, fruits, and nuts. Moneylending becomes a major economic activity of wealthier nomads. Even landowning village farmers often need extra cash for such life-cycle ceremonies as birth, circumcision, marriage, and death: nomads happily lend money at exorbitant rates. Some farmers also purchase trade goods from nomads on credit. Unable to repay, the farmer sometimes loses his land and becomes a tenant to the nomads, who then collect annual rents as they pass through the village.

Often, nomadic groups have traditional grazing rights in lands owned by specific villages. While grazing on the grain stubble, their flocks deposit vast quantities of dung, later plowed under to replace nitrogen lost in the soil. Animal dung remains a primary source of fuel, as well as fertilizer. In the Soviet Union, however, planners in the 1930s considered the nomads parasites and either collectivized or eliminated them, thus in the process eliminating the natural fertilization. This is one reason why the attempted cultivation and exploitation of the "virgin lands" of Kazakhstan has yet to succeed.

Intelligently conceived, vigorously implemented, long-range programs to supplement land reclamation and improved agricultural practices would benefit both nomads and villagers, but the elimination of the economically valuable—and maligned—goat would probably not improve the situation.

ASIAN CULTURE AREAS

Within each area of Asia, different peoples and their patterns can be identified, described, and classified together in many ways. The following represents simply one of many useful classifications.

Southern Siberia. Located south of the circumpolar culture area and east of the Ural Mountains, southern Siberia lies in a transitional area in which tundra or treeless plain becomes taiga or swampy coniferous forest. Farther south, the taiga blends into a forest steppe with alternating stands of trees and open, seasonal grasslands. Summers, except in the extreme south, are too short to permit agriculture; spring floods often inundate the taiga. Farther north are the Altai (Kirgiz) and Sayan (Kazakh) mountains, which abut Central Asia in the west and the Mongolian People's Republic in the east. Farther east are the mixed broadleaf forests of the Amur–Ussuri valleys, with hot summers and severe winters. The Pacific maritime coast and Sakhalin Island complete the geographic picture. The northeastern region of China (formerly called Manchuria) sits to the south.

Mongoloids of Siberia

The indigenous populations are primarily Mongoloid racially, although generally they are outnumbered in most areas by Russian, Ukrainian, and Belorussian immigrants. The aborigines still live mainly in separate,

virtually endogamous villages, though many groups (including reindeer herders) have been collectivized by the Soviets. They speak several languages and dialects of the Uralic-Altaic family, among them Samoyedic, Tungusic, Mongolian, and Yakut (or other Turkic dialects).

The main economic pursuits include small-scale reindeer herding (seven to ten reindeer per herd, small when compared to groups farther north), hunting (deer, elk, squirrel), trapping (sable, fox), and fishing. Traditionally, the inhabitants have lived either in longhouses and semi-subterranean huts in permanent villages or in conical, bark-covered tepees at temporary campsites and, for transportation, have used canoes, rowboats, skis and snowshoes, sleds drawn by reindeer or dogs, and, in the east, horses. Many aborigines now work in mines and other Siberian industrial projects.

Most aboriginal groups are small. The 80,000 Evenki (Tungusic-speaking reindeer herders), for example, live and wander in an area about the size of the continental United States.

Central Asia. Central Asia extends from western Manchuria to the Caspian Sea, along the grasslands, mountains, and plateaus of Inner and Outer Mongolia, Tibet, the Central Asian Muslim republics of the U.S.S.R. (Kirgiziya, Tajikistan, Uzbekistan, Kazakhstan, and Turkmenistan), and Afghanistan north of the watershed of the mountain range of the Hindu Kush. An interesting man-made boundary, the Great Wall of China, exists in the east. Altitudes and local ecology vary considerably from region to region and even within regions. The western Turkistan deserts and steppes lie close to sea level, whereas the upland Mongolian steppes often exceed 5,000 feet (1,500 metres) above sea level. Basically, Central Asia is an arid zone that has inland drainage and a continental climate.

Mongoloids of Central Asia

Racially, all Central Asians are Mongoloid variations, more Caucasoid to the west (Slavic) and southwest (Mediterranean subvariant). Most speak Uralic-Altaic languages—usually Turkic dialects but also Mongolian. Other language families include Sino-Tibetan (Tibetan in Tibet and Chinese near the eastern and southern borderlands) and Indo-European (Tajiki and Dari Persian). Cyrillic has replaced the Arabic script in Soviet Central Asia and Mongolian in Mongolia. Inner-Mongolians now use streamlined Chinese characters, but Tibetans, though under great pressure from the Chinese, cling to their own traditional script.

Islām pressed into Central Asia early in the 8th century AD, culminating in the Arab sack of Samarkand in 712. Most people as far east as Outer Mongolia and south into Chinese Sinkiang became Sunnī Muslims, constituting the main orthodox school of Islām, but pockets of Ismā'īlīs, who consider the Aga Khan as their leader, live scattered in such mountainous areas as the Pamirs and the Hindu Kush. Buddhism centring on teachers known as lamas and heavily laden with animistic and shamanistic beliefs and practices dominates Tibet and neighbouring parts of Mongolia and China. Anti-religious drives by the Communists of the U.S.S.R. and China have been only partly successful.

Often called the belt of "pastoral nomadism," Central Asia also includes sizable sedentary agriculturists, usually existing in the loose-knit symbiotic system described earlier, with modifications wherever Communism has introduced collectivization of herds and land. Political ties are largely kin oriented, and a loosely stratified class system exists along with a hereditary nobility. Slavery was common until the Russians and British virtually wiped it out in the 19th century. Mongol nomadic patterns dominate: groups move year-round, more horizontally than vertically, but within well-defined territories; there are comparatively few fixed economic centres (bazaars and towns).

Local ecology greatly influences livestock distribution. Fundamentally, sheep and goats are mountain animals (though sheep prove to be less adaptable than goats and tend to flounder in snow); cattle and camels appear more numerously in transitional forest steppes and semideserts. Yak and yak–bovine hybrids become important in the Pamirs and the higher mountains of Sinkiang, Mongolia, and Tibet. Donkeys and mules are found mainly in the west, where Central Asia nudges the Middle East. Horses, prestige animals throughout Central Asia, are ridden, and sometimes their milk is drunk, often as *kumys* (fermented mare's milk).

South Asia. South Asia includes the territories of southern Afghanistan, Pakistan, India, Nepal, Sikkim, Bhutan, Bangladesh, Sri Lanka (formerly Ceylon), and the Maldive Islands. From the permanent glaciers of the highest mountains on Earth to the lush hilly jungles of Sri Lanka, and from the barren deserts of Afghanistan to the intricate delta networks of the Brahmaputra River of Bangladesh, four major geographic features dominate the South Asian landscape: the Himalayas, the Indo–Gangetic Plains, the Deccan, and the Indian Ocean. All these centre in India, itself divided by the Narmada River into two major ecological north–south zones. The seasonal monsoonal winds play important roles in most South Asian ecosystems, each monsoon having two phases—the northeast monsoon (January to February cold; March to mid-June hot) and the southwest monsoon (mid-June to mid-September rain; mid-September through December ebbing monsoon).

Caucasoid and Mongoloid in South Asia

India is the easternmost wing of the Caucasoid racial region; Mongoloid influences, however, occur in areas bordering Central Asia, especially in central and northern Afghanistan, Pakistan, Kashmir, Nepal, Sikkim, and Bhutan, in northeastern India, and in eastern Bangladesh. Darker skins occur as one moves south and southeast toward India.

In India, particularly in the south, the higher castes usually have lighter skins than do the lower or peasant castes. In the central hills of India live small bands of Negrito-like hunters and gatherers. The few hundred surviving unmixed indigenous hunters and gatherers of Sri Lanka are primarily Caucasoid but with some Australoid characteristics. Some Mongoloid-looking peoples live over 4,000 feet (1,200 metres) above sea level, but basically, the people of Sri Lanka are dark-skinned, slender Caucasoids. Two interesting Caucasoid subvarieties occur in Punjab (India), Nūrestān (Afghanistan), and Chitrāl (Pakistan): the Nūrestāni and the Kafirs of Chitral, who speak Kafiri dialects of Indo-European, have a high frequency of blond characteristics (almost one-third of the population), and the Punjabi-speaking Sikhs are usually deemed the hairiest people in the world.

Four great language families embrace all the dialects spoken in South Asia: Indo-European (mainly north of the Narmada River and across the whole range of South Asia), Dravidian (mainly south of the Narmada River), Sino-Tibetan (on the northeastern borders), and Austro-Asiatic, including Munda, (some hill tribes of India and Bangladesh). In India itself, the geographical and cultural heart of South Asia, over 500,000,000 people speak 14 official languages, divided into about 80 principal dialects and about 550 subdialects. English still functions as the lingua franca for the intelligentsia of India, Pakistan, Bangladesh, and Sri Lanka.

Sanskrit influences most regional languages in India; all with written traditions are written in Devanāgarī or related scripts derived from Sanskrit—except for Tamil, a distinctive script, and Urdu, which uses the Perso-Arabic script, as do the Indo-European languages of Afghanistan and Pakistan.

Hinduism, Islām, and other religions

Hinduism and elements of caste permeate most of the South Asian cultural scene. The Indian government has outlawed caste, but, like racial discrimination elsewhere, caste refuses to be effectively cowed by legislation. Islām, basically an egalitarian religion, entered the Indian subcontinent in the early 8th century AD. Mass conversions to Islām mainly involved "untouchables" and low-caste Hindus; the Muslim elite grew out of the conquering Islāmic armies. Change of religion, however, did not automatically change economic, political, or social status, and caste elements remain strong among Muslim groups in India, Bangladesh, and Pakistani Punjab and Sind. Occupational caste names refer to local groups, and, although men often freely change occupational and politi-

cal affiliations, endogamy is still the pattern, and social status remains largely group oriented.

The several other religions in South Asia include over 10,000,000 Christians scattered in India, Pakistan, and Sri Lanka (about 120,000 are Anglo-Indian), forty percent of them Roman Catholics; about 200,000 urban, commercially and industrially oriented Parsis (Zoroastrians), about half of them living in Bombay and about 90 percent of them literate; about 2,000,000 Jains, also largely in Bombay; and about 10,000,000 Buddhists, almost wholly in Sri Lanka (India, where Buddhism began, has only a few hundred thousand Buddhists).

Although regional, particularly linguistic, units may be politically strong in India, the extended family usually forms the most important socio-economic unit; all contribute to the extended family and are supported by it. Mountaineers and hillsmen, found in the extreme east and west of South Asia and in the central hills of India, are usually tribal. Most rural peoples are governed by village, camp, caste, or tribal councils.

Wheat and rice are the major crops, but barley, millet, maize, gram, tea, coconuts, rubber, jute, cotton, fruits, and nuts are cultivated where soils and climate permit. India has about one-half (not the better half) of the world's cattle, a ratio of 80 cattle for every 100 persons. Most South Asians drink milk and use dairy products; only the Hindus prohibit the consumption of meat. Water buffalo occur in all wetter areas; some are found as far north as northern Afghanistan. Sheep provide many essentials for the peoples of the Northwest Frontier Province and Baluchistan of Pakistan and Afghanistan; half of India's sheep are concentrated in Madras and the southern Deccan, which have insufficient pasturelands to support cattle. Goats are ubiquitous. Donkeys, mules, and camels live mainly in the northwestern regions of South Asia. Elephants, more romantic than visible, occur mainly in the forests and jungles of northern and northeastern India, Bangladesh, and Sri Lanka. Fishing is important everywhere except in the extreme north and northwest.

The Orissa–Bihār–Bengal triangle in east India provides all the ingredients necessary to support large-scale industrial developments; the only other Asian area equal in potential is Manchuria. Several industrial complexes sprawl across north India.

East Asia. Asia owes debts to China, the dynamic high-cultural centre of East Asia, which has been both innovator and synthesizer. All three East Asian nations—China, Korea, and Japan—have north–south cultural and ecological divisions, although the mountainous islands of Japan and equally mountainous peninsula of Korea tend to be more homogeneous than sprawling China, with perhaps 800,000,000 people, including about 20,000,000 hill peoples in South China common with Southeast Asia.

China as the cultural centre of East Asia

North China, more level than the south, has a relatively dry climate and is influenced by the proximity of western steppes and deserts; the windblown and alluvial soil makes extensive cultivation of wheat possible, along with some soybean and kaoliang (grain sorghum). South China, the predominately rice- and tea-growing area, begins about halfway between the Huang Ho and Yangtze rivers, with the Yangtze Basin being the transitional area of wheat and wet-rice production. In mountainous Korea the 45,000,000 people are squeezed into the narrow coastal plains and fertile inland river valleys, which constitute only about one-fifth of the land area. Similarly, in the 3,350 mountainous islands of Japan, the little cultivable land has to be efficiently farmed. Japan's climate, though generally mild and temperate, has climatic and latitudinal ranges almost identical to those from chilly Maine to humid Georgia.

Mongoloids in East Asia

Physically, East Asians are essentially variations of the Mongoloid. The North Chinese are generally taller and exhibit more Mongoloid traits than southerners. Near China's southern borders live a number of important non-Chinese minorities, such as Miao, Lolo or Yi, Puyi, Shan, Tai, and Mon-Khmer. Other important minority groups in China include Kazakh (and other Turkic speakers), Dungan (Hui) or Tunya, Koreans, Chuang, and Uighur; over 500,000 Japanese live in the north, especially in Manchuria.

Koreans, more homogeneous than either the Chinese (Koreans are shorter) or Japanese (Koreans are taller), have assimilated practically all minority groups. The Japanese are the hairiest Mongoloids, partly because of racial mixture with the modified Caucasoid Ainu, possibly the aborigines of Japan but today relegated to Hokkaido and other islands off south Siberia.

Linguistically, China presents a rather diverse face. At least six (some specialists list as many as nine) dialects of the Sino-Tibetan language family are spoken by the Han or Chinese proper: Mandarin (the major dialect), Cantonese, Wu, Fukienese, Hakka, and Hainanese. Although spoken dialects are usually mutually unintelligible, the 40,000 Chinese characters are common to all dialects; and, to be considered literate, an individual should know at least 5,000. Korean has a phonetic script called Han'gŭl or Ōnmun, using modified Chinese characters. Japanese apparently relates to Korean. The script consists of Chinese characters and a syllabary (*kana*) of Japanese origin. The Ryukyuan language of the northern Ryukyus is similar to Japanese, but Okinawan (in the southern Ryukyu Islands) is not mutually intelligible with Japanese. Culturally, Okinawa leans toward China, as well as Japan. A popular saying is: "China is our father, Japan our mother."

Philosophies and religions of East Asia

Although exposed to all major (and many minor) religions, China remains most greatly influenced by Confucianism, Taoism, and various schools of Buddhism. Throughout history, Chinese rulers encouraged the ideals of Confucianism most helpful to the maintenance of state power, currently disguised as the *Quotations from Chairman Mao Tse-tung*. Ancestor worship and seasonal animistic rites relating to the agricultural cycle have also continued to be widespread in China. The indigenous religion of Korea, Chondo-gyo, consists of several interlocking, seemingly contradictory elements, combining monotheism, shamanism, animism, Confucianism, and Buddhism. About 4,000,000 Koreans practice Buddhism, and about 2,000,000 are Christians. Shintō, Japan's indigenous religion, has been without state sanction since 1945 but remains an important politico-religious force, with at least 40,000,000 adherents. The 65,000,000 Japanese Buddhists usually also practice some sort of Shintō.

In each East Asia cultural zone, cultural traits vary from north to south. For example, in drier north China, wheelbarrows and two-wheeled carts drawn by oxen, donkeys, or horses replace the sedan chairs and river and canal boats of the subtropical, wetter south; water buffalo exist in both areas, but the North Chinese use oxen more extensively for cultivating land; famine has always been more prevalent in the north. The People's Republic has perpetuated at least as much as it has modified in Chinese culture. The family retains its importance, and the village council continues to exist in the communes. The philosophical base of government exhibits familiar patterns, peasants continue to be peasants, and the political hierarchy exists within a framework of divinely sanctioned (Maoist) authority.

World War II divided homogeneous Korea politically into North and South Korea, and, although the southern elite wear Western-style clothing and northerners wear Mao jackets, village culture remains relatively the same. Outside the official government level, local, often kin-oriented political parties control the rural areas in both North and South Korea. The North Korean land reforms of 1946 distributed 56 percent of all farmland to peasants, and another law guaranteed private ownership of small farms; thus few collective farms exist. North Korea has become more highly industrialized.

Japan, a zone of homogeneous cultural blend (like Korea), has embraced much that is Chinese (often filtered through Korea). But Japan is, of course, one of the most industrialized nations in the world, with 90 percent literacy. This modernization began in the late 19th century and resulted in Japanese uniqueness and contrasting patterns: devotion to the emperor coupled with democracy, sumo wrestling and baseball, modern business and indus-

trial giants (*zaibatsu*) combined with the *on-oyabun* (loyalty of client to company, or worker to company). As in most East Asia, though, family and a stratified social system remain strong in Japan, and often political parties and unions reflect these patterns.

Southeast Asia. Southeast Asia—which includes Burma, Thailand, Laos, North and South Vietnam, Cambodia, Malaysia, Indonesia, and the Philippines—has served as a north–south funnel for peoples and cultures for thousands of years, resulting in a hodgepodge of linguistic and ethnic groups. Important influences have flowed from both China and India. Two major geographic worlds exist: the mainland and the archipelagoes of Indonesia and the Philippines. In both, thriving empires and tribes of hunters and gatherers, intensive wet-rice agriculture (at times with elaborate terracing), and slash-and-burn horticulture exist. Plow agriculture is practically nonexistent; the hoe and digging stick are common. Besides the main staple of rice, crops include tea, coffee, maize, opium, cinchona (for quinine), rubber, coconuts, palm oil, sugarcane, tobacco, cotton, teak, and hemp.

Seamanship and navigation reach high peaks of technical perfection in the open seas, especially around Malaysia, Indonesia, and the Philippines. Fishing is an important economic institution. For all Southeast Asians (except Muslims) the pig is important in the diet. All eat chicken. The water buffalo (seldom milked) is the most common animal used to prepare fields for agriculture.

Australoids and Mongoloids of Southeast Asia

Physically, most Southeast Asian peoples are either Australoid or Mongoloid, with significant variations within each group. Australoids include the Negritos of the Philippines, the Semang of Malaysia, the Andamanese, and some Cambodian groups. Possibly several hunters and gatherers represent remnants of the earliest Mongoloids, including the Yumbri (northeast Thailand and Laos), Punan (Borneo), Sakai or Senoi (south of the Semang River), and Kubu (Sumatra). Intermediate Mongoloids include the Thais, Mon (Burma), Malay, Lao, Vietnamese, Cambodians, Filipinos, and Indonesians. The Mongoloid Miao (China, Vietnam, Laos, Thailand) apparently have some Caucasoid admixture. The Burman are physically the most Mongoloid (Burman, an ethnic designation, should not be confused with Burmese, the citizens of Burma). The Thai, Burman, Khmer, and lowland Vietnamese constitute about 75 percent of the mainland population. Since at least the 3rd century BC, Chinese have been moving into Southeast Asia; the more than 20,000,000 now in the area are primarily urban labourers, professionals, merchants, and traders.

Traditionally, anthropologists have classified Southeast Asians in such linguistic categories as Sino-Tibetan (Chinese dialects, Tibeto-Burman, Karen, Miao-Yao), Tai-Kadai (Tai and Kadai, the latter spoken mainly on southern Hainan Island and along the Yunnan–Vietnam border), Austro-Asiatic (widespread in Cambodia and the two Vietnams and including Mon-Khmer, Viet Muong, Senoi-Semang), and Malayo-Polynesian (in Malaysia, Singapore, and Indonesia and among the Cham of South Vietnam and Cambodia).

Various religions pockmark the cultural landscape, with Buddhism either dominant or subtly influencing all others. Varieties of Buddhism are widely practiced in Burma, Thailand, Laos, Cambodia, and Vietnam. Islām, with strong Hindu elements, is strong in Indonesia, Malaysia, Singapore, Mindanao, and other southern Philippine islands. Almost all mountain tribes practice animism and ancestor worship to a certain degree (some to the exclusion of other religions) and have extensive pantheons of nature spirits and supernatural beings. Magic is common, and various categories of shamans, healers, and diviners flourish among both sexes. Roman Catholicism is strong in Malaysia, Singapore, and the Philippines.

Extensive empires developed under Indian impact in the early centuries AD, but today only a few petty princely states survive as legacies.

The village remains the key unit of culture. (Even in North Vietnam, certain tribal groups are permitted to retain their identity, though their leaders are officially linked to the central government.)

ORIGIN AND DEVELOPMENT OF ASIAN PEOPLES AND CULTURES

Paleoanthropological record. Once considered by Charles Darwin and others as the cradle of mankind, Asia, given the present evidence, must relinquish the bassinet to Africa south of the Sahara. Nevertheless, many important prehominid and hominid paleontological finds have been made.

The Asian continent was undoubtedly the scene of a large-scale dispersal of early members of the natural order to which man belongs, the Primates. It is, for instance, the home of the only surviving genus (*Tarsius*) of the tarsiers, small nocturnal primates (about the size of newborn kittens), now confined to the Philippines, Borneo, the Celebes (Sulawesi), and Sumatra, the fossil remains of which are of great importance in the study of man's remote ancestry. Fossils of extinct apes, belonging to the widespread genus *Dryopithecus*, have been recovered from the Siwālik foothills of the Himalayas and are believed to be broadly ancestral to the African apes the chimpanzee and gorilla, though the Asian species may not be the actual ancestral forms. More important are the rare fossils of the genus *Ramapithecus* from the same locality, which appear to be near the ancestral lineage of man and are classified as early Hominidae. From South China the large fossil primate teeth and jaws named *Gigantopithecus* are thought to represent an eastern gorilla-like derivative of *Dryopithecus*.

Primate and hominid remains

Although there is no convincing evidence of Asian fossils of equivalent age and morphology to the African apemen called *Australopithecus*, there is a good fossil record from East Asia of hominid remains classified as *Homo erectus*, which lived in the Pleistocene Epoch, some 200,000 to 500,000 years ago. The most ancient specimens come from Java (first found in 1891 and then named *Pithecanthropus*), and an important group of middle Pleistocene fossils was discovered at Chou-k'ou-tien near Peking (recognized in 1927 and then named *Sinanthropus*); many other forms have since been found. This species, *Homo erectus*, which appears to have been ancestral to modern man, is believed to have walked fully erect, since the thighbones from Java are practically indistinguishable from those of modern man. The skull, however, was very different and enclosed a brain that averaged 1,000 cubic centimetres (the average for modern man is some 1,400 cubic centimetres); it was long and low so that its greatest breadth was as far down as the level of the ears, and the heavy jaws lacked a chin. The fossils found near Peking came from cave deposits and were associated with tools of bone and stone, as well as with hearths.

In late Pleistocene deposits of less than 100,000 years were found important fossils that many authors consider to represent diverse varieties of the modern species, *Homo sapiens*, primarily Australoid and Mongoloid. One probable Australoid skull from Niah Cave, Borneo, dates from about 40,000 years ago, the oldest specimen of *Homo sapiens* yet discovered.

Early *Homo sapiens*

Cultural origins. The earliest traces of human activity in Asia are associated with the fossil man known as *Sinanthropus* found at Chou-k'ou-tien near Peking. Hearths, roughly chipped stone implements, and other crude tools that were found with him are distinct in shape and method of manufacture from the hand axes characteristic of the Lower Paleolithic in Europe, but they are of comparable antiquity, perhaps 400,000 years old. Middle Paleolithic (100,000–30,000 BC) finds occur sporadically in Central Asia and China; scattered Upper Paleolithic (30,000–12,000 BC) sites are found in Afghanistan, China, Japan, Java, the Philippines, and possibly the southern Celebes. In southern Siberia were found traces of a flint- and bone-using culture with pressure-flaked flint points and bone harpoon heads but without the developed cave art of Europe. These cultures are perhaps 10,000 or 15,000 years old. Later, the stone tools of East and Central Asia included many tiny parallel-sided blades with little secondary trimming, intended for mounting in series. These "microliths" also echo the

European and North African development. The microlithic cultures lasted long, for in Siberia, Mongolia, and Manchuria they are often found combined with Neolithic elements in the shape of polished or partly polished stone axes, crude handmade pots with rounded or pointed bottoms, and small triangular arrowheads. With the advent of pottery and stone polishing, probably about 3000 BC, these semisedentary hunting and fishing cultures moved closer to a Neolithic technology.

This northern belt of cultures reaches to the sea in Kamchatka. To the south lies the cradle of the high bronze culture of protohistoric China, based on the easy agriculture of the Huang Ho Valley. The Neolithic cultures that preceded the Bronze Age in this area were much more advanced than the Mongolian Neolithic, and their origins seem to have owed nothing to it. The southernmost zone, comprising South China, Indochina, and the great islands of Southeast Asia, also possessed a distinct prebronze tradition. Excavations in Indochina have produced small chipped-stone tools unlike anything found in the north, as well as individual kinds of stone axes, polished only on the edge, and some rough pottery with impressed decoration. The use of stone in the southern zone continued into historical times, as it did in Japan, which had no share in the early spread of bronze metallurgy, but evidence indicates early Japanese pottery (*c.* 10,000 BC) that possibly predates previous known finds.

In South Asia the typical chipped-stone tools associated with the Paleolithic Period have been found widely distributed. In the north of India some of these are associated with river terraces corresponding with glacial and interglacial periods. Hunting, food-gathering Mesolithic peoples of South Asia, who used microliths to point and barb their arrows and as knives and scrapers, were as widespread as their Paleolithic predecessors. In forest areas, this cultural stage with its use of microliths persisted to the early centuries AD or even later. In the valley of the Son River and the Mahādeo hills and at Singhanpur, rock paintings, physically associated with microliths, have been compared with Paleolithic paintings in Europe; actually the majority are of the early historic period, and none is earlier than 1000 BC. The Neolithic Period in South Asia is the hardest to determine, depending mainly on the advance in agriculture to make its evidence apparent. Digging-stick cultivators would show no change in material possessions from those of the Mesolithic food gatherers. Hoe cultivation, possibly with terracing, needs stone hoes; and slash-and-burn cultivation needs axes. By far the greater number of ground and polished stone axes have been found in the east and south of India. The small amount of copper found at some stone-ax sites in India hardly justifies the cultures as being classed as Chalcolithic, or copper- or bronze-using.

Neolithic Period

The origins, development, and spread of food production, which marked the Neolithic Period, are complicated, and new evidence challenges the priority of the Middle East. Evidence from Āq Kupruk in north Afghanistan indicates the presence of domesticated sheep and goats about 8000 BC and of cattle about 6000 BC. Sickle blades, pecked stone hoes, chisels, hand mills, and pounders suggest at least the collection and preparation of wild grains. Possibly the zone of this animal domestication and grain collection spread along a latitude of 30° to 40° N, at an average altitude of 2,500 feet (750 metres) above sea level, from the northern Hindu Kush to Anatolia and the Aegean. In addition, in the Spirit Cave of north Thailand was found evidence of the cultivation of vegetables, beans, and water chestnuts; if the estimated date of these findings, 10,000 BC, is accurate, then they represent the world's earliest known horticulture. A possible date of 9000 BC for slash-and-burn horticulture on Taiwan adds to the evidence of the birth of cultivation in Asia. Throughout East and Southeast Asia by about 2500 BC, people were cultivating millet, buckwheat, beans, and probably rice, though those on the coasts largely depended on the sea for their livelihood, as attested by large shellfish middens. The South Chinese and Southeast Asians early developed maritime skills, and by about 2500 BC sea trade possibly extended as far west as Bengal.

Early civilizations. Chinese urban civilization began with the Shang or Yin dynasty (*c.* 1766–*c.* 1122 BC) and continued under the Chou (*c.* 1122–221 BC), who marched in from the west, absorbed much from the Shang, and developed the Chinese social and cultural patterns still recognized today. Confucius (551–479 BC) lived during this period. Iron arrived about 500 BC and was cast into plows (the stone hoe previously dominated) and into weapons. In India, Indo-European-speaking tribes, commonly referred to as Aryans, began their invasions from the west about 1700 BC, which contributed to the destruction of the Indus Valley (Harappan; *c.* 2300–1700 BC) Bronze Age civilization. The centuries following witnessed several immigrations, culminating in the entry of the Vedic Aryans in the 14th to 12th centuries BC. Iron probably was introduced in the 7th or 6th century BC. The Vedic period came to a close at the end of the 6th century BC, to be succeeded by a plethora of dynasties and states, briefly united under the Mauryan Empire of Aśoka the Great (*c.* 265–*c.* 238 BC).

With urban civilizations firmly established, several patterns developed that became recurrent through time. Asian urban society had—and still has—an economy based primarily on agriculture. Control of water for irrigation and the upkeep and spread of the intensive canal and drainage systems fostered the growth of authoritarian governments with varying degrees of despotism. Periodically, invasions from the steppes and deserts of Mongolia and Central Asia overthrew existing dynasties, and then the conquerors in turn were culturally conquered by the fallen civilizations. Empires expanded and fell, and others rose in a continual process of political fusion and fission. A charismatic tribal ruler would unite tribes into a confederation and then spread his political mantle as far as military power, diplomatic duplicity, and sexual alliances could carry him. With the death of the charismatic leader, the empire he created usually began to break up; at times, fission began even before his death.

Invasions and rise and fall of empires

A few such individuals left lasting monuments—or ruins—on the social, political, and cultural landscape. Alexander the Great smashed through Afghanistan, moved unsuccessfully into Central Asia, and down to the Punjab in the 4th century BC. Hellenism resulted; Greco-Bactrian and Indo-Greek kingdoms rose and fell; some lasted well into the 1st century BC. More important, Buddhism mingled with Greco-Roman humanism, ideologically and artistically, and spread along the Silk Route, which stretched from ancient Cathay to Greece and Rome. The Silk Route was a highway of culture, as well as commerce.

In the 13th century AD, Genghis Khan and his Mongol army left great ruins in their wake but also revitalized the eastern Islāmic world. Timur in the 14th century destroyed much while fostering a Central Asian cultural renaissance. China received cultural stimulation from its perennial conflicts with the nomads beyond the Great Wall, a symbol of the borderland rivalries existing all over Asia.

European contacts and modern developments. Political insecurity, particularly after the Turco-Mongol invasions, broke up the Silk Route and led European navigators around the South African cape in search of new routes to the East.

Beginning in the early 19th century, European imperialism began to replace Asian imperialism. Tsarist Russia drove to the Pacific Ocean, conquering the marginal hunter-fishermen of Siberia and the ancient and sophisticated but jaded Muslim khanates of Central Asia; after the Russian Revolution, Communists replaced the tsarists. The British rapidly gained control of the Indian subcontinent; the French moved into Indochina; the Dutch occupied the East Indies; the Spanish ruled the Philippines until the United States took over after the Spanish-American War and thus intruded into the Asian scene, to remain there until the present. China, culturally satisfied but politically introvert, feebly resisted the capitulations. Japan moved slowly from its self-imposed isolation and

then rapidly to imperialism, which culminated in its military defeat in World War II and subsequent economic revival.

The European imperialists split Asia into either colonies or zones of influence. Even those few nations (such as Afghanistan and Thailand) that continued to remain independent had their external boundaries drawn by imperialists.

Modern Asian states and territories

Empires need bookkeepers, and the Europeans in Asia trained large numbers of local personnel, gradually permitting them at least limited entrance into their political, economic, and, to a lesser extent, social worlds. The bookkeepers, exposed for varying periods of time to the ideological, political, and economic models of their masters, began to adopt them, often unconsciously, as their own. The ideals of individual freedom and economic well-being led to Asian nationalism and, particularly after World War II, brought about political independence, and the internal imperialist replaced the foreign imperialist. Asian leaders, having adopted the political and economic models of the West, tend to think in Western terms and react like their Western counterparts when faced with internal problems. Regional autonomy and tribalism are anathema to most Asian leaders, and they insist on the rights of the individual and his obligations to the state under various brands of democracy and Socialism. They forget, however, that when European boundary-hardening exercises settled down after World War I, most European nations were based on reasonably uniform linguistic units. Rarely, however, do linguistic units in Asia coincide with national boundaries, which are legacies of European imperialism. In addition, the reciprocal rights and obligations between government and governed that theoretically dominate Western political processes occur at the kinship, tribal, and regionally oriented level in Asian society. Few Asian governments can even begin to replace the delicate networks woven by these kin systems.

Modern Asian governments range from Socialistic (People's Republic of China) to constitutional monarchies (Afghanistan), developing democracies (India) to transient military dictatorships (Pakistan), but, in most cases, whatever their form, they involve the masses of the people only minimally.

The rising demands for Asian regionalism, the creation of autonomous units within federations, are often either ignored or suppressed by central government in the name of "national integrity." The successful creation of Bangladesh in 1971 after the Pakistani military regime attempted to crush a movement for regional autonomy is a classic example. Such demands will probably increase. Even in Soviet Central Asia, cries for Tadzhik power, Uzbek power, and so on are getting louder and stronger. Perhaps only three Asian nations have recognized the utility of linguistic and tribal semi-autonomous units inside their national boundaries: India, China, and North Vietnam.

When—and if—foreign military adventures end in Asia, meaningful international economic blocs may develop and permit the Asian nations to present a unified economic front to the more developed Western world.

BIBLIOGRAPHY. The works of CARLETON S. COON, *The Story of Man*, 2nd rev. ed. (1962), *The Origin of Races* (1962), and *The Living Races of Man*, with E.E. HUNT, JR. (1965), probably offer the best surveys of fossil man and the development and distribution of modern Asian physical types, although controversy exists over some of Coon's conclusions. Two classic Asian geographies are L.D. STAMP, *Asia: A Regional and Economic Geography*, 12th ed. (1967), and GEORGE B. CRESSEY, *Asia's Lands and Peoples*, 3rd ed. (1963). A good regional study is HARRY ROBINSON, *Monsoon Asia* (1966). Few archaeological works exist on Asia in general, but many specific studies have been written. Several works, however, do have important survey materials, including GRAHAME CLARK, *World Prehistory*, 2nd ed. (1969); J. HAWKES, *History of Mankind: Prehistory*, vol. 1, pt. 1; LEONARD WOOLLEY, *History of Mankind: The Beginnings of Civilization*, vol. 1, pt. 2 (1963). FRANCOIS BORDES, *La Paléolithique dans le monde* (1968; Eng. trans., *The Old Stone Age*, 1968), admirably outlines the early periods in Asia. The socio-political processes and ecologic patterns that led to the development and repeated rise and fall of authoritarian regimes in Asia is ably detailed in K.A. WITTFOGEL, *Oriental Despotism* (1957). Another monumental work is GUNNAR MYRDAL, *Asian Drama*, 3 vol. (1968), which evaluates past and potential development. For general reading and easy reference, D.N. WILBER (ed.), *The Nations of Asia* (1966), is indispensable. Finally, RALPH LINTON, *The Tree of Culture* (1955), remains a basic work on Asian cultural patterns.

(L.Du.)

Asians, Prehistoric

The traditional picture of Asia as the cradle of mankind provided the impetus for Western scholars to search this part of the world for the vestiges of man's biological and cultural beginnings. The romanticism of this notion has been tempered by scientific discoveries indicating that man had inhabited Europe and Africa as long as Asia, but Asia was nonetheless important in its broad spectrum of climatic zones, from Arctic to temperate and tropical, its mountain ranges and grassland steppes, its inland seas and Arctic tundra, plus other ecological settings to which prehistoric and modern man had to adapt both biologically and culturally. Since the formation of the present geological features of Asia during the period of mountain-building activity in the Middle Tertiary (about 35,000,000 BP [Before Present]) the continent gave rise to major taxa of primates that included the dryopithecine apes of the Miocene–Pliocene age (21,000,000 BP) and the earliest known hominid, *Ramapithecus punjabicus* (see HOMINIDAE). With the onset of the Pleistocene (about 2,000,000 BP), Asia was one of the habitats of the australopithecines, who, by middle Pleistocene times (500,000–200,000 BP), were replaced by *Homo erectus* forms. Evidence of *Homo sapiens neanderthalensis* (Neanderthal man) and later *Homo sapiens sapiens* (modern man) appeared during the upper Pleistocene, which began about 150,000 years ago. Thus, the major events of human evolution are represented in the paleontological and archaeological record of Asia.

This article is divided into the following sections:

The fossil record of prehistoric man in Asia
  Areas of human occupation
  Remains of Neanderthal man
  Homo sapiens sapiens remains
Morphology of Asian fossil remains
  Neanderthal morphology in western Asia
  Neanderthal morphology in eastern Asia
  The Asian *sapiens* fossils
Life styles of prehistoric man in Asia
  Stone-tool cultures
  Fire, shelter, and cultural data
Phylogenetic affinities of Asian fossils to modern man

THE FOSSIL RECORD OF PREHISTORIC MAN IN ASIA

Areas of human occupation. Human occupation in Asia from 100,000 years ago until about 35,000 years ago, a period that includes the geological-climatic events of the third interglacial and the initial glaciation of the Würm (Last) Glacial Period, was restricted to a considerable degree by the extension of ice sheets that covered portions of eastern Europe, the Tibetan plateau, and the diagonal mountain chains of Central Asia during periods of peak glaciation. These conditions blocked passage between Europe and eastern Asia save for narrow corridors not covered by ice. For this reason, high-altitude glaciated regions in western Asia and in the Himalayas are poor in prehistoric sites. The Bosporus remained dry land during much of the Pleistocene and formed the main avenue of communication between the Near East and the Black Sea. During periods of maximum glaciation, the Caspian Sea rose 250 to 300 feet, and in this condition of flooding by waters fed by glacial melt it formed with the Volga River a spillway into the Black Sea. The Aral Sea enlarged in rhythm with the Caspian, and to the south and the east of the Ural Mountains a vast swamp marked the limits of glacial ice. With the onset of interglacials, of which the third is the one relevant to the period of time under consideration, these landlocked bodies of water became lowered while the Black Sea was elevated, since it was fed by oceans expanding in size from water liberated by melting marine glaciers.

During these peak glaciation periods, portions of Asia were isolated from Europe.

Thus, with the onset of the Würm glaciation some 70,000 years ago, climatic conditions effectively separated the Neanderthal populations of western Europe from those Neanderthals of western Asia whose occupation sites are recognized today in the Zagros Mountains of Iran and just north of the Elburz Mountains and Hindu Kush in Iran, Soviet Central Asia, and Afghanistan. With the onset of the second phase of glaciation (Würm II Stadial) 40,000 years ago and during a short period of milder climates in Europe and western Asia, the Neanderthal occupations were replaced by the sites of early modern hominids.

The Bering Strait formed an effective land bridge between northeastern Asia and the New World during the glacial epochs but ceased to serve as such with the elevation of sea levels during the Third Interglacial and again after the Würm glaciation. The former continental landmasses of the Sunda Shelf and Sahul Shelf in Southeast Asia were separated at the close of the Pleistocene as sea levels rose. These areas were dry during the Riss and Würm glaciations and provided a passageway for populations moving from Southeast Asia to Australia. The geographical boundary of Wallace's Line, running between Bali and Lombok, Borneo and the Celebes, Mindanao and Sangi, isolated the Australian fauna, including man, from Southeast Asia during the upper Pleistocene. South Asia was affected by glacial activity in the Himalayan region but in the south that area retained a tropical ecological setting throughout most of the Pleistocene. The upper Pleistocene of China is characterized by deposits of yellow earth, or loess, upon which man settled. The climate remained cold and dry for much of this period in China.

Neanderthaloids from western Asia

Remains of Neanderthal man. More than a dozen sites with fossils of *Homo sapiens neanderthalensis* have been located in the Near East. Prehistoric research began there in 1864, when Louis Lartet (who later discovered Cro-Magnon man) rediscovered in Syria an ancient settlement observed some 30 years earlier. In 1878 an inventory of Stone Age artifact discoveries was compiled, but the first fossil find of early man in western Asia was made in 1900 at Grotte d'Antelias, in Lebanon. The remains of a seven-year-old child were recovered in 1938 in Lebanon at Ksar 'Akil, a rock-shelter near Beirut dated to 43,750 years BP. Discovery of an adult Neanderthal skull was made in 1925 at el-Zuttiyeh (Robbers' Cave) at Lake Tiberias in Israel—the "Galilee Skull" now datable to 70,000 years BP. In its vicinity is the cave of Har Qedumim (Jebel Qafzeh), where Neanderthal bones of comparable antiquity were recovered in 1933–35. The cave at Amud, also in this region, yielded in 1961 bones dated to the period of the Würm I Stadial of the last Ice Age (70,000–50,000 BP).

Near Jerusalem the Shuqbā (Shukbah) Cave contained human remains that had been deposited at the end of Neanderthal occupation of this part of Asia—*i.e.*, at about 35,000 years BP. The most critical discoveries in Israel are from two caves adjacent to one another in the Mt. Carmel range—Maghārat at-Tabūn, excavated in 1929–34, and Maghārat as-Skhūl, excavated in 1931–32. The age of occupation deposits at at-Tabūn ranges from 70,000 to 37,750 years BP. The occupation of as-Skhūl may have been slightly later, but both caves have yielded Neanderthal fossils with many physical features characteristic of later *Homo sapiens* hominids.

The original designation *Palaeoanthropus palestinus* assigned to the Mt. Carmel hominids was later dropped from use. The Turkish sites of Karain near Adala, excavated in 1949, and Musa Dağı have yielded teeth that may belong to the Würm I hominids, but the dating of these sites remains uncertain. Shanidar Cave in Iraq was excavated from 1953 to 1960. Its deposits range in age from 60,000 to 44,950 years BP and include the bones of a child and several adults. The two sites of Würm I date in Iran are the Kermanshah Cave, near Bisitun, and the Tamtama Cave, near Rezā'īyeh. Excavation of these fossil-bearing deposits began in 1949.

Fossils of Neanderthals were found in the Soviet Union in 1924 at the cave of Kiik-Koba, in Crimea. Two skeletons, both missing skulls, appear to have been purposeful burials. The deposit dated to the early part of Würm I, while at another Crimean site, called Staroselye, excavated in 1952, the fossil remains date to 35,000 years BP. Teshik-Tash cave, in Uzbekistan, yielded in 1938–39 a child burial of the Würm I–II Interstadial, but the skeletal remains unquestionably belonged to a Neanderthal population. Less certain is the dating of an alluvial deposit containing some prehistoric human remains found in 1925 along the lower Volga River at Undory. Aman-Kutan Cave, in Samarkand, has the earliest dated fossil Neanderthal in Central Asia.

Neanderthaloids from Southeast Asia and eastern Asia

In eastern and Southeast Asia, hominid fossils resembling the western Asiatic *neanderthalensis* hominids have been recovered from sites that date from the early part of the upper Pleistocene and even into the beginning of post-Pleistocene, or Recent, times (10,000 BP). The taxonomic status of these hominids is uncertain, but they have been given the colloquial appellation of Neanderthaloids, even though some populations of this hominid group were contemporaries of *Homo sapiens sapiens*, who emerged in western Asia, Europe, and Africa by the beginning of the Würm II Stadial of the last Ice Age, about 40,000 years BP.

Since the excavation in 1922 at Sjara-osso-gol River in Ordos, China has provided numerous loci with human remains. This site may have been inhabited contemporaneously with the occupation of the other Ordos site, Ti-shao-kou-wan, which was investigated in 1957. Both appear to be of upper Pleistocene date (*c.* 150,000–10,000 BP).

Earlier than these is the Ting-ts'un site, in Shansi, excavated in 1954 and dated to the Third Interglacial (100,000–70,000 BP). Still more ancient is the human skull from the cave at Ma-pa, in Kwangtung, which may be late middle Pleistocene; *i.e.*, *c.* 125,000 years BP. Other series of Neanderthaloids appear in the fossil record of Java at the upper Pleistocene site of Ngandong, on the Solo River.

At the time of the excavation of 11 skulls from Ngandong in 1931–32, the specimens were named *Homo* (*Javanthropus*) *soloensis*, but modern systematists regard these fossils as similar to the Chinese Neanderthaloids and to the Neanderthals of western Asia. Two skulls from a limestone terrace overhanging an ancient lake near the village of Wadjak, in Java, were discovered in 1890, and these constitute the earliest discovered fossil hominids from Asia, although they were not reported until 1921. They have been dated to the very end of the Pleistocene or possibly to the beginning of post-Pleistocene times (*c.* 10,000 BP), but they resemble the Ngandong Neanderthaloid specimens in a number of striking ways. Fossil evidence of *Homo sapiens neanderthalensis* has not been reported from southern Asia, although this region contains artifactual materials made by hominids of the period of time under consideration.

A few bone fragments from a late middle Pleistocene deposit (*c.* 125,000 BP) at Ushikawa, on the Japanese island of Honshū, were reported in 1957, but the bones are too incomplete to make possible a reliable identification of their taxonomic status.

Homo sapiens sapiens remains. In western Asia the first appearance of *Homo sapiens sapiens* coincides with the onset of the Würm II Stadial of the last Ice Age, about 40,000 years BP, approximately the same period of time that *sapiens* hominids replace Neanderthals in Europe and Africa. As noted above, Neanderthaloid hominids persisted in eastern and southeastern Asia until the close of the Pleistocene, but they appear to have constituted isolated populations in this part of the world, where modern-type *sapiens* had also appeared by late upper Pleistocene times (*c.* 30,000 BP).

The earliest date for *Homo sapiens sapiens* in Southeast Asia is 37,650 years BP, assigned to a skull specimen discovered in 1959 at Niah Cave, in Borneo. Late upper Pleistocene is also the date for the Chinese fossils

found in 1930 at the Upper Cave of Chou-k'ou-tien, near Peking, in 1951 at Tzuyang (formerly Tzeyang), in Szechwan, in 1956 at Kai-t'o-tung in Laipin, Kwangsi, and at Ch'i-lin Shan, in the same province. Fossils were also found in 1958 at Chaisha (formerly Liukiang), also in Kwangsi; this site may have lower deposits dating to a period contemporary with the Würm I Stadial (70,000–50,000 BP), as is certainly the case at the site of Ch'ang-yang (Lungtung), in Hupeh Province, investigated in 1957.

Japan has yielded a *sapiens* specimen in the fossil record taken from Aichiken, in Honshū, in 1958, but the date can only be specified as upper Pleistocene. In 1921 a skull with Negrito or Pygmy features was reported from an alluvial deposit of the Río Pasig near Manila, but the dating of this specimen from the Philippines is uncertain.

Southeast Asia's record of *sapiens* fossils does not commence until the beginning of the geological Recent period (*i.e.*, after 10,000 BP), save for a single hominid lower molar found in northern Indochina in a deposit of uncertain antiquity. Similarly, southern Asia does not provide a *sapiens* fossil series for this period; the earliest skeletal remains, coming from Sai-Nahar-Rai in Uttar Pradesh date to about 8000 BC.

Near Eastern *sapiens* fossils

The Late Stone Age (Mesolithic) site of Bellanbandi Palassa in Sri Lanka (Ceylon) has yielded a dozen skeletal specimens of *H. sapiens sapiens* from a deposit dated about 5000 years BC.

In the Near East, Lebanon offers two Pleistocene sites with *sapiens* fossils: a cave at Abri Bergy, near Antilas, which was excavated in 1948, and the previously mentioned site of Ksar 'Akil, which has a middle Würm deposit (*c.* 40,000 BP) superimposing the portion of the shelter from which the Neanderthal specimens had been removed in 1938. Similar hominids came from an upper level of the Har Qedumim cave, in Israel, as well as from the Mugharet el-Kebareh and Wādī Maghārah (Mugharet el-Wad) caves of the Mt. Carmel Range, where excavations in 1931 provided data on late upper Pleistocene occupations. At Malta, in Siberia, a late Würm (*c.* 15,000 BC) *sapiens* specimen was recovered in 1929, but other sites in the Soviet Union have provided few skeletal remains from this period.

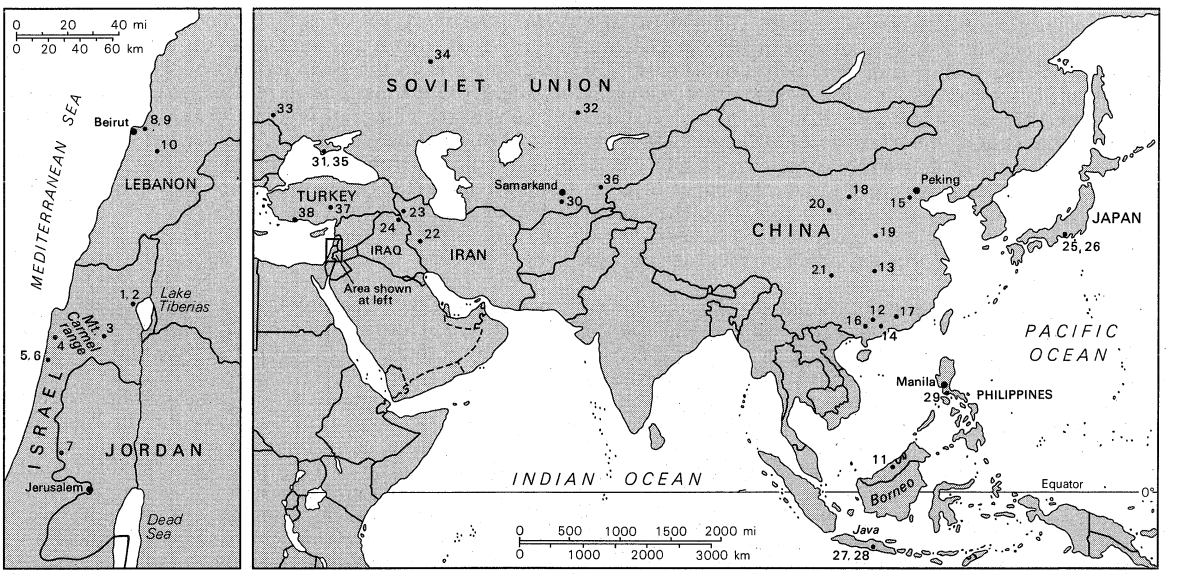
MORPHOLOGY OF ASIAN FOSSIL REMAINS

Neanderthal morphology in western Asia. The classification of upper Pleistocene Asian hominids as Neanderthal, Neanderthaloid, and sapient forms becomes meaningful when the physical anthropologist compares the osteological and dental anatomy of the fossils from the sites just discussed. Some Neanderthal specimens of western Asia closely resemble the Würm I Neanderthals of Europe with regard to the possession of a massive supraorbital torus (browridges), a low-lying forehead, large face with large eye sockets (orbits), large nose, dolichocranic or mesocranic cranial form, chinned or chinless mandible, bulging of the occipital bone (at the back of the skull) into a "chignon," taurodont molars, a cranial capacity with a mean of about 1,550 cubic centimetres (95 cubic inches), and a range from 1,270 to 1,700 cubic centimetres, inclusive, of both sexes, a modern-like dentition (apart from taurodonty), and a body conformation that was burly and muscularly well developed. The misconception that Neanderthals walked with slightly flexed knees and a hunched posture was based upon the examination of the Neanderthal skeleton from La Chapelle-aux-Saints, in France, an arthritic and aged male. The retroversion of the head of the tibia (large bone of the lower leg) was also considered to be indicative of a slumped posture and plodding gait.

Features of the skull

Recent reinvestigation of the La Chapelle specimen and other Neanderthal skeletons does not indicate that this hominid's locomotor pattern was significantly different from that of later and modern hominids. Stature ranged from about 154 centimetres for females to 173 centimetres for males.

The Asiatic specimens that show the most striking physical resemblances to European specimens of Neanderthal man (such as those from La Chapelle-aux-Saints and La Ferrassie, in France, Spy, in Belgium, and the Neander Valley specimen, in West Germany) are from Amud and Shuqbā, in Israel, as well as some of the skeletons from the series of the Mt. Carmel caves, from Shanidar, in



ISRAEL
1. Amud
2. el-Zuttiyeh
3. Har Qedumim
4. Mt. Carmel (Maghārat at-Tabūn Maghārat as-Skhūl)
5. Mugharet el-Kabareh
6. Wādī Maghārah

JORDAN
7. Shuqbā

LEBANON
8. Abri Bergy
9. Grotte d'Antelias
10. Ksar 'Akil

BORNEO
11. Niah

CHINA
12. Chai-sha
13. Ch'ang-yang
14. Ch'i-lin Shan
15. Chou-k'ou-tien
16. Kai-t'o-tung
17. Ma-pa
18. Sjara-osso-gol
19. Ting-ts'un
20. Ti-shao-kou-wan
21. Tzu-yang

IRAN
22. Kermanshah
23. Tamtama

IRAQ
24. Shanidar

JAPAN
25. Aichi-ken
26. Ushikawa

JAVA
27. Ngandong
28. Wadjak

PHILIPPINES
29. Rio Pasig

SOVIET UNION
30. Aman-Kutan
31. Kiik-Koba
32. Malta
33. Molodova
34. Oundory
35. Staroselye
36. Teshik-Tash

TURKEY
37. Karain
38. Musa Dağı

Major sites of Upper Pleistocene hominid remains in Asia.

Iraq, and from Kiik-Koba, Aman-Kutan, and Teshik-Tash in the Soviet Union.

Other Neanderthal fossils of western Asia, however, bear closer phenotypic similarity to the Third Interglacial Neanderthals of Europe as represented in the sites of Ehringsdorf, East Germany, and Steinheim, in West Germany, and Fontéchevade, in France, where the specimens exhibit less morphological specialization than do the Würm I inhabitants of Europe. These Third Interglacial Neanderthals are more like modern hominids in their possession of lower mean cranial capacity, with a range of 1,100–1,450 cubic centimetres for both sexes, the absence of an occipital "chignon," less muscular marking of the cranial vault, longer faces, and a tendency toward linearity of body build. The postcranial long bones are not bowed, as is the case with many Neanderthal specimens of Europe and western Asia of Würm I times. With these more modern-looking Neanderthals may be associated the western Asiatic fossils from Ksar 'Akil, in Lebanon, Har Qedumim, el-Zuttiyeh, and the majority of specimens from the Mt. Carmel series in Israel and those from the Soviet site of Staroselye. The specimens from Maghārat at-Tabūn and Maghārat as-Skhūl have been studied in greatest detail and are recognized by most physical anthropologists as representative Neanderthal populations evolving into the sapient type of humanity that has dominated this part of the world for the past 40,000 years. For the other sites of western Asia noted earlier, either the fossil evidence is too fragmentary or the comparative anatomical studies are not sufficiently complete to permit classification beyond the statement that a Neanderthal phenotypic pattern seems to be represented.

Neanderthal morphology in eastern Asia. For the Neanderthaloids of eastern Asia, the fossil record is less complete. Incisor teeth from Sjara-osso-gol and Tingts'un are "shovelled," a feature found in high frequency among living peoples of modern Asia and not unknown among *Homo erectus* specimens at the middle Pleistocene site of Chou-k'ou-tien. Cranial bones from Ti-shao-kou-wan and Ma-pa are thick, the skull from the southern China locality having a capacity of 1,225 cubic centimetres as well as robust features similar to the skull vaults from Ngandong. Yet the Javanese skulls, which are 11 in number and the only case of a population series for this period of time, have a mean cranial capacity well under this value, the means for males and females being 1,158 cubic centimetres and 1,042 cubic centimetres, respectively. The range of cranial capacity for the series is 1,035–1,255 cubic centimetres. In this feature the Chinese and Javanese Neanderthaloids do not resemble the Neanderthaloids of Broken Hill and Saldanha Bay, in Africa, but are more like the middle Pleistocene *Homo erectus* specimens from Asia, from which phylogenetic line they most likely evolved. Similarities with *Homo erectus* are also obvious in the pronounced angularity of the occipital bone (at the back of the skull), platyrrhinic nasal indices (indicating wide nasal openings in the skull), the thickness of the parietal bones (at the sides of the skull), and the location of the maximum width of the skull at points just below the parietal area. Yet the frontal torus (browridge) is of the divided pattern, as it is in Neanderthals, rather than forming a continuous bar. It is with the Wadjak specimens that cranial capacities increase to 1,550 cubic centimetres for the female and 1,650 cubic centimetres for the male. Specific features of large and chinned mandibles, broad and flat faces with broad noses and massive browridges, alveolar prognathism (forward projection of the tooth-bearing portion of the upper and lower jaws), and deep palates have led physical anthropologists to note the close similarity of the Wadjak specimens to modern aboriginal Australian and Melanesian populations rather than to modern Asians. Even less is known of the postcranial anatomy of these Neanderthaloids, but stature estimates based on two modern-looking tibial bones (from the lower leg) directly associated with the Ngandong crania suggest a height of about 178 centimetres for the male. The stature of the Ti-shao-kou-wan specimen is estimated for a thick-walled femoral (thigh bone) fragment as 167 centimetres if the specimen was male and 160 centimetres if female. The humerus (upper arm bone) from Ushikawa is thought to belong to a female less than 140 centimetres in height, and it differs from the humeri of modern Japanese by its narrowing of the proximal (shoulder) end and the pronounced thickening of the cortical area of the bone. The short stature and stocky body build of the greater number of Neanderthals and Neanderthaloids is considered by some anthropologists to have been an adaptive mechanism for surviving cold climatic conditions, where increased body bulk helps conserve body heat. Certain features of the Neanderthal-Neanderthaloid face have been explained as thermal adaptations to cold stress. The taller and more linear body conformation of the Solo specimens is especially interesting in view of the fact that this particular population lived in the tropical belt, where cold adaptation was not a factor of survival.

Skull morphology

Body build of Asian Neanderthals

The Asian sapiens fossils. The *Homo sapiens sapiens* fossil specimens are not unlike contemporary Asiatic populations, with the exception of the skull from Niah Cave, which resembles more closely the skulls of Australians than Borneans in structural features. Three of the seven specimens from the Upper Cave site at Chouk'ou-tien have been compared to contemporary populations of Ainus, Melanesians, and Eskimos, and a broad range of phenotypic and polymorphic features may have been common in this part of Asia during later Pleistocene times (30,000–10,000 BP). Very similar to the skeletal anatomy of present occupants of eastern Asia are the specimens from Tzu-yang, Kai-t'o-tung, Ch'ang-yang, and Chai-sha, Kwangsi, save that the male postcranial skeleton from the latter site suggests a low height of 150 centimetres, which is just on the upper border of Pygmy stature. The skull fragments from Aichi-ken in Japan are sapient, as are the other fossil remains from upper Pleistocene sites in western Asia noted above.

LIFE STYLES OF PREHISTORIC MAN IN ASIA

Stone-tool cultures. The way of life of the Asiatic hominids of this period can be reconstructed in large part from the study of the archaeological data of preserved stone tools and weapons. There are two technological traditions represented in deposits dated to the middle Pleistocene (*c.* 125,000 BP)—the bifacial-hand-ax and cleaver-tool industries of western and central Asia and the chopper–chopping-tool traditions of eastern and southeastern Asia. The transitional zone of these two traditions occurs in India, where it has been defined as the "Movius Line," after H.L. Movius, an American prehistorian who has conducted research in Asia. The bifacial-hand-ax and cleaver-tool tradition is called Acheulean in Europe, Africa, and western Asia, where it appears stratigraphically earlier than the flake-tool industry called Levalloisian, although both flakes and bifaces occur together in some sites in this tricontinental area.

Acheulean and Levalloisian tool traditions

Acheulean hand axes have been found in deposits in the caves of Mt. Carmel, in Israel, in Jordan and Syria, Turkey, Arabia, Iran, Iraq, Afghanistan, and eastward to Armenia, Crimea, and the Caucasus Mountains. The Narbada Valley of India is particularly rich in hand axes, the first discovery of a paleolithic tool in South Asia having been made in the Madras area of India as early as 1863. Levallois flakes made by a prepared core technique occur in increasing frequency with the dawn of the upper Pleistocene in India, where they are a final phase (dating from *c.* 150,000 years ago) of another flake-tool tradition, called Soan. Apart from a few isolated cases of hand axes in China, the only place in Southeast Asia where these tools appear is in Java, where they are associated with the Patjitanian tool industry, which also has certain Levalloisian features. Chopper–chopping tools also appear in Java during this period, as do tool industries called the Sangiran flake culture and the Ngandongian culture. These are marked by the presence of antler picks, stingray barbs, and bola stones as well as small stone flake tools.

With the advent of colder climatic conditions and the

onset of the Würm I Stadial (70,000 BP) in western Asia, the bifacial-hand-ax industries were replaced by flake-tool traditions; one of these was an evolved Levalloisian and another the Mousterian tool industry, in which stone flakes were retouched to make tools but not specifically manufactured from prepared stone cores. The Mousterian tradition is intimately associated with *Homo sapiens neanderthalensis* in western Asia, as is seen at Ksar 'Akil, the caves of Mt. Carmel, a number of sites in Jordan and Syria, Kiik-Koba in Crimea, Shanidar in Iraq, Teshik-Tash in central Asia, Molodova in the western Ukraine, and far eastward in the Ordos region near the Great Wall of China. In southern Asia, the later phase of the Soan tool tradition resembles the Mousterian. Elsewhere in eastern and southern Asia, the chopper–chopping tool tradition persisted until the close of the Pleistocene (*c.* 10,000 BP). In Burma, in the vicinity of the Irrawaddy River, a silicified wood or tuff was manufactured into choppers and retouched flakes. These tools represent the Anyathian culture. The Tampanian culture of Malaysia is another local variation of the chopper-chopping tool tradition. On the shores of the Sea of Azov in southern U.S.S.R. this tradition merged with the limits of diffusion of the Mousterian tradition. At the upper Pleistocene localities of Chou-k'ou-tien, quartz flakes with working at both ends occurred with choppers, chopping tools, and small scrapers. A tool tradition of flaked stone balls and bifacial choppers but no bipolar flaking is called the Fenho Complex, and its Third Interglacial date marks the transition to the upper Pleistocene traditions that have features of the Asiatic Mousterian in combination with older styles. In eastern and southeastern Asia many elements of this culture persist into terminal Pleistocene times, when Mesolithic cultures, characterized by small-blade microliths, replaced the older traditions.

Mousterian tool industries

Fire, shelter, and cultural data. Human occupation during the upper Pleistocene included the areas inhabited by *Homo erectus* of the middle Pleistocene as well as extensions into areas that earlier hominids had not entered. Biological and cultural adaptation to cold climatic conditions permitted settlement of higher altitude regions of Asia, and with the retreat of the ice sheets man followed game fauna into the newly opened country. Cold climates were made more tolerable by the use of caves and rock-shelters, some of which served as places of burial as well as hearths and industry sites. Open-air encampments continued to be used in some regions, however. Fire, which has been reported in Asia as early as 360,000 years ago from evidence of charcoal and charred bone at the *Homo erectus* deposits of Chou-k'ou-tien, in China, was controlled by the hominids of the upper Pleistocene, thus enabling them to move into wider areas of settlement and survive under a wide range of ecological settings. The control of fire by man, along with cave dwelling and the wearing of skins sewn together with bone awls, enabled prehistoric man to survive the cold conditions of the later upper Pleistocene. Traces of fire are found in most of the open-air, cave, and rock-shelter sites where skeletal and artifactual remains have been preserved, but the traces of charcoal and calcined bones do not appear in disturbed deposits, as at Ngandong. Among the undisturbed sites are the caves at Mt. Carmel, Aman-Kutan, Molodova, and Ch'i-lin Shan. Neanderthal man practiced burial of the dead at the Mt. Carmel caves, Kiik-Koba, Teshik-Tash, in the Upper Cave of Chou-k'ou-tien, Ksar 'Akil and Malta. At Molodova, the remains of a hut can be identified by the oval of mammoth bones and tusks enclosing some 15 separate hearths. This is evidence of open-air housing during a warm phase of the upper Pleistocene, although caves and rock-shelters continued to be the most common form of habitation in areas where they were naturally present. Evidence of cannibalism occurs at Solo, as it does in several Neanderthal sites of Europe. The presence of diseased and aged individuals in the Neanderthal community at Shanidar speaks of a more humane aspect of upper Pleistocene life. Pictorial or plastic art does not appear in the archaeological record for Neanderthal man and emerges only with the presence of *Homo sapiens sapiens*.

Fire and shelter

Throughout a broad geographical range of habitation in the Asiatic landmass, upper Pleistocene hominids were successful hunters of large game animals. For the western and central Asiatic Neanderthals these were Paleoarctic fauna, which can be identified as cold adapted or warm adapted according to their occurrence in periods of high glaciation or interglacial–interstadial recessions of the ice. In Israel, the warm-climate fauna are marked by remains of hippopotamus and rhinoceros, cold-climate fauna by deer and antelope. In central Asia the cold-adapted woolly rhinoceros, mammoth, and reindeer were abundant. South and Southeast Asian fauna were distributed also in southern China and merged with the Paleoarctic fauna of northern China. In India, some of the fauna of the middle Pleistocene survived until well into the late upper Pleistocene, as suggested by the presence of certain species of bovids. The persistence of tropical conditions throughout much of the Pleistocene in southern and Southeast Asia gives a very different faunal picture from that of western and Central Asia.

Game animals

PHYLOGENETIC AFFINITIES OF ASIAN FOSSILS TO MODERN MAN

Physical anthropologists are cautious in assuming phylogenetic affinities of living human populations to specific hominid fossil specimens of the Pleistocene, although some scholars have asserted that the living races of man can be recognized in the ancestral hominid record as far back as the middle Pleistocene. Beyond a few thousand years, however, physical resemblances of particular osteological or dental features between living and extinct populations become fewer in number and more tenuous as reliable data for establishing phylogenetic lines and classifications, and it is no longer a common practice in physical anthropology to attempt a "racial phylogeny" for a population beyond the limits of a few millennia. The traditional racial categories of Veddoid, Caucasoid, Mongoloid, Australoid, and the like are no longer regarded as particularly meaningful in relation to knowledge of prehistoric man.

In western Asia, the Natufian people of the Mesolithic of Israel do not resemble the present populations of the Near East; hence, it is not surprising that still earlier hominids, such as the Neanderthals of western Europe, fail to reflect obvious connections with contemporary Israelis, Lebanese, Syrians, etc. In Southeast Asia, the Neanderthaloid specimens from the sites of Solo and Wadjak in Java and Niah Cave in Borneo do not resemble the present inhabitants of mainland and island Southeast Asia, although some features of Neanderthaloid crania appear in the contemporary native populations of Australia and Melanesia. With the coming of Mesolithic (Middle Stone Age) post-Pleistocene cultures (after 12,000 BP) in Southeast Asia, the populations bearing these new traditions are represented in the osteological record by only a few specimens that are similar in anatomically significant ways to modern Southeast Asians. The earliest known fossil specimen of a Pygmy or Negrito population in Asia has come from Indochina, but it is only as ancient as the Neolithic Period of culture (*c.* 8000 BP). The antiquity of Pygmy populations in Malaysia, the Philippines, Andaman Islands, and New Guinea remains unknown, but it no longer seems reasonable to conceive of a "Pygmy race" binding these dwarfed Asiatic groups to Pygmies of Africa. Pygmy populations of Asia bear many more resemblances of physical characters to the macropopulations surrounding them than to more distant Pygmy populations.

Skeletons that can be reasonably identified as Chinese first appear in the Far East around 3000 BC, although some phenotypic variables, such as shovel-shaped incisors, which appear in high frequency in populations of Asiatic descent, occur as well in *Homo erectus* fossils from Chou-k'ou-tien. The high frequency of Chinese characters in the indigenous populations of Southeast Asia is explained by historic movements of Chinese into regions to the south of their ancient cultural area. In

southern Asia, the hominid skeletal record has been recognized as useful in drawing some phylogenetic affinities between living Indians and their ancestors of a few thousand years ago, but the evidence for Pleistocene man in India was established upon the discovery of stone industries rather than upon a rich fossil record. It is with the food-producing Neolithic (New Stone Age) cultures of Asia that biological similarities between extinct and contemporary human populations become apparent, but this represents an antiquity of only four or five millennia. It was from Asia that man moved last to the uninhabited vastness of Oceania, Australia, and the Western Hemisphere. Movements of Asian people into Europe has continued to the present day in eastern Europe and the Aegean.

BIBLIOGRAPHY. Literary sources on prehistoric Asia are numerous, but the most important ones constitute portions of more general works or appear as technical articles in scientific journals. General writings about prehistoric cultures of the Old World, which include discussions of Asiatic sites, are F. BORDES ,*La Palèolithique dans le monde* (1968; Eng. trans., *The Old Stone Age*, 1968); and J.G. CLARK, *World Prehistory: A New Outline* (1969). Regional studies are best represented by C.S. CHARD, "An Outline of the Prehistory of Siberia," *SWest. J. Anthrop.*, 14:1–33 (1958); H.L. MOVIUS, "Palaeolithic and Mesolithic Sites in Soviet Central Asia," *Proc. Am. Phil. Soc.*, 97:383–421 (1953), and "Palaeolithic Archaeology in Southern and Eastern Asia, Exclusive of India," *J. Wld. Hist.*, 2:257–282, 525–553 (1955); H.D. SANKALIA, *Prehistory and Protohistory in India and Pakistan* (1963). Critical sources on dating, stratigraphy, and cultural chronology include F.C. HOWELL, "Upper Pleistocene Stratigraphy and Early Man in the Levant," *Proc. Am. Phil. Soc.*, 103:1–65 (1959); and K.P. OAKLEY, *Frameworks for Dating Fossil Man* (1964). The physical anthropology of prehistoric Asians is discussed in the classic by M. BOULE and H.V. VALLOIS, *Les Hommes fossiles* (1921; Eng. trans., *Fossil Men*, 1957); and in C.S. COON, *The Origin of Races* (1962). Coon interprets the biological data according to a particular phylogenetic school that is not favoured by all contemporary anthropologists, but his discussion of the fossil record is excellent. Regional physical anthropology of prehistoric Asians is represented by K.A.R. KENNEDY, "The Search for Early Man in India," *S.S. Sarkar Memorial Volume* (1972); and by H.N. MICHAEL (ed.), *Ethnic Origins of the Peoples of Northeastern Asia* (1963). More specialized topics are treated by K.A.R. KENNEDY, "Paleodemography of India and Ceylon Since 3000 B.C.," *Am. J. Phys. Anthrop.*, 31:315–319 (1969); and in a series of articles in English, French, and German that appear in G.H.R. VON KOENIGSWALD (ed.), *Hundert Jahre Neanderthaler, Neanderthal Centenary, 1856–1956* (1958).

(K.A.R.K.)

Aśoka

Aśoka, one of the most significant early devotees of Buddhism, was the last major emperor in the Mauryan dynasty of India during *c.* 265–238 BC (others give *c.* 273–232 BC), his dominions extending over Afghanistan and the entire subcontinent except some southern territories.

The account below of Aśoka's life is based on his own numerous inscriptions or edicts, some short, some long, found at many places, rather than on Buddhist legends that are mostly fanciful or exaggerated. He conquered the Kaliṅga country (modern Orissa State) on the east coast after a very sanguinary war in the eighth year of his reign, but the terrible sufferings the war inflicted on the defeated people moved him to such great remorse that he renounced armed conquests forever. It was at this time that he came in touch with Buddhism and adopted it. Under its influence and prompted by his own dynamic temperament, he resolved to live according to, and preach, the *dharma* (principles of right life) and to serve his subjects and all humanity. He declared his nonaggressive intentions to his neighbours, assuring them of his good will toward them and sent envoys to distant kings bearing his good will and the message of the *dharma*. These actions he called his new policy of "conquest by *dharma*."

Aśoka's view of the *dharma*

By *dharma*, as he repeatedly declared, he understood the energetic practice of the socio-moral virtues of honesty, truthfulness, compassion, mercifulness, benevolence, nonviolence, considerate behaviour toward all, "little sin and many good deeds," nonextravagance, non-acquisitiveness, and noninjury to animals. He spoke of no particular mode of religious creed or worship, nor of any philosophical doctrines. He spoke of Buddhism only to his coreligionists and not to others.

Toward all religious sects he adopted a policy of respect and guaranteed them full freedom to live according to their own principles but he also urged them to exert themselves for the "increase of their inner worthiness." He, moreover, exhorted them to respect the creeds of others, praise the good points of others, and refrain from vehement adverse criticism of the viewpoints of others.

In order to gain wide publicity for his teachings and his work he made them known by means of oral announcements and also engraved them on rocks and pillars at suitable sites. From these inscriptions—the Rock Edicts and Pillar Edicts (*e.g.*, the lion capital of the pillar found at Sarnath, which has become India's national emblem)—mostly dated in various years of his reign and containing statements regarding his thoughts and actions his life and acts are known. There is such a ring of frankness and sincerity in the utterances of Aśoka that they appear to be true.

To practice the *dharma* actively Aśoka went out on periodical tours preaching the *dharma* to the rural people and relieving their sufferings; he ordered his high officials to do the same, in addition to attending to their normal duties; he exhorted administrative officers to be constantly aware of the joys and sorrows of the common folk and to be prompt and impartial in dispensing justice. A special class of high officers, designated "*dharma* ministers," was appointed to foster *dharma* work by the public, relieve sufferings wherever found, and look to the special needs of women, of people inhabiting outlying regions, of neighbouring peoples, and of various religious communities. It was ordered that matters concerning public welfare were to be reported to him at all times. The only glory he sought, he said, was for having led his people along the path of *dharma*. No doubts are left in the minds of readers of his inscriptions regarding his earnest zeal for serving his subjects. More success was attained in his work, he says, by reasoning with people than by issuing commands.

Aśoka's public works and services to Buddhism

Among his works of public utility were the founding of hospitals for men and animals and the supply of medicines; planting of roadside trees and groves, digging of wells, and construction of watering sheds and rest-houses. Orders were also issued for curbing public laxities and preventing cruelty to animals.

With the death of Aśoka the Maurya Empire disintegrated and his work was discontinued. His memory survives for what he attempted to achieve and the high ideals he held before himself.

Most enduring were Aśoka's services to Buddhism. He built a number of *stūpas* (commemorative burial mounds) and monasteries and erected pillars on which he ordered inscribed his understanding of religious doctrines. He took strong measures to suppress schisms within the order (the Buddhist religious community) and prescribed a course of scriptural studies for adherents. Tradition recorded in the Ceylonese chronicle *Mahāvaṃsa* says that when the church decided to send preaching missions abroad, Aśoka helped them enthusiastically and sent his own son and daughter as missionaries to Ceylon. It is as a result of Aśoka's patronage that Buddhism, which until then was a small sect confined only to particular localities, spread throughout India and subsequently beyond the frontiers of the country.

A sample quotation that illustrates the spirit that guided Aśoka is: "All men are my children. As for my own children I desire that they may be provided with all the welfare and happiness of this world and of the next, so do I desire for all men as well."

BIBLIOGRAPHY. AMULYACHANDRA SEN (ed.), *Asoka's Edicts* (1956), deals with all aspects of Aśoka's life and work on the basis of archaeological and literary materials. D.R BHANDARKAR, *Asoka*, 3rd ed. (1955); and R.K. MOOKERJEE, *Asoka*, 3rd ed. (1962), are also good studies based on historical materials.

(A.Se.)

Asquith, H.H.

Liberal British prime minister from 1908 to 1916, Herbert Henry Asquith was responsible for the Parliament Act of 1911 (limiting the power of the House of Lords) and led Britain during the first two years of World War I.

Radio Times Hulton Picture Library



Asquith.

Asquith was born at Morley, Yorkshire, on September 12, 1852, the second son of Joseph Asquith, a small businessman in the wool trade and an ardent Congregationalist, who died in 1860. Asquith was educated at the City of London School from 1863 to 1870 when he won a classical scholarship at Balliol College, Oxford. At Balliol he obtained the highest academic honours, and became a fellow of his college in 1874. Deciding upon a legal career, he entered Lincoln's Inn and was called to the bar in 1876. The following year he married Helen Melland, daughter of a Manchester doctor, by whom he had four sons and one daughter. His early days at the bar were difficult, but from about 1883 onward he became highly successful.

A keen Liberal, Asquith entered the House of Commons for East Fife in 1886 and remained its member for 32 years. He commanded the attention of the House from the first, concentrating particularly upon the Irish question. In 1888 he achieved celebrity as junior counsel for the Irish leader Charles Stewart Parnell, when Parnell was accused, before a parliamentary commission, of condoning political murder. In 1892 Gladstone made Asquith home secretary. Before that, in 1891, his wife had died of typhoid fever, leaving him with a family of young children. Less than three years later he astounded the social and political world by marrying Margot Tennant, who was 12 years younger and the centre of social and intellectual circles far removed from those in which Asquith and his first wife had moved.

His three years as home secretary, though in general an unhappy period for the Liberals, established Asquith's reputation as an administrator and a debater. By 1895 he had become one of the leading figures of his party. Defeated at the polls, the party spent the next 11 years in opposition. Asquith earned during this time a large income at the bar, but the lack of any private means obliged him to refuse the party leadership when it was offered to him in 1898, and Sir Henry Campbell-Bannerman succeeded instead. Asquith did not see eye to eye with the new leader on all questions of foreign and imperial policy. Their divergence became open and public during the South African War (1899–1902), when Asquith, along with Lord Rosebery, Sir Edward Grey, and R.B. Haldane, formed the Liberal League to advocate an imperial policy in support of the government's expansionism. The conflict was temporarily healed after the end of the war and following the Liberals' victory at the polls in 1906, Asquith served as chancellor of the exchequer under Campbell-Bannerman.

Early in April 1908 Campbell-Bannerman resigned and died some days later. Asquith, generally regarded as his inevitable successor, became prime minister and was to hold the office for nearly nine years. He appointed David Lloyd George to the Exchequer and made Winston Churchill president of the Board of Trade. The chief problem confronting him at home was the opposition of the House of Lords to Liberal reforms, and the consequent danger of a rebellion from the frustrated radicals in his own party; abroad there was a growing naval competition with Germany. When Lloyd George endeavoured to raise money for naval increases and social services in his "radical budget" of 1909, the budget was vetoed by the House of Lords.

Parliamentary reform

At this stage Asquith took over the conduct of a constitutional struggle. In 1910 he announced a plan to limit the powers of the upper house, and, after two general elections, persuaded King George V to threaten to create enough new pro-reform peers to swamp the opposition. The resulting Parliament Bill, passed in August 1911, ended the Lords' veto power.

The three years between the end of this episode and the outbreak of World War I were extremely harassing for the prime minister. Abroad, the international situation deteriorated rapidly; at home, controversy was caused by charges of corruption in his government, the disestablishment of the Anglican Church in Wales (1914), and the conflict between Home Rulers and Unionists in Ireland, which nearly led to civil war in 1914. Asquith's policies did little to improve the situation in Ireland.

Though convinced that a German victory over France would be disastrous to the British Empire, Asquith delayed Britain's entry into World War I until public opinion had been aroused by the German attack on Belgium. In war, he trusted his military experts and in general favoured the school that maintained that victory could be won only on the western front.

In May 1915 Asquith had to reconstruct his Cabinet on a coalition basis, admitting Unionists as well as Liberals, and appointing Lloyd George minister of munitions. The coalition was not successful under his leadership. The Dardanelles expedition failed and there was no sign of a breakthrough in the west. At the end of 1915 Asquith substituted Sir Douglas Haig for Sir John French as British commander in chief in France, and appointed Sir William Robertson as the new chief of the imperial general staff. But 1916 was an even unhappier year: the Easter Rising in Dublin caused a grave domestic crisis, and the battle of the Somme led to a complete impasse on the western front. After a protracted struggle, conscription was belatedly introduced. But there was a general aura of dissatisfaction by the autumn, and Asquith was assailed by a strident press campaign. In December he resigned and was replaced by Lloyd George. He never held office again, though he remained leader of the Liberal Party until 1926. In this capacity he often opposed the policies of his successor.

Asquith accepted a peerage as earl of Oxford and Asquith in 1925, and was created a Knight of the Garter shortly afterward. In the last years of his life he was relatively impoverished and wrote a number of books to make money, the best known being *The Genesis of the War* (1923), *Fifty Years of Parliament* (1926), and *Memories and Reflections* (1928). He died at Sutton Courtenay on February 15, 1928, and was buried in the graveyard of the parish church there.

Asquith was a competent statesman, but not a great one. He had no original or innovating genius and lacked the sense of the dramatic needed to convince Britain that it was in good hands in a time of national crisis.

BIBLIOGRAPHY. J.A. SPENDER and CYRIL ASQUITH, *Life of Herbert Henry Asquith, Lord Oxford and Asquith*, 2 vol. (1932), is the official biography, cautious, discreet, and highly favourable. The most up-to-date life is ROY JENKINS, *Asquith* (1964), shorter and more critical than the official biography, though in general pro-Asquith. ROBERT BLAKE, *The Unknown Prime Minister: The Life and Times of Andrew Bonar Law, 1853–1923* (1955), gives the Conservative side of the period;

THOMAS JONES, *Lloyd George* (1951), and FRANK OWEN, *Tempestuous Journey* (1954), give the Lloyd Georgite point of view. Other books worth consulting are A.J.P. TAYLOR, *English History, 1914–1945* (1965); RANDOLPH CHURCHILL, *Winston S. Churchill: The Young Statesman, 1901–14* (1967); and MARTIN GILBERT, *Winston S. Churchill, 1914–16* (1971).

(B.)

Assam

One of the 21 states of the Indian Union, Assam is located in the northeast of India, from which it is almost isolated on the west by Bangladesh. It has an area of 30,452 square miles (78,890 square kilometres) and a population (1971) of about 14,625,000. Until an administrative reorganization in 1972, the Union Territories of Arunachal Pradesh (formerly the Northeast Frontier Agency [NEFA]) and Mizoram and the state of Meghalaya were part of Assam.

Assam is bounded to the north by Bhutan and Arunachal Pradesh; to the east by Arunachal Pradesh, Nagaland and Manipur; to the south by Mizoram and Meghalaya; and to the west by West Bengal, Tripura, and Bangladesh (formerly East Pakistan). To the northwest, a narrow corridor running through the foothills of the Himalayas connects the state with India. Geographically, Assam has no natural boundaries. The capital of the state, formerly located at Shillong (now the capital of Meghalaya), was temporarily shifted to Dispur, a suburb of Gauhāti, in 1972. (For associated physical features, see BRAHMAPUTRA RIVER; HIMALAYAN MOUNTAIN RANGES; for historical background, see INDIAN SUBCONTINENT, HISTORY OF THE.)

History. In the earliest times the territory and its environs were known as Kāmarūpa, a state that had its capital at Prāgjyotiṣapura (modern Gauhāti, in Assam on the southern bank of the Brahmaputra). Ancient Kāmarūpa included roughly the Brahmaputra Valley, Bhutan, the Rangpur district (now in Bangladesh), and Cooch Behār, in West Bengal. King Narakāsura and his son Bhagadatta were famous rulers of Kāmarūpa in the Mahābhārata period, at least 1000 BC. A Chinese traveller, Hsüan-tsang, who visited the country when it was ruled by Bhāskaravarman, about AD 640, left a vivid account of the country and its people. After this, information about ancient Assam is very meagre for a few centuries, but the discovery of several copperplates, clay seals, and stone inscriptions that date from the 7th to the mid-12th centuries has thrown some light on the condition of the country at that time. These archaeological records indicate that the inhabitants of the region attained considerable power and a fair degree of civilization. The copperplates further provide clues to the locations of important ancient settlements and also the routes connecting them.

Rule by the Ahoms

Assam was ruled by various dynasties—the Pālas, Koches, Kachāris, and the Chutiyas—but, because of the constant warfare amongst these princes, there was no stable government in the area until the coming of the Ahoms in the 13th century. The Ahoms crossed the Pātkai Range from Burma and conquered the local chieftains who then ruled the Upper Assam Plain. In the 15th century the Ahoms, who gave their name to the country, were the dominant power in Upper Assam. Two centuries later, after defeating the Koches, the Kachāris, and other local rulers, they also became the rulers of Lower Assam up to Goālpāra. The power and prosperity of the Ahoms reached a zenith during the rule of King Rudra Singh (1696–1714).

Dissension and jealousy among the princes gradually weakened the central administration, with the result that, in 1786, the ruling prince Gaurinath Singh, in despair, sought the aid of the British in Calcutta. Captain T. Welsh, sent by the British governor general in India, restored peace and was recalled in spite of the protests of the Ahom king. Internal strife then caused one crisis after another until, in 1817, the Burmese entered Assam in response to the appeal of the *bar phukan* (governor) Badan Chandra, a rebel against the king. They swept over the country thrice, bringing destruction and misery.

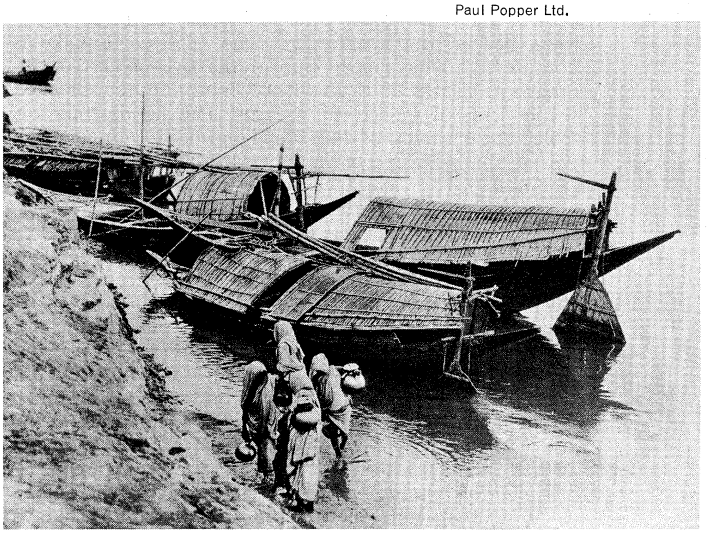
The British, whose interests elsewhere were threatened by these developments, ultimately drove out the Burmese invaders, and, after the Treaty of Yandabo was concluded with Burma in 1826, Assam became a part of British India.

A British agent, representing the governor general, was appointed to administer first the devastated area of Assam. In 1838 Assam was incorporated into British-administered Bengal. By 1842 the whole of Assam Valley had come under British rule. In 1874 a separate province of Assam was created under a chief commissioner, with its capital at Shillong. In 1905, on the initiative of the British viceroy, Lord Curzon, the province was amalgamated with the eastern part of Bengal. The subsequent partition of Bengal was, however, resented, and in 1924 this arrangement was reversed, with Assam once more made a separate province. During World War II, Assam became one of the main supply routes for Allied forces operating in Burma, and within the territory of Assam decisive battles were fought in 1944 in a number of places (including Bishenpur, in Manipur, and Kohīma, in Nāgāland). These battles were instrumental in preventing the Japanese advance into India.

After India achieved independence in 1947 and as a result of the ensuing partition, the district of Sylhet (excluding the Karīmganj subdivision) was ceded to Pakistan. In 1961 and 1962 Chinese armed forces, disputing the McMahon Line as the boundary between India and Tibet, occupied part of the NEFA. In December 1962, however, they voluntarily withdrew to Tibet. In 1963 the Nāga Hills district, formerly a part of Assam, became the 16th state of the Indian Union, under the name of Nāgāland (*q.v.*). A part of Tuensang, a former territory of NEFA, was also added to Nāgāland. In 1970, after the tribal people of the Gāro district and the United Khāsi and Jaintia Hills district, located in the centre of Assam, had also demanded a separate state, an autonomous state named Megalaya (*q.v.*) was created with these two districts. In 1972, the North East Frontier Agency was renamed Arunachal Pradesh and was separated from Assam and the Union Territory of Mizoram was created in the south.

The landscape. *Physiography.* Except for the districts of Mīkīr Hills and North Cāchār Hills, Assam is generally comprised of plains and river valleys. The state may be divided into three principal physical regions—the Brahmaputra Valley in the north, the Barāk Plain in the south, and the hilly regions that lie between these two valleys.

Paul Popper Ltd.



Sampan skiffs on the Brahmaputra River, Assam, India.

The valley of the Brahmaputra River (*q.v.*) is the dominant physical feature of Assam. The river, which is known as the Ya-lu-tsang-pu Chiang in Tibet, enters Assam near Sadiya, after which it runs directly westward for nearly 450 miles, traversing every district across the

length of Assam before turning due south to enter the plains of Bangladesh. The river valley, rarely more than 50 miles wide, is studded with numerous groups of hills that rise abruptly from the plain. The valley is surrounded on all sides, except on the west, by mountains and is intersected by many streams and rivulets that flow down from the neighbouring hills to empty into the Brahmaputra.

To the southeast of Meghalaya is the small but important Barāk Valley, about 125 miles long and about 60 miles wide. The valley supports a dense population.

Geologically, the Brahmaputra Valley and the Barāk Valley are entirely alluvial in origin and consist of sand, sandstone pebbles, clay, and sometimes a mixture of sand and clay containing decomposed vegetable matter. The Tertiary deposits (from 2,500,000 to 65,000,000 years old) include a similar variety of rocks, among them hard sandstone, soft and loose sand, conglomerates, coal seams, shales, and sandy clays. The lowest Tertiary deposits in some parts of Assam include limestone.

The Mīkīr Hills and part of the North Cāchār Hills form part of the Meghalaya Plateau, which may have been an extended portion of Gondwanaland (an ancient landmass in the Southern Hemisphere that once grouped together South America, Africa, Australia, and part of the Indian subcontinent). This truncated portion of the main plateau displays a rugged topography, with shales of the Barail series exposed in valley bottoms. It has, roughly, a northerly slope, with the outer ranges of the Mīkīr Hills reaching an average elevation of about 1,500 feet (450 metres) above sea level. Higher average elevations of about 3,300 feet (1,000 metres) are reached in the central portion of the hill district east of the Kopili River. (The river has formed plain embayments into the interior of the Mīkīr Hills, largely isolating this section from the main plateau.)

The northern ranges, which extend from Dabaka in the southwest to Bokākhāt in the northeast, attain an average elevation of 2,000 feet (600 metres). Major peaks in the north include Basundhari Parbat (2,540 feet), Raisang (2,420 feet), Mehekongthu (2,095 feet), and Kud Parbat (2,055 feet).

The southern ranges, known as the Rengma Hills, have an average elevation of about 3,000 feet (900 metres). Main peaks in the south are Chenghehishon (4,200 feet) and Khunbaman Parbat (4,300 feet).

Earthquakes

Earthquakes are a common phenomenon in Assam. In modern times the most important Assamese earthquakes have been those that occurred in 1897, with the Shillong Plateau as the epicentre; 1930, with Dhubri as the epicentre; and 1950, with Ch'a-yü (Rima) in Tibet at the NEFA border as the epicentre. The 1950 earthquake is considered to have been one of the five biggest earthquakes in history. It created heavy landslides that blocked the courses of many hill streams. The floods that followed the bursting of these artificial dams caused more loss of life and property than the earthquake itself. (H.D.)

Climate. The average temperature is moderate—about 84° F (29° C) in the hottest month (August). In the cool season, when the northeast monsoon (rain-bearing) wind blows down the Brahmaputra Valley, the average valley temperature in January is 61° F (16° C). At this season the climate in the valley is marked by fogs and a little rain. Assam escapes the normal Indian hot, dry season. Some rain occurs from March onward, but the real force of the monsoon winds is felt from June onward, when they blow at right angles to the Assam hills. By the time the winds have crossed the hills they bring little rain to the valley; Gauhāti receives only about 67 inches. (L.D.S./Ed.)

Vegetation and animal life. The forested area covered 18,500 square miles before the creation of Meghalaya and Mizoram and formerly represented nearly 39 percent of the total area of the state, of which reserved forest, which is efficiently managed by the government, constituted only about 6,400 square miles, or about 12 percent of the total area. The creation of both Meghalaya and Mizoram, however, substantially reduced the state's forest area.

A portion of the reserved forest in Assam is devoted to wildlife sanctuaries. There are several such sanctuaries, of which Kāziranga, the stronghold of the fast-disappearing one-horned rhinoceros, is the most famous. The most important forest products are timber and bamboo, firewood, and lac (shellac). There are about 74 species of timber, of which 49 varieties are commercially exploited. The forests are inhabited by many wild animals.

Population. *Demography.* During the 70 years from 1901 to 1971 the population of what is now Assam increased from 3,290,000 to 14,625,000—an increase of 345 percent, compared with an average increase of 130 percent for India during the same period. The distribution of population is uneven owing to the hilly terrain, the number of rivers, the forests, the small amount of cultivable land, and the lack of industrialization. In the plains the average density is 577 per square mile, whereas in the hills it was only 77.

The unusually rapid growth of population in the state has been due mostly to the immigration into Assam of tea-garden labourers, Nepali graziers, West Bengal Muslims, and Hindu refugees from what is now Bangladesh.

The rural element amounted to some 92 percent of the population of Assam in 1971. Villages with a population of 2,000 or more in 1961 accounted for 50 percent, and 42 percent lived in smaller villages or clusters.

Urbanization

The increase of urban population during the 1960s was significant. The growth of many industries, increased commercial activity, and the desire of the Bangladesh refugees to live near towns were some of the causes of the increase in urban population. Only Gauhāti had a population of more than 100,000. There are 11 towns of more than 20,000 persons. Some areas, such as Badarpur *thānā* (police district), with a 1961 density of 1,388 persons per square mile, have become overpopulated.

Ethnic composition, linguistic patterns, and religious affiliations. The people of the plains districts of the Brahmaputra and Barak valleys are mainly of Indo-Iranian stock, although many are of Mongoloid stock, particularly in the upper Assam region. By the time of their arrival in the Brahmaputra Valley, it would appear that the original Aryan people of Assam had become intermixed with other peoples. Assamese is the principal language and is regarded as the lingua franca of the whole of northeast India. Although scholars trace the history of the Assamese literature from the beginning of the second millennium AD, yet an unbroken record of literary history is traceable only from the 14th century. The majority of the people of the Cāchār district in the Barak Valley speak Bengali.

About two-thirds of the Assamese are Hindus; about a quarter of the population is Muslim. The Muslims are mostly recent settlers from East Bengal, or converts belonging to the lower strata of the peoples of Indo-Iranian origin. A majority of the Hindus accept Vaiṣnavism, which is based on the deity Vishnu.

The tribes in the hills, and also those in the plains, are of Mongoloid stock. They speak dialects of Tibeto-Burman origin, with the exception of the Khasi dialect, which is derived from the Austroasiatic branch of languages. Many of the hill tribes have been converted to Christianity by missionaries, but the majority still observe the customs and festivals of their traditional religion, which is based on animism and has a close affinity to the ancient form of Hindu worship. The Mikirs and Kachāris of the Mīkīr and North Cāchār Hills are mostly Hindus; although they speak dialects of Tibeto-Burman origin, they have adopted Assamese as their first language.

Administration and social conditions. The state government has a unicameral legislature under a governor and a Cabinet. The state of Assam comprises 10 districts —Goālpāra, Kamrup, Nowgong, Darrang, Sibsāgar, Dibrugarh, and Lakhīmpur in the Brahmaputra Valley; Cāchār in the Barak Plain; the Mīkīr Hills; and the North Cāchār Hills. (H.D./Ed.)

Education

In 1968–69 an estimated 2,151,000 children were receiving education, of whom more than 60 percent were boys; 90 percent came from rural areas. Primary educa-

tion was being given to about 75 percent and secondary education to about 25 percent of the children in the respective age groups. Compulsory primary education was in force for children between the ages of six and 12 in 13 towns and about 4,400 villages. The 1971 census recorded that literacy among males was 37 percent and among females only 18 percent. There are universities in Gauhāti, Jorhāt, and Dibrugarh. In 1955 birthrates and death rates were, respectively, about 40 and about 16 per 1,000 persons. Seventeen welfare-extension projects, operating through more than 80 centres and covering 300,000 people, were providing recreational and cultural facilities for women and children. (S.B.L.N.)

The economy. *Agriculture.* Agriculture is of basic importance to Assam. In the early 1970s, about 56 percent of the total working population were directly engaged in agriculture, with another 10 percent employed on the plantations, in forestry, or in other occupations related to agriculture.

Rice accounted for about 70 percent of the sown area. In 1968 to 1969 tea covered an area of almost 520,000 acres and jute about 270,000 acres. These two crops are important foreign-exchange earners. In the early 1960s, Assam grew nearly 50 percent of India's tea and 25 percent of its jute. Cultivation in the Brahmaputra Valley is extensive; other crops grown there include oilseeds, pulses (leguminous plants, such as peas, beans, or lentils), sugarcane, rape (an oil-yielding plant, the leaves of which are used for fodder), mustard, and potatoes.

Fruits that are grown include oranges, pineapples, and bananas. The state produces a surplus of cereals, but there is a shortage of oilseeds and pulses. Double-cropping (growing two crops a year) and other improved methods of cultivation were being introduced in the early 1970s.

The tea gardens

Tea cultivation in Assam was begun as early as 1835. Before Assam's reorganization in 1972, there were almost 750 tea gardens, producing nearly 44,000 pounds of tea annually.

Minerals. In the early 1970s the known minerals of the state were petroleum, coal, limestone, fireclay, china clay, and feldspar. Of these, oil, coal, and limestone were being commercially exploited.

Oil is found at Digboi, Naharkatia, Hugrijan, Moran, Rudrasagar, and Lakowa—all in Upper Assam. Coal is found in the Dibrugarh district of Upper Assam and in the Mīkīr Hills. Railways, plantations, and steamships are the main consumers of Assam coal. Limestone is quarried in the Mīkīr Hills.

Industry. In spite of its raw materials, industrial growth in Assam is in its infancy, for—with the exception of tea and oil—there are few industries of significance. Industrial development is handicapped mainly by Assam's isolation from the rest of India, by a bad transport system, by a small local market, and by the lack of sufficient capital. A number of industrial enterprises have nevertheless been started. These include a fertilizer plant at Namrup, a jute mill at Silghāt, a sugar mill at Dergaon, a paper mill at Jogighopa, a spun-silk mill at Jagiroad, and a cement factory at Bokajān. There are also many sawmills and plywood and match factories that make use of the timber supplies of Assam. A number of food-processing units for rice, oil, and fruit were being established in the early 1970s. The oil refinery at Digboi was built as early as 1899; the Noonmati refinery near Gauhāti started production in 1962. The Bāruni refinery is in Bihār state but is supplied crude oil from Assam through a pipeline. A petrochemical complex is planned for Bongaigaon.

Transport and communications. A poor transport and communication system has been the main hindrance to the economic development of Assam. Though the four recognized forms of transport—road, rail, air, and water—all exist, none satisfactorily serves the needs of the region. The isolated location of Assam, linked as it is with the rest of India by only a narrow corridor between the Himalayas on one side and Bangladesh on the other, has to a great extent prevented the development of quick and efficient road and rail systems in the region. The growth of an inland water-transport system has similarly been checked by the inclusion of the lower part of the Brahmaputra in Bangladesh. Assam had about 1,400 miles of railways, about 12,000 miles of motorable roads, and 6,000 miles of inland waterways in the early 1970s. There was also a considerable amount of airborne traffic between Assam and Calcutta. The important airports are Borjhār (Gauhāti), Mohanbāri (Dibrugarh), Lilābāri (North Lakhīmpur), Rowroyāh (Jorhat), Sāleni (Tezpur), and Kumbhirgrām (Silchar). The airports are used by the Indian Airlines Corporation (IAC) for civil air transport of passengers and freight.

Cultural life and institutions. The cultural life of Assam is interwoven with the activities of a number of cultural institutions and religious centres like *satra* (seat of a religious head, the Satrādhikār) and *namghar* (prayer hall). Several *satras* established at different places in Assam have been looking after the religious and social well-being of the villagers as well as of some urban dwellers for the last 400 years. The Assamese people observe all the pan-Indian religious festivals such as Durgā-pūjā, Dol-jātrā or Holī (*i.e.*, festival when colour is sprinkled with joy and a sentiment of brotherhood), and Janmastami (birthday of Lord Krishna). But the most important social and cultural celebrations are the three Bihu festivals observed with great enthusiasm irrespective of caste, creed, and religious affinity. The Bihus were originally agricultural festivals observed by the villagers at different seasons of the year. The Bohāg Bihu, celebrated in mid-April with the commencement of the new year (first day of the Bohāg or Baiśākh month which falls usually on the 14th of April) during the spring season, is the most important one. It is also known as Rangāli Bihu (*rang* means merry-making and fun). It is observed by dancing and singing in open spaces as well as in the houses of villagers. The mother or other ladies of the family present a hand-woven *gāmochā* (towel) to each and every member of the family on this day.

The second important Bihu, the Māgh Bihu, celebrated in mid-January (in the month of Māgh), is a harvest festival. It is performed with community feasts and bonfires. It is also known as Bhogāli Bihu (*bhog* means enjoyment and feasting). The third Bihu festival, the Kāti Bihu, is observed in mid-October. It is also known as the Kangāli Bihu (*kangāli* means poor) because by this time of a year, the house of a common man is without foodgrains as the stock is usually consumed before the next harvest.

Another important aspect of the cultural life of the people of Assam, particularly of the women, is weaving of fine silk and cotton cloths with various floral and other decorative designs. Every Assamese house, irrespective of caste, creed, and social status, has at least one loom and each grown-up girl is required to know the art of weaving.

BIBLIOGRAPHY

General references: Census of India, 1961, vol. 3, *Assam*, part 1-A, *General Report* (1964), a general collection of data on various geographical, economic, cultural, and social aspects of the state; H.P. DAS, *Geography of Assam* (1970), an illustrated geographical account of the state; JOHN P. WADE, *An Account of Assam*, 2 parts, ed. by BENUDHAR SHARMA (1927), an important reference book including the historical geography of important places.

Specialized references: P.C. CHOUDHURY, *The History of Civilization of the People of Assam to the 12th Century, A.D.*, 2nd ed. (1966), a study of prehistoric and ancient Assam; S.K. BHUYAN, *Anglo-Assamese Relations, 1771–1826* (1949), a detailed analysis of the late Ahom and early British period; see also *Ahom-Buranji* (1930), a chronicle of Ahom rule in the Ahom language with a parallel English translation; D.N.D. GOSWAMI, *Geology of Assam* (1960), a systematic treatment of the subject; P.D. STRACEY and H.P. DAS, *Assam's Economy and Forest* (1949); and M.C. JACOB, *The Forest Resources of Assam* (1940), classifications and evaluations of forest resources; DEPARTMENT OF ECONOMICS AND STATISTICS, ASSAM, *Statistical Handbook, Assam* (annual), a pocketbook collection of important data; P.C. GOSWAMI, *The Economic Development of Assam* (1963), an analysis of various resources and their utilization.

(H.D.)

Association Football (Soccer)

Association football, popularly known as soccer, is the only form of football played with a round ball, which can be more readily controlled and encourages a more open game than the oval-shaped variety. Because tackling is less violent than in other forms—American football, for example—soccer can be played in school playgrounds and streets. In Rio de Janeiro the beaches are given over each weekend to matches between barefoot teams, and the quality of play explains why Brazil produces such fine players.

The highly skilled professionals and international players are the apex of a vast pyramid embracing millions of schoolboys, youths, amateurs, and minor professionals of all grades and qualities.

The patterns and clarity of purpose of the game make it easy to understand and easy to play. Scoring a goal is a rare and dramatic achievement because the ball has to be propelled into a target 8 yards (7.3 metres) wide and 8 feet (2.4 metres) high, guarded by a goalkeeper and protected by defenders. While soccer has an attraction for the casual onlooker and occasional player, it also has a depth of virtuosity that delights the connoisseur. The game looked very simple when played by the great Hungarian side (team) of the 1950s but, in fact, their easy style was based on supreme skill and common understanding, the result of years of training individually and together.

Soccer orginated in Great Britain, where it was refined into a sophisticated team game. The first competition between England and Scotland was held in 1872. Eleven years later it was enlarged into the British Championship, with Wales and Northern Ireland taking part. Since then a host of international tournaments, at club and national level, has sprung up.

The governing body of association football, the Fédération Internationale de Football Association (FIFA) has more members than the United Nations. Its membership was brought to 135 at the 1970 congress, compared with 126 in the UN. It is estimated that 600,000,000 to 1,000,000,000 people followed the 1970 World Cup in Mexico by television and radio. These facts are striking evidence of the tremendous popularity of the game, both as a participating sport and an entertainment.

This article is intended for the general reader who may have no knowledge of the sport. Therefore, in addition to tracing the history and outlining the present status of association football, the article is designed to help a spectator understand a game he may be watching. For information on the specific rules or how to play, the reader should consult the works listed in the bibliography.

History of the game. Where football began nobody knows for certain. A primitive form is mentioned in a Chinese military text of the 3rd or 4th century BC. Roman soldiers also had a variety of football called *harpastum,* and there are evidences of football-like games in ancient Greece, Mexico, and Japan.

The Romans spread football through Europe, where it developed during the Middle Ages, notably in Italy, into a rough sport often played between towns, involving hundreds of players and with the goals perhaps half a mile apart. These mob games, called *mêleés* or mellays, were boisterous and dangerous. There were many attempts to ban them, including a proclamation by Edward II of England in 1314: "We command and forbid, . . . on pain of imprisonment, such games to be used . . . in future." As indicated by Shakespeare's reference in his *Comedy of Errors,* Act II, however the game persisted:

> Am I so round with you as you with me
> That like a football you do spurn me thus?
> You spurn me hence and he will spurn me hither;
> If I last in this service you must case me in leather.

British public schools and universities, particularly Cambridge, took up football in the first half of the 19th century. It was the period of "muscular Christianity," of cold baths, rigorous discipline, and fair play, and the game was a means of achieving these virtues. Different rules persisted despite an attempt by Cambridge University to unify them in 1846.

Development of organization. The Football Association (FA) was brought into being on Oct. 26, 1863, for the express purpose of establishing a uniform set of rules. Although William Webb Ellis sowed the seeds of Rugby football by carrying the ball at Rugby School as early as 1823, the FA hoped to fuse both codes by permitting a player who caught the ball to run with it. Discussions over three months, however, failed to bridge the gap and Rugby adherents, led by the Blackheath club, broke away.

Separation from Rugby

Soccer became one of Britain's greatest exports, spread by soldiers, sailors, merchants, engineers, priests. The Danes were the first to take to it, forming the Boldspil Union in 1889. Sailors took the game to Brazil, and two Englishmen launched it in Russia through their cotton mill in Orekhovo-Zuyevo. In the United States the universities were the first disciples, and the names of early clubs, such as the Kensington F.C. of Saint Louis and Shamrock F.C. of Cincinnati, indicated their British origin (see also FOOTBALL, AMERICAN AND CANADIAN).

By the turn of the century, the standard had risen dramatically throughout the world; and in 1902 Austria and Hungary played the first of their more than 100 international tournaments, to be followed three years later by Argentina and Uruguay. Europe looked to England to take the lead in establishing a world governing body. The English FA, however, "in a monumental example of British insularity" as the FA's official history admits, coldshouldered the approach.

On May 21, 1904, Belgium, Denmark, France, Holland, Spain, Sweden, and Switzerland founded FIFA in Paris. After years of mind changing, the British countries, led by progressive FA secretary Sir Stanley Rous, joined the fold in 1946. Sir Stanley left the FA to become FIFA President in 1960 and continued his efforts to unify the game's adherents. By this time, alarming differences in interpretation of rules had developed, particularly in South America, where the shoulder charge, harassing the goalkeeper, and tackling from behind were abhorred, while interfering with an opponent not playing the ball was tolerated. The differences explained the unseemly clashes in unofficial world club championship play between representatives of Europe and South America from 1960 on.

FIFA's growth was steady. By its 25th birthday there were 40 members, and by 1963 there were more than 100. FIFA has headquarters in Zurich and is association football's ultimate governing body. The power to alter the rules, however, still lies mainly with the British countries, which have four votes to the rest of the world's two on the International Board. There are six regional confederations under FIFA—Europe, South America, Central and North America, Asia, Africa, and Australasia—and each has wide control over discipline and competitions in their geographical areas.

Rule changing authority

In England the mob football, which had been refined by the upper classes at public schools and universities, was taken back avidly by the people, with a significant effect on the rest of the world. Professionalism was legalized in 1885 by the English FA, and that in turn led to the Football League being formed by some of the strongest clubs three years later in order to stimulate interest and thereby attendances. Teams played each other both at home and away, receiving two points for a win and one for a draw. There were 12 founder clubs, including the still famous Everton, Aston Villa, Derby, and Wolverhampton Wanderers. Today the league has 92 members, divided into four divisions. At the end of a season, the top teams of the lower divisions are promoted into the places of low-ranking teams in the upper divisions.

Practically every country now has a professional league modelled on England's. In Brazil the vast distances have led to separate leagues, one embracing the teams in the Rio de Janeiro area and the other those around São Paulo. The U.S.S.R. and Germany used to have regional leagues, but improved transportation permitted national mergers.

The standard of play rose sharply through the stimulus of competition and full-time training of the players. The game also became increasingly commercialized, and outstanding players were idolized and paid princely wages. Pele, of Brazil, probably the finest footballer the world has known, earned around $250,000 a year at his peak, from the game itself, from advertising, from a television series, from selling his name for use on a brand of coffee, and so on. Ironically, England clung to a maximum wage of $56 a week until 1961, when a strike threat by the Players' Union forced wage hikes.

Large sums of money are paid for the transfer of a player from one club to another. The record fee of over a million dollars was paid by Juventus of Turin to Varese for the Sicilian forward Piero Anastasi. The deals are not confined to teams in the same association, and Italian clubs have been the importers on the grandest scale. They raided Argentina in the 1930s, not merely signing several players but also fielding them in the 1934 World Cup winning team, claiming they had Italian ancestry. Because the wholesale use of foreign stars hindered development of their own youngsters, the national team's performance declined, and import was banned in 1964.

In England there is often a conflict of interests between the big clubs, which look to their parent body, the league, for support, and the association, which is in the main an amateur organization. The association usually selects the international team, but the clubs may claim that they have first call on their players because they produce, groom, and pay them.

Soccer in the United States

The intense interest shown in the televising of the 1966 World Cup competition encouraged sponsors to believe that soccer would flourish in the United States. Foreign players were imported to make up teams. Unfortunately, the formation of rival leagues, the United Soccer Association and the National Professional League, split what audience there was, and clubs lost up to $1,500,000 each in the opening season. Most Americans continued to regard association football as an alien game played only by ethnic groups.

Three years later the U.S. leagues merged and the 17 teams were reduced to six. More important, the drive to foster the game at the grass roots was paying off, and more high schools and colleges organized teams.

Competition. One of the aims of FIFA was to organize a world championship, and the move took substance after soccer was a success in the 1920 and 1924 Olympic Games. Jules Rimet, president of FIFA from 1921 onward, was a driving force in establishing the championship, and the first gold trophy was named after him.

In 1930, because of travelling difficulties and the unwillingness of clubs to release players, only 13 countries took part in the first tournament in Uruguay; in the competitions in 1966 and 1970, by contrast, there were more than 70 entries, necessitating regional tournaments to reduce the final number to 16. Uruguay became the first champion and won again in 1950—a remarkable achievement for a country with fewer than 3,000,000 inhabitants.

World Cup competition was to be staged every four years between Olympic Games, and the second was in Italy, where it was blatantly used—as were the Berlin Olympics of 1936—for Fascist propaganda. Guided by its manager Vittorio Pozzo, Italy won in 1934 and won again in France in 1938.

Brazil dominated post-World War II competitions, which were renewed in 1950. True, they lost to Uruguay in the 1950 final before 200,000 spectators in their own Maracana stadium in Rio. And in Switzerland in 1954, they were knocked out by Hungary in an unsavoury match in which three players were sent off and the fighting continued in the dressing rooms. Then they won three of the next four tournaments, 1958 in Sweden, 1962 in Chile, and 1970 in Mexico, to make the Jules Rimet Trophy their own. The adventurous freedom of Brazil's forwards was a refreshing contrast to the cautious, defensive football of the period.

Yet Brazil did not reach the heights of Hungary, which went from 1949 to 1955 with only one defeat, which came in the 1954 World Cup final against West Germany. Hungary enjoyed four world-class players and was the first foreign side to beat England on their own soil, by a score of 6–3 in 1953.

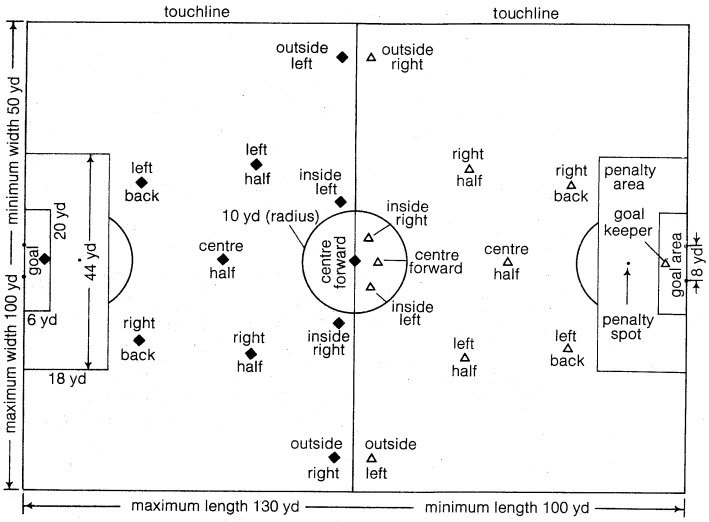


Figure 1: Soccer field and the positions of players at the kickoff.

The severity of the defeat was just punishment for England's stagnating insularity. They had an earlier warning when they entered the World Cup for the first time in 1950 and suffered a humiliating 1–0 defeat by the United States in a sensational giant-killing act. The reverse by Hungary caused a salutary revival, and England won the World Cup as host country in 1966.

The Olympic Games

The Olympic association football tournament was first staged in the 1908 games in London and has been in every Olympics since then, except at Los Angeles in 1932. England won the first two tournaments and then withdrew, mainly because of objections to broken-time payments (compensation for loss of regular salary) to amateurs. FIFA evaded the problem of defining an amateur and left it to the consciences, many of them easy-going, of the individual nations. Uruguay's title-winning sides of 1924 and 1928 were virtually professional.

The cynical attitude toward amateurism made the International Olympic Committee often look dubiously at the inclusion of soccer in the games, but the power of FIFA and the high attendances at matches made it difficult to eliminate. After World War II, the Iron Curtain countries dominated Olympic play because they claim they have no professionals, although realists dub their players "government amateurs." Hungary won in 1952, 1964, and 1968; the Soviet Union in 1956; and Yugoslavia in 1960. FIFA tried to redress the balance by banning players with World Cup experience from the Olympics.

The European Cup originated in 1955 for league champions of the 33 countries in membership with the European FA. The first five tournaments were won by Real Madrid, possibly the finest club side of all time. They were marshalled by Alfredo di Stefano, an Argentinian with phenomenal technique and stamina. The Latin countries dominated the cup until Celtic of Glasgow won in 1967. The success of the European Cup led to the launching of the Cup Winners' Cup for winners of the national cup competitions and the European Fairs Cup for 64 leading clubs. In 1960 South America started the equivalent of the European Cup competition, the Copa de Libertadores, and opened the way for an unofficial world club championship (beginning 1960), determined annually by a match between the winners of the European and the South American cups.

Playing the game. The aim is to propel the ball into the goal, using any part of the body except the hands and arms; the side scoring more goals wins. The ball is a round, leather-covered, inflated rubber bladder 27–28 inches (about 68–71 centimetres) in circumference and 14–16 ounces (435–497) grams) in weight. The players

Rules

move the ball by hitting it with head or foot from one teammate to another or by dribbling—a series of very short kicks, usually with the instep or side of the foot. Only the goalkeeper of the 11 players is allowed to handle it, and he is restricted to the penalty area, a rectangular area in front of the goal, 44 yards (40.2 metres) wide and extending 18 yards (16.5 metres) into the field. The game is of 90 minutes duration and is divided into equal halves. The teams change ends after a 5-minute half-time interval. For international matches, the pitch (playing field) must be 110–120 yards (approximately 100–110 metres) long and 70–80 yards (64–75 metres) wide. Markings on the field and the usual positions of players at the start of the game are shown in Figure 1.

UPI



World-famed soccer star Pelé (Brazil) has characteristically darted past the defender to intercept the ball and take a shot at the goal during the 1970 World Cup final against Italy at Mexico City. Brazil won 4–1 to become world champions for the third time.

Women and schoolboys usually play a shorter game on a smaller pitch. The game is controlled by a referee, who is also the time keeper, and two linesmen who patrol the touchlines, or sidelines, signalling when the ball goes out of play.

While solo runs and dribbles are spectacular, association football is essentially a team game, based on accurate passing of the ball among members of the same side. Passing produces the fluid movements and varied patterns that make the game an attractive spectacle. A distinctive feature is the development of heading, and many thrilling goals are scored by a forward jumping to hit the ball with his forehead.

Free kicks are awarded for fouls or violations of rules; all players of the offending side must be ten yards from the ball. Free kicks may be either direct, from which a goal may be scored, for more serious fouls such as kicking an opponent, tripping, or handling the ball; or indirect, from which goals cannot be scored until after the ball has touched another player, awarded for lesser violations such as obstruction (interfering with an opponent while not playing the ball). A penalty kick, a direct free kick awarded to the attacking side in the penalty area, is taken from a spot 12 yards from the centre of the goal, with all players other than the defending goalkeeper and the kicker outside the penalty area.

Considering the tremendous advances in play in the 20th century, it is remarkable that there have been only two major alterations of the rules in the period—those related to "offside" and to player substitution. A man is offside if he receives a forward pass from a colleague in the opponent's half with fewer than two opponents—three before the rule was changed—between himself and the goal line. The interpretation often causes controversy, even though the linesmen generally guide the referee by positioning themselves alongside the foremost attacker on each team.

The flagrant use of substitutes in many countries forced the FIFA's International Board to alter the rule limiting teams to 11 players. At first it was hoped to restrict changes to injured players, but it is now permissible to replace any two at any time. Substitutes were used for the first time in the World Cup in 1970—two from five on the reserve bench—and it proved an invaluable innovation because of the heat and altitude of Mexico City, site of the contest.

The biggest changes have been in uniforms and equipment. Clothing has become attenuated, and heavy boots have given way to shoes. Many players have discarded shin guards. Conformity of dress is not essential, provided that the sides can be identified and the goalkeeper is distinguishable. The rules also permit players to participate barefoot.

Styles of play

Although professionalism and the league system brought benefits, they are partly responsible for the growing blight of defensive football. Herbert Chapman, shrewd manager of the Arsenal team in the 1930s, pulled back the centre halfback to be a third back and operated with four forwards instead of five. Italy, possibly the most commercialized soccer country, introduced the Catenaccio system (see Figure 2), with an extra man, named the sweeper or libero (free man), behind another four defenders, plugging holes where they developed.

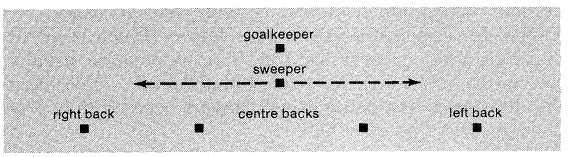


Figure 2: Catenaccio defense, showing the positions of players and the range of the sweeper.

Ideally, with both tactics the solid defenses were the springboard for quick counter attacks and, in Scotland's Alex James, Arsenal had the man to organize them. But with most imitators it was solely a case of stopping the other side from scoring.

A defensive outlook was also encouraged by the European club tournaments, which were played on a home and away basis, the winning side being that with the greater aggregate of goals. As a result, visiting teams did not go for a win but concentrated on keeping down the opponent's score in the hope of obtaining more goals in the return match. A bold attempt to reverse this trend was made by the North American Soccer League in 1969, awarding six points for a win, three for a draw, and a bonus point for each goal up to three. A side losing heavily still had an incentive to go for goals.

One of the few attacking ideas to come forward after World War II was the withdrawn centre forward (see Figure 3) of Hungary's Nandor Hidegkuti in the 1950s.

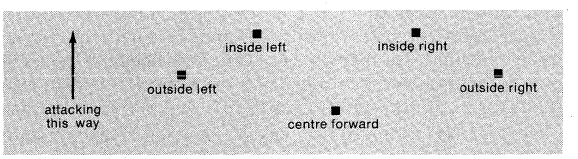


Figure 3: The positions of players for the withdrawn centre forward attack.

He played behind the attack, luring the centre halfback with him, and the inside forwards were the spearhead. The line had an M shape instead of the fashionable W formation.

Teams used to have specialists who stuck to their roles—wingers out on the touchlines, backs deep in defense, and centre forwards lying upfield. They gradually gave way to all-around players who realized their responsibilities in all departments of the game. Work rate became as important as skill, and method football took over. Brazil grouped its outfield players in a 4–2–4 formation when winning the 1958 World Cup and still gave outlet to its brilliant individualists. England won the World Cup in 1966 playing 4–3–3—without a winger—and four years later often operated 4–4–2, fielding only two strikers up front.

Too much can be made of numbers and systems. Basically, the aim is to put a man with time and room to spare in a shooting position near the opponent's goal, and it does not matter whether the man is nominally a forward or a defender. Attitude of mind is as important as technical excellence in achieving the aim.

BIBLIOGRAPHY. A.H. FABIAN and G. GREEN (eds.), *Association Football*, 4 vol. (1960), the most comprehensive book on the game, covering history, laws, administration, amateur and schoolboy football, British and international football, and all major competitions; B. GLANVILLE, *Soccer: A Panorama* (1968), a dramatic presentation of the development of the game in Britain and its growth abroad; G. GREEN, *History of Association Football* (1960), the most authoritative account of the development of the association that refined the game into its modern form and set international standards; R.C. CHURCHILL, *English League Football* (1961), a history of the Football League, all its clubs, and league tables from 1888; B. GLANVILLE and J. WEINSTEIN, *World Cup* (1958), a history of the first five tournaments, 1930–54; B. JOY, *Soccer Tactics* (1962), diagrams and photographs illustrating how ballcraft and intelligent positioning are harnessed into teamwork, including the development of tactics from the mid-19th century to current styles used by leading foreign teams; A. CSANÁDI, *Soccer*, 2 vol. (Eng. trans. 1967), an analysis of the game, covering individual skills and tactics and coaching, with diagrams and action photographs.

(B.Jo.)

Asterales

The order Asterales consists of a single very large family of flowering plants, the Asteraceae, also called Compositae. The Asteraceae family is one of the largest families of flowering plants, perhaps the largest. The number of species is not accurately known; but estimates of 15,000 to 20,000 are current. The only other family with a comparable number of species is the orchid family, Orchidaceae. Asterales is the name-bringing order for the subclass Asteridae in the class Magnoliopsida (dicotyledons) of the division Magnoliophyta (angiosperms, or flowering plants). The order includes many familiar garden ornamentals, such as asters, chrysanthemums, dahlias, daisies, marigolds, sunflowers, and zinnias. Some other members of the group, such as dandelions, ragweeds, and thistles, are familiar weeds. Lettuce and safflower are examples of economically important members of the Asterales order.

GENERAL FEATURES

Unique flowers a distinguishing characteristic

The most obvious and outstanding general feature of the Asterales is that the flowers are characteristically grouped into compact inflorescences (heads) that superficially resemble individual flowers. The name Compositae refers to this feature; what appear to be individual flowers are actually composite flowers. Each such head is ordinarily subtended by a usually green involucre of small modified leaves (bracts), which bear the same relation to the flower head that the green sepals do to an individual flower in other families. Furthermore, in more than half the members of the order, the flowers in the outermost row or rows of the head have a modified, mainly flat and elongate corolla (the collection of petals) that more or less resembles an individual petal of an ordinary flower. The "petals" of a daisy or sunflower are actually these outermost flowers of the head. Various members of the order may produce anywhere from one to several hundred or even more of these composite-flowered heads.

Several other families or individual genera of flowering plants have the flowers grouped in heads more or less resembling those of the Asterales, but they lack the syndrome of other features that characterize the order, and most of them especially lack the complex pollen-presentation mechanism (see below *The flower*).

Structural diversity and distribution. Members of the Asterales order are diverse in habit and habitat. They are most characteristically herbs of sunny places in temperate to subtropical regions, but they are not at all so restricted. Not many of them are large trees, and not many of them are adapted to life in undisturbed moist tropical forests, but otherwise they are ubiquitous. They are trees, shrubs, herbs, or woody or herbaceous vines. The herbs are annual, biennial, or perennial, with or without an active cambium (a zone of growth contributing to stem thickness). Some of them are megaphytes (coarse, simple, or sparsely branched shrubs with thick, soft, almost herbaceous stems). Many of the perennial herbs have long, creeping rhizomes (underground rootlike stems) from which large colonies (clones) can develop by vegetative multiplication; others are taprooted, or fibrous-rooted from a caudex (stem base) or short rhizome. The leaves are simple or less often compound, and their arrangement along the stem may be opposite, alternate, or less commonly whorled; not infrequently they are opposite toward the base of the stem and alternate above. Species occur from the Arctic to the Antarctic and from above the timberline to the ocean shores. In addition to the more ordinary habitats, some of them are adapted to growth in sand dunes; others to cliff crevices; others to talus slopes; others to seleniferous, gypsiferous, or alkaline soils; and others to fields or disturbed sites around human habitations. A few species are aquatic.

Centres of diversity

The greatest centres of diversity in the order are the dry highlands of Mexico, where the primitive tribe Heliantheae is especially well represented, and the Mediterranean–Near East region. South Africa is an important secondary centre of diversity. In most temperate regions more than 10 percent of the species of angiosperms belong to the Asterales. In tropical regions the percentage is smaller, but the numbers are still significant.

Economic importance. The greatest economic importance of the Asterales order lies in the use of many of its members as garden ornamentals. Species and garden hybrids of *Aster*, *Bellis* (English daisy), *Callistephus* (China aster), *Chrysanthemum*, *Cosmos*, *Dahlia*, *Helianthus* (sunflower), *Rudbeckia* (coneflower, black-eyed Susan), *Tagetes* (marigold), and *Zinnia* are well-known garden favourites. *Achillea* (yarrow), *Ageratum*, *Anaphalis* (pearly everlasting), *Anthemis*, *Artemisia*, *Calendula*, *Centaurea*, *Echinops* (globe thistle), *Erigeron* (daisy), *Eupatorium*, *Gaillardia* (blanketflower), *Helichrysum* (strawflower, everlasting), *Liatris*, *Ratibida* (coneflower), *Santolina*, and *Stokesia* are also familiar in gardens. The florists' cineraria, a popular wintertime pot flower, is *Senecio cruentus*, originally from the Canary Islands.

The most important food plant in the Asterales is lettuce, *Lactuca sativa*, a European cultigen, followed by the common sunflower, *Helianthus annuus*, a native of the United States. Sunflower seeds are excellent poultry food, and a light-golden-yellow oil made from them is used as a salad oil and in cooking and the manufacture of margarine, soap, paint, and varnish. The oil cake is fed to livestock, and the whole plant is used as ensilage. Flowers of safflower, *Carthamus tinctorius*, are the source of a red and a yellow dye, and the seeds produce an edible oil that is also used in soap, paint, and varnish. Several other members of the order, such as the artichoke (*Cynara scolymus*) and the Jerusalem artichoke (*Helianthus tuberosus*), are of minor importance as food plants.

Chemical compounds derived from Asterales

Pyrethrum, an insecticide that does not produce the environmental problems associated with DDT and other synthetic products, is obtained from the flowers of several species of *Chrysanthemum*, notably *C. cinerariaefolium*. Extracts from several species of *Artemisia*, notably *A. cina* from the Middle East, have been much used to expel intestinal worms, whence the common name wormwood applied to this genus. *Artemisia absinthium* is the source of a poisonous oil used to give the liqueur absinthe its distinctive character.

A few other members of the order have minor economic uses, and several of them have excited interest as a possible source of rubber. Guayule (*Parthenium argentatum*) and the Russian dandelion (*Taraxacum koksaghyz*) are two species that have been studied in this regard, but neither is economically profitable when other sources of rubber are available.

The ragweeds (*Ambrosia*), dandelions (*Taraxacum*), and thistles (*Carduus, Cirsium,* and *Onopordum*) are the most troublesome weeds in the Asterales order. Among the wind-pollinated weeds, *Ambrosia artemisiifolia* (common ragweed) and *Ambrosia trifida* (giant ragweed) are two of the most important plant species causing the allergic reaction known as hay fever.

NATURAL HISTORY

Pollination. Pollination is effected by diverse agents, most commonly various sorts of insects. The individual flowers of most species are relatively small, and the nectar within the corolla tube is thus readily available to most insect visitors. The pollen itself is freely exposed on the surface of the head, and such heads are likely to be visited by diverse kinds of insects. A considerable minority of members of the order are wind pollinated; these generally have small and inconspicuous flower heads.

Some species are pollinated by both wind and insects. *Solidago speciosa,* one of the common goldenrods of eastern United States, for example, produces a considerable amount of airborne pollen in addition to attracting insect visitors. The goldenrods, like the ragweeds, generally flower in late summer and fall; and because they are common and conspicuous when the ragweeds are pollinating, they have often been blamed for the allergies that are actually caused primarily by the ragweed.

A relatively few species are regularly self-pollinated; the genus *Psilocarphus* provides an example of this method. Bird pollination is also uncommon, but the tropical American genus *Mutisia* is bird pollinated.

Production of fertile seeds asexually

Various genera and individual species of the order are known to be reproduced by apomixis, the setting of seed without fertilization, either completely or in addition to normal sexual means. The genus *Antennaria* (pussytoes), well-known in the Northern Hemisphere, is dioecious (male and female heads on separate plants), and some of the species are represented in large parts of their range only by pistillate (female) plants. In this genus, normal sexual reproduction produces equal numbers of staminate and pistillate plants, but apomictic reproduction yields only pistillate plants. Among the members of the Asterales order, as in other orders, apomixis is often associated with polyploidy (the presence of three or more complete sets of chromosomes in every cell) and a past history of hybridization.

Seed dispersal. The fruit of the Asterales is an achene; *i.e.*, it is dry, contains only one seed, and does not open at maturity. The apparent seeds of the sunflower, for example, are actually achenes. The hull is the achenial wall, and the proper seed coat surrounding the embryo is thin and insignificant. In speaking of seed dispersal of the Asterales, it should be realized that it is actually the achenes, each containing a seed, that are dispersed.

The seeds of many Asterales species are distributed by wind, having fluffy or parachute-like structures that provide buoyancy. In others, such as *Coreopsis* (tickseed), the achene is thin and flat, and the surface area is increased by the presence of an even thinner expanded margin (wing). Some have barbed structures or are provided with hooks or spines, as in cocklebur (*Xanthium strumarium*) or burdock (*Arctium* species), that engage man or animals as means of transport. Other means of seed dispersal are less common. The achenes of species that grow in wet places may be carried in mud on the feet of migrating waterfowl; those of some streamside species have minute air cavities in the wall and are buoyant, achieving dispersal by floating until they become waterlogged. In *Centaurea* and some related genera, the achenes are attractive to ants, which carry them about and feed upon special parts of the wall. The achenes of some field weeds have been widely distributed by becoming mixed with the seeds of cultivated crops. Many other members of the order have no obvious means of seed dispersal.

FORM AND FUNCTION

The flower. The flowers of the Asterales are sympetalous; *i.e.*, the petals are joined together by their margins, forming a tubular or mainly strap-shaped corolla that often has apical teeth representing the petal tips. The flowers are also epigynous; *i.e.*, the other floral parts are attached to the top of the ovary rather than beneath it. These two features occur individually in various other groups of flowering plants, but their occurrence in combination is much more limited.

The calyx (collection of sepals) of the Asterales is so highly modified, in contrast to that of other orders, that it is given a different name, the pappus. The pappus consists of one to usually several or many dry scales, awns (small pointed processes), or capillary (hairlike) bristles; in some, the scales may be joined by their margins to form a crownlike ring at the summit of the ovary. In only a few genera (*e.g.*, *Marshallia*) of the primitive tribe Heliantheae does the calyx consist of five regularly placed scales that are obviously homologous with sepals. Often the pappus is completely wanting. When the pappus consists of numerous capillary bristles, as in the common dandelion (*Taraxacum officinale*), it facilitates wind distribution of the achenes. In some other genera, such as *Bidens* (beggar ticks), the pappus awns are barbed, permitting them to stick in fur or clothing, and some achenes are thus transported by animals. In many other genera, especially those with the pappus of scales or a crown, the function of the pappus is obscure.

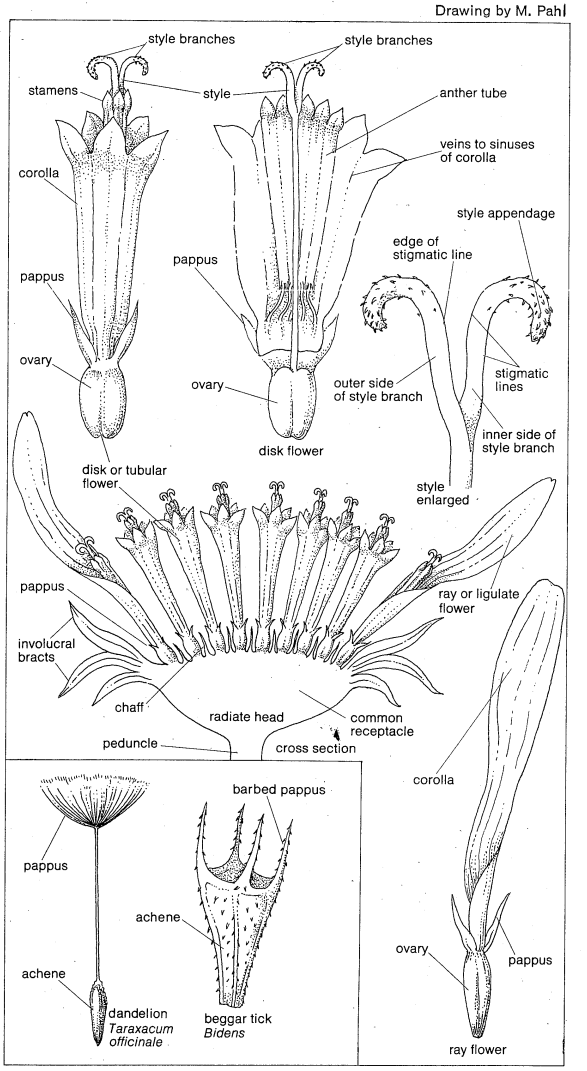


Figure 1: Floral structures of the Asterales.

Pollen-presentation mechanism

The pollen-presentation mechanism of the Asterales is especially characteristic, being shared only by the orders Campanulales and Calycerales and a few genera of the family Rubiaceae, in the order Gentianales. The stamens (male reproductive structures) are generally joined to-

gether by their anthers (pollen producing region of stamen) to form a tube; the stamens open toward the centre by lengthwise slits, releasing the pollen into the tube. The style (part of the female structure) grows up through the tube, pushing the pollen out ahead of it. The style usually has two branches that separate after it has grown up through the anther tube. The stigmatic surfaces (pollen receiving regions) are usually arranged in lines along the inner margins of these branches, well back from the sterile tips (appendages), which are the structures that push out the pollen. The stigmatic lines later become exposed to the air, when the style branches spread apart.

The pistil (female structure) is composed of two carpels, united to form a compound ovary with a terminal style. There is usually a nectar-producing region (nectary) in the form of a minute ring surrounding the style atop the ovary. The ovary has only one locule (seed cavity), with a single ovule arising from the base. The fact that the ovule is basal provides the best single distinction of the Asterales from the related order Calycerales, which also has involucrate heads with a similar pollen-presentation mechanism but which has the ovule pendulous from the top of the ovary.

As in most flowering plants, the ovule is anatropous; *i.e.*, it is curved back on itself so that the micropyle (the apical opening) is alongside the funiculus (the stalk). As in other members of the subclass Asteridae, the ovule has a single, rather thick integument (outer covering layer, the forerunner of the seed coat) rather than the double integument found in so many other orders. The seed has no endosperm; its reserve food is stored largely in the two cotyledons (seed leaves) of the embryo.

The inflorescence. The secondary inflorescence of the Asterales (*i.e.*, the arrangement of the flower heads) is typically cymose (determinate). The terminal head blooms first, followed by the terminal heads of the main branches. After that the sequence is less clear, not infrequently being mixed, with both cymose and racemose (indeterminate) components. Only rarely, and then clearly as a derived condition, is the secondary inflorescence racemose throughout, with the lowest heads blooming first and the terminal ones last.

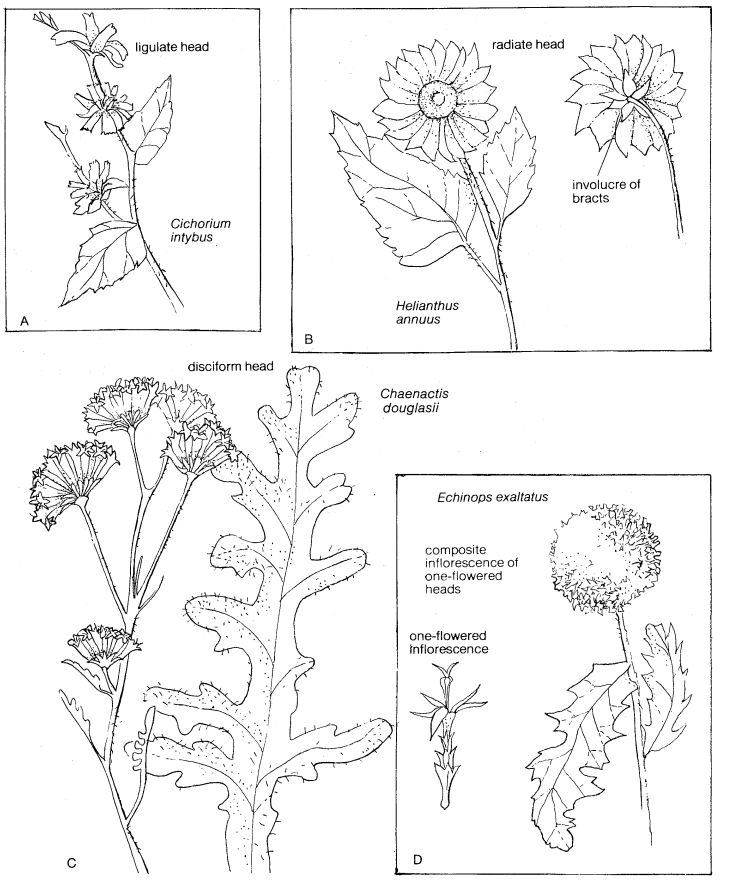


Figure 2: Inflorescence types in the order Asterales.
Drawing by R. Findahl based on (A,B, radiate head, D) G.H.M. Lawrence, *Taxonomy of Flowering Plants* (1951), The Macmillan Company; and (B, involucre of bracts) J. Craighead, F. Craighead Jr., and R. Davis, *A Field Guide to Rocky Mountain Wildflowers*, reprinted by permission of the publisher Houghton Mifflin Company

Sequence of flowering

The sequence of flowering within the individual heads, on the other hand, is always racemose (indeterminate, centripetal). The outer flowers bloom first, and there is a progressive spiral of flowering thence to the centre of the head. The head is in essence a compact shoot, with spiral (alternate-leaved) phyllotaxy. The involucral bracts are more or less modified leaves, and very often they are green and leafy in texture. Sometimes the flowers in the outermost row of the head are borne in the axils (angles) of involucral bracts, sometimes not.

In many members of the order, especially the relatively primitive tribe Heliantheae, the common receptacle (the short stem tip, on which all the flowers of the head are borne) is also provided with bracts, to which the individual flowers are axillary. The receptacle is said to be chaffy when these bracts are present and naked (or sometimes bristly) when they are not. Sometimes the receptacle is chaffy only toward the margin or only toward the centre. In some genera (notably of the tribe Cynareae) the receptacle is bristly; *i.e.*, it is beset with bristles that are difficult to interpret in terms of bracts, although that may be their eventual evolutionary origin.

Flowering head types. Individual heads of most members of the Asterales order are said to be ligulate, radiate, discoid, or disciform, according to the kinds of flowers they contain. The simplest type is the discoid head, with all the flowers having a regular, tubular corolla, the generally four or five apical teeth representing the tips of the petals. This kind of flower is called a disk flower. Ordinarily the flowers in a discoid head are all perfect (with reproductive parts of both sexes) and fertile. In a comparison of the flower head to the individual flowers of other kinds of plants, the discoid head lacks the marginal, petal-like flowers. Thistles and ageratums are examples of Asterales species with discoid heads.

The radiate head has disk flowers in the centre, surrounded by one or more marginal rows of another kind of flower, the ray flower. The corolla of ray flowers is very irregular. It is tubular at the base but prolonged on the outer side into a generally flat projection, the ray, or ligule. These rays are the petallike parts, in a comparison of the flower head to an ordinary flower. The ray in radiate heads represents only three lobes of the corolla; often there are two or three minute apical teeth. The other two corolla teeth (those toward the centre of the head) are much reduced or, often, wanting.

The details of size and structure of ray flowers in a radiate head vary, but the position of the ray flowers with respect to the disk flowers is absolutely constant. The disk flowers are always central and the ray flowers marginal. Ray flowers in radiate heads are never perfect (bisexual). Instead they are pistillate (with female reproductive parts only) or neutral (with a vestigial, nonfunctional ovary and no style). Disk flowers in a radiate head are usually perfect, but sometimes they are functionally staminate, with a normal pollen presentation mechanism but without a viable ovary.

"Double-flowered" plants

Occasional mutant forms of species that normally have radiate heads have most or all of the disk flowers more or less transformed into ray flowers, with only a few (or no) normal disk flowers in the centre. These "double-flowered" forms do not survive long in nature, but they are valued and perpetuated horticulturally because of their more showy flower heads. The "daisy-flowered" chrysanthemums, with only a single marginal row of ray flowers, have the normal type of radiate head, but the more commonly cultivated kinds of mums are double flowered. The garden dahlia is another member of the Asterales order that is cultivated in both normal and double-flowered types, with the double-flowered ones being more frequent. China asters, marigolds, and zinnias are also commonly cultivated in double-flowered forms.

The disciform head is a special derivative of the radiate type. It resembles the discoid head in lacking the mar-

ginal rays, but the outer flowers are pistillate, with a tubular, rayless corolla. The corollas of these marginal pistillate flowers are generally very slender, in comparison to the normal disk corollas, and they do not have the four or five well-defined teeth of the disk corollas. *Artemisia stelleriana*, one of several plants called dusty miller, has disciform heads, as do plants of the genus *Gnaphalium* (cudweed). Some species, such as *Erigeron compositus*, show a complete series of transitions from the radiate to the disciform type of head, the ray part of the marginal corollas varying from elongate to short to very short to lacking.

Radiate, discoid, and disciform heads occur in various tribes of the Asteraceae family. The ligulate head, on the contrary, is almost entirely restricted to one tribe, the Lactuceae (Cichorieae), and is found in all members of that tribe. Ligulate heads consist entirely of one kind of flower, the ligulate flower. Ligulate flowers superficially resemble the ray flowers of radiate heads in having the corolla tubular at the base and prolonged on the outer side into a flat, strap-shaped ligule. They differ from ray flowers in being ordinarily perfect (with functional stamens as well as a pistil) and especially in the structure of the ligule itself, which consists of all five lobes of the corolla and generally shows five terminal teeth. The dandelion is a familiar plant with ligulate heads. The term ligulate flower is often loosely used to cover both ray flowers and the type here described as ligulate. It is helpful, however, to have separate terms for these two kinds of ligule-bearing flowers, and no other terms for them are in general botanical use.

Still another kind of flower is found nearly throughout another tribe, the Mutisieae. This tribe is largely tropical, and only one of its genera, *Gerbera*, is familiar in cultivation in temperate regions. Most members of the Mutisieae have some or all of the corollas bilabiate (two-lipped), with a large, three-lobed (sometimes four-lobed) outer lip and a smaller, two-lobed (or one-lobed) inner lip. These bilabiate flowers may be either pistillate or perfect. When pistillate, they are always external to any perfect flowers that may be present in the head. Often they are much like ordinary ray flowers, except that there are two small teeth at the top of the corolla tube, opposite the ligule. The Mutisieae tribe shows every transition from the typical disk flower to the typical ray flower and the typical ligulate flower.

Disk corollas in the Asterales generally have an unusual sort of vascular system, scarcely found outside the order. There are as many longitudinal veins as corolla lobes, typically five, but the veins are directed to the sinuses of the corolla rather than to the teeth. At the sinus each vein forks, one fork following along the margin of each corolla tooth and meeting the vein from the other side of the tooth at the tip. Ray flowers and ligulate flowers are similar, but some of the larger rays have more numerous veins, and many of the smaller ones have reduced or no veins.

Numbers of flowers. A curious feature about the flower heads of the Asterales is the frequency with which the number of ray flowers (and sometimes also involucral bracts) reflects the Fibonacci series, a mathematical series that proceeds 1, 1, 2, 3, 5, 8, 13, 21, 34, . . ., with each succeeding number being the sum of the two previous ones. Fibonacci numbers also appear in efforts to explain the arrangement of leaves on a stem (phyllotaxy) in angiosperms in general. The appearance of Fibonacci numbers as absolute numbers in heads of the Asterales is thought to reflect the fact that the flower head is a compact shoot. Given the existence of normal phyllotaxy on this shoot, it is easier to produce a symmetrical arrangement with 5, 8, 13, or 21 rays than with other numbers such as 7, 9, 11, 19, etc. Heads with 13 rays, for example, probably have a 5/13 phyllotaxy (5 turns around the stem to pass through 13 leaf bases, before the next leaf is directly above the first one). Given the same phyllotaxy, a head with 14 rays would have the 14th ray on top of the first one, and one with only 12 rays would have a gap in the series.

Fibonacci numbers in composite flower heads

These Fibonacci numbers, when they appear, often represent tendencies rather than fixed numbers, especially when the number is 13 or more. Five-rayed or 8-rayed heads may be very constant, with few or no examples of other numbers; but it is not uncommon to see within one local population (or on one plant) a range of 10 to 16, with 13 as the high point of the curve of distribution. Sometimes there is even a bimodal curve, with high points on adjacent Fibonacci numbers, such as 13 and 21, or 21 and 34.

EVOLUTION

Fossil record. The fossil record of the Asterales is not especially useful in helping to understand the evolutionary history of the group. The floral structures that would permit identification of the order on critical technical features are rarely fossilized, and with few exceptions fossils can be referred to the group only by a close resemblance to modern species or genera. A fossil (*Palaeanthus problematicus*) from the Upper Cretaceous Period (about 75,000,000 years ago) in New Jersey is more or less suggestive of a sunflower head, but it could equally well be any of several other things, and its status as a member of the Asterales order is dubious at best. The oldest generally accepted fossil representatives of the order are some dandelion-like achenes of Oligocene age (about 30,000,000 years ago). There is also a recently discovered fossil from the Oligocene–Miocene boundary region (26,000,000 years ago) in Montana that looks very much like a head of the modern genus *Viguiera* (tribe Heliantheae) and has been so interpreted in the literature. The order is poorly represented in Miocene (about 13,000,000 years ago) and even Pliocene (about 5,000,000 years ago) deposits, not becoming abundant until the Pleistocene (beginning about 2,500,000 years ago), but this situation doubtless reflects the difficulty of identification rather than a real absence of the group.

Oldest fossils of aster order

Phylogeny. Phylogenetic speculations in the absence of a useful fossil record are hazardous, but there is a fair degree of agreement among students of the Asterales (synantherologists) about the evolutionary trends within the order. The interpretation necessarily depends mainly on comparison of living members of the group, in the light of generally accepted principles of angiosperm evolution. Present evidence permits taxonomists to believe that the currently accepted broadscale organization and evolutionary concepts concerning the Asterales order are probably correct, but some major uncertainties remain. In the following discussion, as elsewhere in this article, the terms primitive and advanced refer to phyletically anterior, or primary (*i.e.*, early appearance in a group), as opposed to subsequent, or secondary, status and imply nothing at all about the adaptive significance of the features in question.

It is generally believed that the ancestral prototype of the Asterales was woody, perhaps a small tree. Some of the woody members of the group (*e.g.*, the sagebrush, *Artemisia tridentata*) are clearly only secondarily woody, deriving from herbaceous ancestors within the order, but many more or less woody Mexican species of the Heliantheae tribe are considered to be primitively woody. Many of the temperate zone herbs in the order, such as the common sunflower, develop a considerable amount of wood, but the stem does not persist from one year to the next.

It is believed that opposite leaves are primitive in the Asterales, and alternate or whorled leaves derived. Numerous transitional forms from opposite to alternate leaves occur in the Heliantheae, a tribe considered on a syndrome of features to be relatively primitive within the order. Many of the Heliantheae have the lower leaves opposite and the upper ones alternate. It is thought that a progressive phyletic change from opposite to alternate phyllotaxy began in the flower heads and spread progressively farther down the stem until all the leaves were alternate.

The presence of a well-developed resin system (a system of ducts or intercellular spaces lined with specialized cells that secrete resin, a sticky organic substance insoluble in water) in the vegetative parts is considered to be a primitive feature in the Asterales. Reduction and

Resin and latex systems

loss of the resin system proceed from the top downward. Presence of a latex system, on the other hand, is secondary, and its phyletic development proceeds from the top downward. Elaboration of the latex system is usually associated with reduction of the resin system, but the two systems are anatomically and genetically independent. A few genera and species have both the latex and the resin system.

Since each head represents a complete inflorescence, it is assumed that primitive members of the order had relatively few heads, each with relatively numerous flowers. The cymose secondary inflorescence (the arrangement among the heads) is also primitive.

Synantherologists are agreed that aggregation and reduction in the inflorescence have been pervasive trends within the order. The culmination of these trends is the production of composite composites, such as in *Echinops* (tribe Cynareae) and *Lagascea* (tribe Heliantheae), which have numerous one-flowered, individually involucrate heads aggregated into a secondary head with its own secondary involucre.

Another assumption based on the premise that the flower head is a complete inflorescence is that the involucre primitively consisted of several series of green, more or less leaflike bracts. Involucres with dry, nongreen bracts, or with the bracts in only a single series, are thus considered as derived rather than primitive. The situation is not perfectly simple and straightforward, however. It is not always easy to distinguish the outer involucral bracts from the uppermost leaves of the peduncle (flower stalk), and in diverse species in various tribes the involucre has evidently been augmented by modification of these uppermost peduncular leaves into bractlike structures.

The presence of chaff on the receptacle is uniformly considered to be a primitive character in the order and its absence to reflect a loss. Only one tribe, Heliantheae, has more representatives with the receptacle chaffy than with it naked.

The discoid head, with the flowers all tubular, has traditionally been considered the primitive type within the family, but this concept is now held to be too simplistic. Scattered throughout the several usually radiate headed tribes (Heliantheae, Astereae, Anthemideae, Arctotideae, Inuleae, Senecioneae, and Calenduleae) are special subtribes, genera, species, infraspecific taxa (varieties, races, etc., within a species), and individuals with discoid heads. Among the radiate tribes the absence of rays represents a loss, and the discoid head is phyletically secondary. The consistently discoid tribes Cynareae, Eupatorieae, and Vernonieae are more difficult to explain as secondarily discoid, but neither is there any reason to suppose that they are primitive in other respects.

Evolutionary appearance of the ray flower

There is no doubt that the ray flower is, phyletically, a modification of the disk flower, but its time of appearance during the evolutionary development of the order is not known. If, as is thought by some, it occurred in pre-Asteralean ancestors and was transmitted to the earliest members of the Asterales order as a normal feature of the group, then the discoid tribes have lost their rays, and all discoidy in the order is secondary. On the other hand, if the change took place in the ancestors of the Heliantheae tribe, after this group had already diverged from the ancestors of the present discoid tribes, then discoidy is primary in the discoid tribes and secondary only in the radiate tribes.

The Mustisieae tribe is a special case. It shows all transitional stages in the development of ray flowers from disk flowers, and the changes evidently occurred within the tribe rather than in remote pre-Asteralean ancestors. The Mutisieae are advanced in other features, however, notably the caudate ("tailed") anthers, and the radiate tribes could not have evolved from them. Evolution of rays in the Mutisieae shows how it may have happened elsewhere in the order, but it is not the direct source of the rays in other tribes.

Synantherologists consider fertile disk flowers to be more primitive than functionally staminate ones. Similarly, pistillate, fertile ray flowers are considered to be more primitive than ray flowers that are sterile but that still have a style; the neutral ray, lacking even a style, is still further advanced.

Yellow floral pigments have been considered to be primitive in the Asterales, and anthocyanic pigments, which produce blue to red colourings in flowers and other plant parts, to be derived. This hypothesis has been extended to suggest that the sequence is from carotenoids (yellow to red pigments) to anthoxanthins (yellow to scarlet flavonoid pigments, visually similar to the carotenoids) to anthocyanins (blue to purple or crimson flavonoid pigments). Thus the major chemical change, from carotenoids to flavonoids, occurs without much influence on the actual flower colour, and the major change in flower colour reflects a very minor chemical change.

The restriction of the stigmatic surfaces to parts of the style well back from the tip is part of the pollen presentation mechanism syndrome. The condition of the Heliantheae tribe, in which the stigmatic lines are often not well differentiated from the nonstigmatic part of the style branches, is considered to be primitive in this regard. Both ray flowers and ligulate flowers generally have well-defined stigmatic lines extending all the way to the tip of the style branches, which are without a sterile appendage. This condition is thought by some to argue for the evolution of both of these floral types very early in the evolutionary history of the order, before the style tips were sterilized.

The Asterales, like the angiosperms in general, are so beset with parallel evolution (*i.e.*, the independent appearance of similar structures in related groups) that it is often difficult to distinguish independently evolved similarities from those inherited from a common ancestor. There is, in fact, no sharp line between these two conceptually different types of similarity, and close parallelism among phyletic lines from closely related ancestors is very common indeed. Before all the members of an admittedly natural group can be traced back to a common ancestry, it is likely that the path will lead outside the confines of the group. These considerations apply with especial force to the genera and tribes of the Asterales. A genus that combined all the primitive features found in various genera of the tribe Astereae would probably have to be referred to the tribe Heliantheae. The Astereae, like the other tribes, probably evolved as a set of closely related genera undergoing similar sorts of evolutionary changes, rather than by diversification from a single genus that had all the characteristic features of the tribe.

For the reasons noted above, there may never have been an original member of the Asterales, primitive in all respects within the order but possessing all the essential features of the group. Nevertheless, it is conceptually useful to consider what such a plant might have been like and from what other group it might have originated.

Description of proposed Asterales ancestor

The hypothetical ancestral prototype of the Asterales, subject to the limitations noted in the preceding paragraph, may be considered to have been a small to medium-sized tree, probably in or near the dry highlands of Mexico. It had a well-developed resin system and opposite, simple, exstipulate leaves with anomocytic stomates (lacking specialized supporting cells around the guard cells). It had relatively few heads, terminal to the branches and more or less cymosely arranged. Each head had more or less numerous yellow flowers, collectively subtended by an involucre of several series of herbaceous bracts. The head may have been radiate or less probably discoid. The rays, if it had them, were pistillate and fertile. The receptacle was chaffy, with each bract subtending a flower. The flowers were arranged in a close spiral, blooming from the outside of the spiral toward the centre. The flowers were epigynous (*i.e.*, the floral parts attached to the top of the ovary), and the disk flowers were perfect. The pappus consisted of five scales. The disk-corollas were regular and of five lobes, with the midveins to the teeth and lateral veins to the sinuses well developed. The anthers were joined into a tube, into which the pollen was released; they did not have a sterile tail. The style grew up through the anther tube, pushing out the pollen, and there was no sharp differentiation of the

style branches into stigmatic lines and sterile appendages. The ovary was of two carpels, with a single basal ovule.

No presently existing genus combines all of these features, but more of them are found together in the Heliantheae than in the other tribes, and if such a plant did exist it would probably be referred to that tribe. The Heliantheae tribe is thus generally considered to be the most primitive tribe in the order. The Astereae, Anthemideae, Arctotideae, Inuleae, and Senecioneae tribes are generally considered to be derived directly from the Heliantheae, and the Calenduleae tribe is thought to be a specialized derivative of Senecioneae. These several tribes collectively constitute the radiate tribes.

The phyletic position of the discoid tribes (Cynareae, Eupatorieae, and Vernonieae) is less certain, but the Arctotideae do provide a sort of link to the Cynareae, suggesting that this latter tribe, at least, may also be derived eventually from the Heliantheae tribe. The Mustisieae tribe is thought to be allied to the Cynareae, and it also provides a tenuous link to the Lactuceae, the most isolated and distinctive tribe of the order.

Looking backward for possible ancestors to the Asterales, it is found that the families Rubiaceae (order Gentianales) and Caprifoliaceae (order Dipsacales) come the nearest to providing the required features. Many authorities have thought that these two families are themselves closely related. Some of the other members of the Dipsacales order, such as the family Dipsacaceae, have often been compared to the Asterales; but they are already too advanced to serve as possible ancestors, being herbaceous and having a terminal rather than a basal ovule.

In the past, the Asteraceae family was often included in the order Campanulales, but that order is basically herbaceous and alternate leaved and thus can scarcely be ancestral to the Asterales. The small order Calycerales, which has involucrate heads much like those of the Asterales, is also excluded by being herbaceous and alternate leaved. The Calycerales order, furthermore, has a terminal ovule. The Campanulales and Calycerales do have a pollen-presentation mechanism similar to that of the Asterales, and this is probably the main reason that the Asteraceae family has often been associated with the order Campanulales; but a similar mechanism is also present in some members of the Rubiaceae family.

Functional significance of evolutionary adaptations

A comprehensive interpretation of the evolutionary origin and diversification of the Asterales order in terms of adaptation and survival value is difficult and has never been seriously attempted. The flower head obviously acts as a single large flower, with the ray flowers serving the function of petals to make the flowers conspicuous and attractive to a wide range of insects. The pollen-presentation mechanism, which makes the pollen available to all kinds of pollinators, fits well into such a scheme. The pappus in many genera is clearly adapted to seed dispersal. On the other hand, the repeated reduction and loss of both the ray flowers and the pappus are more difficult to account for, although presumably there could be a series of different explanations for the different examples. The functional significance of both the resin system and the latex system is obscure. There is no obvious value in the change from opposite to alternate leaves, nor in the frequent reduction and aggregation of the flower heads to form clusters of small heads. Indeed, there would seem to be some wastage of energy in composite composites such as *Lagascea* and *Echinops*, which continue to produce apparently useless individual involucres around the one-flowered heads that make up the compound head.

CLASSIFICATION

The most widely accepted arrangement of the species of the Asteraceae family into genera and tribes dates essentially from the work of George Bentham, in 1873. Many new genera and species have been described since that time, and the limits of many genera have had to be reconsidered, but botanists still turn to his treatment for an overall view of the family and for information on the characteristics and limits of genera that have not been recently revised. Bentham recognized 13 tribes. One of these, the Helenieae, is now submerged by many students of the family in the tribe Heliantheae, but the others are still generally accepted, with only minor modifications.

Annotated classification. The annotated classification presented here is based essentially on that of Bentham, with minor modifications that recognize recent work by other authorities.

ORDER ASTERALES

Individual flowers epigynous, perfect or unisexual, sympetalous, regular or irregular, commonly 5-merous, the calyx wanting or more often represented by a set of 1 to many hairs, bristles, or scales collectively called the pappus; stamens alternate with the corolla lobes, the filaments attached in the lower part of the corolla tube, the anthers generally elongate and united into a tube; ovary bicarpellate but unilocular, with a single, erect, basal, unitegmic ovule; style growing up through the anther tube, usually 2-cleft, the branches commonly with ventromarginal stigmatic lines that do not reach the tip; fruit an achene; flowers sessile in a centripetally flowering head on a common receptacle, sometimes individually subtended by a small bract (chaff), and almost always collectively subtended by an involucre of few to many bracts; herbs, shrubs, or trees with the heads arranged in various sorts of basically cymose inflorescences. One family, about 900 genera, and 15,000 to 20,000 species with worldwide distribution.

Family Asteraceae (Compositae)

The only family of the order.

Subfamily Asteroideae

Heads radiate, discoid, or disciform, or (in Mutisieae) with the flowers bilabiate, but never (save in a few Mutisieae) strictly ligulate; plants mostly with well-developed resin ducts and without latex, or less often with latex in isolated cells or pockets, only rarely in a system of anastomosing canals; pollen grains variously spiny or nearly smooth to sometimes lophate with 30 or more lacunae, not appearing angular.

Tribe Heliantheae. Leaves (at least the lower ones) tending to be opposite, but sometimes wholly alternate; heads radiate, less commonly discoid or disciform, predominantly yellow; involucral bracts tending to be herbaceous and in several series, but often variously modified; receptacle chaffy in most genera, less often naked, the genera with naked receptacle often taken as a separate tribe, Helenieae, but actually consisting of several diverse groups separately related to the remainder of the Heliantheae; pappus of scales, or a crown, or a few firm awns, or none; anthers basally obtuse to sagittate, but not tailed; style branches often more or less short-hairy throughout and with the stigmatic lines poorly defined, varying to those as are found in the tribes Astereae or Anthemideae. About 230 genera, predominantly in the New World. Familiar genera include *Ambrosia* (ragweed), *Bidens* (beggar ticks), *Coreopsis* (tickseed), *Cosmos*, *Dahlia*, *Helenium* (sneezeweed), *Helianthus* (sunflower), *Rudbeckia* (coneflower), *Tagetes* (marigold), *Xanthium* (cocklebur), and *Zinnia*.

Tribe Astereae. Leaves alternate, generally entire or toothed, seldom dissected; heads radiate, less often discoid or disciform, variously yellow or anthocyanic, often with yellow disk and anthocyanic or white rays; involucral bracts generally in 2 or more series, herbaceous to papery in texture; receptacle naked; pappus of diverse sorts, often of numerous capillary bristles; anthers basally obtuse to sagittate, but not tailed; style branches with ventromarginal stigmatic lines and a terminal, usually externally short-hairy appendage. About 120 genera, most abundant in the New World. Familiar genera include *Aster*, *Baccharis* (groundsel tree), *Bellis* (English daisy), *Callistephus* (China aster), *Erigeron* (daisy, fleabane), *Haplopappus* (goldenweed), and *Solidago* (goldenrod).

Tribe Anthemideae. Leaves alternate, mostly more or less dissected, varying to entire and then mostly small; heads radiate to often discoid or disciform, generally yellow, or with white rays; involucral bracts mostly rather dry and scarcely herbaceous, with very thin and more or less transparent or translucent margins and tip; receptacle chaffy or naked; pappus none, or of a few small scales or a crown; anthers not tailed; style branches with ventromarginal stigmatic lines, mostly ending abruptly in a tuft of minute hairs; plants mostly with a characteristic odour, unique to this tribe. About 60 genera, mostly of the Old World, especially the Mediterranean region and South Africa. Familiar genera include *Achillea* (yarrow), *Anthemis*, *Artemisia*, and *Chrysanthemum*.

Tribe Arctotideae. Leaves alternate, often more or less spiny, as is also the involucre; heads radiate to discoid, most commonly yellow; receptacle naked or sometimes chaffy; an-

thers not tailed; style branches resembling those of the Cynareae (to which the Arctotideae show some transitional forms) or *Ursinia*, of the Anthemideae. About 15 genera, all of the Old World, mostly in South Africa.

Tribe Inuleae. Leaves mostly alternate and entire, very often white-woolly; heads radiate or discoid to very often disciform, generally yellow or yellowish; involucral bracts herbaceous to often dry and papery, sometimes brightly coloured; receptacle chaffy or naked; pappus mostly of numerous capillary bristles, rarely of scales or none; anthers prolonged at the base into slender, sterile tails; style branches generally smooth or nearly so and without definite appendages, truncate to broadly rounded, the stigmatic lines often confluent around the tip. About 160 genera, most abundant in the Old World. Familiar genera include *Anaphalis* (pearly everlasting), *Gnaphalium* (cudweed), *Helichrysum* (strawflower, everlasting), *Inula* (elecampane), *Leontopodium* (edelweiss).

Tribe Senecioneae. Leaves alternate or sometimes opposite; heads radiate to less often discoid or disciform, most often yellow; involucral bracts mostly equal and uniseriate, often with a calyculus of much reduced outer bracts, occasionally in several series; receptacle generally naked; pappus of numerous capillary bristles; anthers not tailed; style branches with ventromarginal stigmatic lines, most often truncate and with a tuft of minute hairs at the end; achenes mostly all alike. About 60 genera, widely distributed. Familiar genera include *Arnica* and *Senecio* (one of the largest genera of flowering plants, with perhaps 1,500 species).

Tribe Calenduleae. Like the Senecioneae, but the pappus wanting and the achenes mostly of different shapes, those of the outermost flowers of the head distinctly unlike the others, often of unusual form. The smallest tribe, with about 9 genera, almost entirely African, especially South African. *Calendula* is the most familiar genus.

Tribe Eupatorieae. Leaves mostly opposite or whorled; heads strictly discoid, usually anthocyanic or white, never bright yellow; receptacle usually naked; anthers obtuse or rarely sagittate at the base; pollen grains minutely spiny, not lophate (crested or with ridges); style branches with short, inconspicuous stigmatic lines and a more or less elongate, papillate, often club-shaped (thicker distally than below) appendage. About 45 genera, largely confined to the New World, most abundant in tropical and subtropical countries. *Eupatorium*, a large genus of about 600 species; *Ageratum;* and *Liatris* (blazing star) are familiar genera.

Tribe Vernonieae. Leaves alternate; heads strictly discoid, usually anthocyanic or white, not yellow; receptacle naked; anthers more or less strongly sagittate at the base, sometimes almost tailed; pollen generally lophate, with a network of usually spiny ridges enclosing 30 or more lacunae; style branches elongate and gradually attenuate, minutely short-hairy outside, generally with short, inconspicuous, ventromarginal stigmatic lines toward the base. About 50 genera, most abundant in the New World, especially in tropical and subtropical countries. *Vernonia*, with about 600 species, is the only really familiar genus. *Stokesia*, of the southeastern U.S., is occasionally cultivated as a garden flower.

Tribe Cynareae. Leaves generally alternate, often more or less spiny, as also the involucre; heads discoid, most commonly anthocyanic; receptacle bristly to less commonly chaffy or naked; anthers tailed at the base; pollen grains spiny to smooth, not lophate; style with a thickened, often minutely hairy ring and an abrupt change of texture below the branches, papillate thence to the tip, the branches commonly more or less united below. About 50 genera, mostly in the Old World, especially in the Mediterranean region and the Near East, less frequent in North America. *Arctium* (burdock), *Carduus* (thistle), *Carthamus* (safflower), *Centaurea* (knapweed), *Cirsium* (thistle), and *Echinops* (globe thistle) are familiar genera.

Tribe Mutisieae. Leaves alternate; heads generally with some or all of the corollas 2-lipped, less often all regular but very deeply cleft; anthers more or less strongly tailed; other features diverse. About 70 genera, mostly in the Southern Hemisphere or in equatorial regions, especially in the mountains of South America. *Gerbera* is the only genus familiar in cultivation.

Subfamily Cichorioideae

Heads ligulate; plants with latex in a more or less well-developed system of anastomosing canals, but without resin ducts, or occasionally with resin ducts in the roots; pollen most often lophate, appearing angular, with about 6 to 21 lacunae enclosed by a network of usually spiny ridges; leaves mostly alternate; involucral bracts often subequal in 1 or 2 series.

Tribe Lactuceae (Cichorieae). The only tribe of the subfamily. About 65 genera, mainly in the Northern Hemisphere. *Cichorium* (chicory), *Hieracium* (hawkweed), *Lactuca* (lettuce), *Sonchus* (sow thistle), *Taraxacum* (dandelion), and *Tragopogon* are familiar genera.

Critical appraisal. Although the order Asterales is morphologically (structurally) and ecologically very diverse, it is absolutely sharply defined. There is not a single species about which there is any doubt as to its inclusion in or exclusion from the order. Botanists dealing casually with the order have often been so impressed by its diversity that many have suggested it must be an artificial rather than a natural group, but serious students, both past and present, are agreed that it is highly natural.

Unity of the aster order

The great majority of species of Asterales could be included in a single genus if the standards of sharpness of distinction that are useful in some other families were rigorously applied. Only the tribe Lactuceae (Cichorieae) stands somewhat apart from the others, and it is connected to the rest by the tribe Mutisieae. In practice botanists have found it necessary, in order to have any mental organization at all, to recognize genera and tribes in the Asterales that are imperfectly defined and are connected by transitional members. The tribes more nearly represent syndromes of correlated tendencies than morphologically definable groups. Each of three tribes (Eupatorieae, Senecioneae, Vernonieae) is dominated by a single very large genus, with which most of the other genera are associated as satellites. A number of the smaller genera may prove to be misplaced, and there is need for a careful reconsideration of generic and subtribal groupings throughout the family, especially in the Heliantheae.

BIBLIOGRAPHY. ARTHUR CRONQUIST, "Phylogeny and Taxonomy of the Compositae," *Am. Midl. Nat.*, 53:478–511 (1955), the most widely accepted interpretative synthesis for the group; GEORGE BENTHAM, "Notes on the Classification, History, and Geographical Distribution of the Compositae," *J. Linn. Soc.*, 13:355–577 (1873), the running explanation to go with the Bentham and Hooker *Genera Plantarum*, a technical work including an authoritative and still useful treatment of the Asteraceae (Compositae) family; SHERWIN CARLQUIST, "Wood Anatomy of Compositae: A Summary, with Comments on Factors Controlling Wood Evolution," *Aliso*, 6:28-44 (1966), interpretation and conclusions based on the author's previous papers dealing with the individual tribes; JAMES SMALL, "The Origin and Development of the Compositae," *New Phytol.*, 16:157–177, 198–221, 253–276 (1917), 17:13–40, 69–94, 114–142, 200–230 (1918), and 18:1–35, 65–89, 129–176, 201–234 (1919), a scholarly and thorough effort to interpret the Compositae in terms of the now discredited Age and Area Hypothesis.

(A.Cr.)

Astrology

Defined as either a science or a pseudoscience, astrology —the forecasting of earthly and human events by means of observing and interpreting the fixed stars, the Sun, the Moon, and the planets—has exerted a sometimes extensive and a sometimes peripheral influence in many civilizations, both ancient and modern. As a science, astrology has been utilized to predict or affect the destinies of individuals, groups, or nations by means of what is believed to be a correct understanding of the influence of the planets and stars on earthly affairs. As a pseudoscience, astrology is considered to be diametrically opposed to the findings and theories of modern Western science.

Nature and significance. Astrology is a method of predicting mundane events based upon the assumption that the celestial bodies—particularly the planets and the stars considered in their arbitrary combinations or configurations (called constellations)—in some causal or significant way, either determine or indicate changes in the sublunar world. The theoretical basis for this assumption lies historically in Hellenistic (*c.* 300 BC–*c.* AD 300) philosophy and radically distinguishes astrology from the celestial *omina* ("omens") that were first categorized and cataloged in ancient Mesopotamia. Originally, astrologers presupposed a geocentric universe in which the planets (including the Sun and Moon) revolve in orbits

whose centres are at or near the centre of the Earth, and in which the stars are fixed upon a sphere with a finite radius whose centre is also the centre of the Earth. Later, the principles of Aristotelian physics were adopted, according to which there is an absolute division between the eternal, circular motions of the heavenly element and the limited, linear motions of the four sublunar elements: fire, air, water, and earth.

Special relations were believed to exist between particular celestial bodies and their varied motions, configurations with each other, and the processes of generation and decay apparent in the world of fire, air, water, and earth. The complexity of these relations, reflected in the complexity of the phenomenal world, was sometimes regarded as providing a system of variables that no human mind could completely control; thereby, the astrologer might be readily excused for any errors. A similar set of special relations was also assumed by those whose physics was more akin to that of the Greek philosopher Plato. For the Platonic astrologers, the element of fire was believed to extend throughout the celestial spheres, and they were more likely than the Aristotelians to believe in the possibility of divine intervention in the natural processes through celestial influences upon the Earth, since they believed in the deity's creation of the celestial bodies themselves.

The role of the divine will

The role of the divine in astrological theory varies considerably. In its most rigorous aspect, astrology postulates a totally mechanistic universe, denying to the deity the possibility of intervention and to man that of free will; as such it was vigorously attacked by orthodox Christianity and Islām. For some, however, astrology is not an exact science like astronomy but merely indicates trends and directions that can be altered either by divine or by human will. In the interpretation of Bardesanes, a Syrian Christian scholar (154–222)—who has often been identified as a Gnostic (a believer in esoteric salvatory knowledge and the view that matter is evil and spirit good)—the motions of the stars govern only the elemental world, leaving the soul free to choose between the good and the evil. And for Gnostics in general, the perceptible universe was viewed as an elaborate mechanical cage in which a bit of the divine has been entrapped by the malevolent creator of the universe, a deity believed to be in opposition to the good, true, and spiritual God, and subsequently divided into individual souls. Man's ultimate goal, then, is to attain emancipation from this astrologically dominated material world. Some astrologers, such as the Harranians (from the ancient Mesopotamian city of Harran) and the Hindus, regard the planets themselves as potent deities whose decrees can be changed through supplication and liturgy or through theurgy, the science of persuading the gods or other supernatural powers. In still other interpretations—*e.g.*, that of the Christian Priscillianists (followers of Priscillian, a Spanish ascetic of the 4th century who apparently held dualistic views)—the stars merely make manifest the will of God to those trained in astrological symbolism.

Purposes of astral omens and astrology. *Purposes of astral omens.* The view that the stars make manifest the divine will is closest to the concept that lies behind the ancient Mesopotamian collections of celestial omens. Their primary purpose was to inform the royal court of impending disasters or success. These might take the forms of meteorological or epidemic phenomena affecting entire human, animal, or plant populations. Frequently, however, they involved the military affairs of the state or the personal lives of the ruler and his family. Since the celestial *omina* were regarded not as deterministic but rather as indicative—as a kind of symbolic language in which the gods communicated with men about the future and as only a part of a vast array of ominous events—it was believed that their unpleasant forebodings might be mitigated or nullified by ritual means or by contrary omens. The *bāru* (the official prognosticator), who observed and interpreted the celestial *omina*, was thus in a position to advise his royal employer on the means of avoiding misfortunes; the omens provided a basis for intelligent action rather than an indication of an inexorable fate.

Purposes of astrology. The original purpose of astrology, on the other hand, was to inform the individual of the course of his life on the basis of the positions of the planets and of the zodiacal signs (the 12 astrological constellations) at the moment of his birth or conception. From this science, called genethlialogy (casting nativities), were developed the fundamental techniques of astrology, which were later applied to a variety of other problems. The main subdivisions of astrology that developed after genethlialogy are general, catarchic, and interrogatory.

Principal subdivisions of astrology

General astrology studies the relationship of the significant celestial moments (*e.g.*, the times of the occurrences of vernal equinoxes, eclipses, or planetary conjunctions) to social groups, nations, or all humanity. It answers, by astrological means, questions formerly posed in Mesopotamia to the *bāru*.

Catarchic (pertaining to beginnings or sources) astrology determines whether or not a chosen moment is astrologically conducive to the success of a course of action begun in it. Basically in conflict with a rigorous interpretation of genethlialogy, it allows the individual (or corporate body) to act at astrologically favourable times and, thereby, to escape any failures predictable from his (or its) nativity.

Interrogatory astrology provides answers to a client's queries based on the situation of the heavens at the moment of his posing the questions. This astrological consulting service is even more remote from determinism than is catarchic astrology; it is thereby closer to divination by omens and insists upon the ritual purification and preparation of the astrologer.

Other forms of astrology, such as iatromathematics (application of astrology to medicine) and military astrology are variants on one or another of the above.

Astral omens in the ancient Near East. The astral omens employed in Mesopotamian divination were later commingled with what came to be known as astrology in the strict sense of the term and constituted within astrology a branch described as natural astrology. Though lunar eclipses apparently were regarded as ominous at a somewhat earlier period, the period of the 1st dynasty of Babylon (18th to 16th centuries BC) was the time when the cuneiform text *Enūma Anu Enlil*, devoted to celestial *omina*, was initiated. The final collection and codification of this series, however, was not accomplished before the beginning of the 1st millennium BC. But the tablets that have survived—mainly from the Assyrian library of King Ashurbanipal (7th century BC)—indicate that a standard version never existed. Each copy had its own characteristic contents and organization designed to facilitate its owner's consultation of the omens.

The four omen categories

The common categories into which the omens of *Enūma Anu Enlil* were considered to fall were four, named after the chief gods involved in the ominous communication: Sin, Shamash, Adad, and Ishtar. Sin (the Moon) contains omens involving such lunar phenomena as first crescents, eclipses, halos, and conjunctions with various fixed stars; Shamash (the Sun) deals with omens involving such solar phenomena as eclipses, simultaneous observations of two suns, and perihelia (additional suns); Adad (the weather god) is concerned with omens involving meteorological phenomena, such as thunder, lightning, and cloud formations, as well as earthquakes; and Ishtar (Venus) contains omens involving planetary phenomena such as first and last visibilities, stations (the points at which the planets appear to stand still), acronychal risings (rising of the planet in the east when the Sun sets in the west), and conjunctions with the fixed stars.

Though these omens are often cited in the reports of a network of observers established throughout the Assyrian Empire in the 7th century BC, they seem to have lost their popularity late in the period of the Persian domination of Mesopotamia (ending in the 4th century BC). During the later period new efforts were made, in a large number of works called *Diaries*, to find the correct correlations between celestial phenomena and terrestrial

events. Before this development, however, portions of the older omen series were transmitted to Egypt, Greece, and India as a direct result of the Achaemenid domination (a dynasty ruling in Persia 559–330 BC) of these cultural areas or of their border regions.

Astral omens in Egypt, Greece, India, China, and Islām. The evidence for a transmission of lunar omens to Egypt in the Achaemenid period lies primarily in a demotic (simplified form of Egyptian hieratic writing) papyrus based on an original of about 500 BC. A more extensive use of Mesopotamian celestial omens is attested by the fragments of a book written in Greek in the 2nd century BC and claimed as a work addressed to a King Nechepso by the priest Petosiris. From this source, among others (including perhaps a work of Eudoxus of Cnidus, a Greek astronomer of the 4th century BC), the contents of *Enūma Anu Enlil* were included in the second book of the *Apotelesmatika*, or "Work on Astrology" (commonly called the *Tetrabiblos*, or "Four Books"), by Ptolemy, a Greek astronomer of the 2nd century AD, the first book of an astrological compendium, by Hephaestion of Thebes, a Greco-Egyptian astrologer of the 5th century AD, and the *On Signs* of John Lydus, a Byzantine bureaucrat of the 6th century. Yet another channel of transmission to the Greeks was through the Magusaeans of Asia Minor, a group of Iranian settlers influenced by Babylonian ideas. Their teachings are preserved in several classical works on natural history, primarily that of Pliny the Elder (*c.* AD 23–79), a Roman natural historian, and the *Geoponica* (a late collection of agricultural lore).

In various Near Eastern languages there also exist many texts dealing with celestial omens, though their sources and the question as to whether they are directly descended from a Mesopotamian tradition or are derived from Greek or Indian intermediaries is yet to be investigated. Of these texts the most important are those ascribed to Hermes Trismegistus (the Greek name for the Egyptian god Thoth) by the Harranians and presently preserved in Arabic, the *Book of the Zodiac* of the Mandeans (a still existing Gnostic sect in Iraq and Khuzistan), the *Apocalypse*, attributed to the Old Testament prophet Daniel (extant in Greek, Syriac, and Arabic versions), and *The Book of the Bee* in Syriac.

Introduction of Mesopotamian literature to India

The transmission of Mesopotamian omen literature to India, including the material in *Enūma Anu Enlil*, apparently took place in the 5th century BC during the Achaemenid occupation of the Indus Valley. The first traces are found in Buddhist texts of this period, and Buddhist missionaries were instrumental in carrying this material to Central Asia, China, Tibet, Japan, and Southeast Asia. But the most important of the works of this Indian tradition and the oldest extant one in Sanskrit is the earliest version of the as yet unpublished *Garga-saṃhitā* ("Compositions of Garga") of about the 1st century AD. The original Mesopotamian material was modified so as to fit into the Indian conception of society, including the system of the four castes and the duty of the upper castes to perform the *saṃskāra*s (sanctifying ceremonies).

There are numerous later compilations of omens in Sanskrit—of which the most notable are the *Bṛhat-saṃhitā*, or "Great composition," of Varāhamihira (*c.* 550), the Jaina *Bhadrabāhu-saṃhitā*, or "Composition of Bhadrabāhu" (*c.* 10th century), and the *Pariśiṣṭa*s ("Supplements") of the Atharvaveda (perhaps 10th or 11th century)—though these add little that is new to the tradition. But in the works of the 13th century and later, entitled *Tājika*, there is apparent a massive infusion of the Arabic adaptations of the originally Mesopotamian celestial omens as transmitted through Persian (*Tājika*) translations. In *Tājika* the omens are closely connected with general astrology; in the earlier Sanskrit texts their connections with astrology had been primarily in the fields of military and catarchic astrology.

Astrology in the Hellenistic period (3rd century BC to 3rd century AD). In the 3rd century BC and perhaps somewhat earlier, Babylonian diviners began—for the purpose of predicting the course of an individual's life—to utilize some planetary omens: positions relative to the horizon, latitudes, retrogressions, and other positions at the moment of birth or of computed conception. This method was still far from astrology, but its evolution was more or less contemporary and parallel with the development of the science of genethlialogy in Hellenistic Egypt.

Equally obscure are those individuals who, living in Egypt under the Ptolemies (a Greek dynasty ruling 305–30 BC), mathematicized the concept of a correspondence between the macrocosm (larger order, or universe) and the microcosm (smaller order, or man) as interpreted in terms of Platonic or Aristolelian theories concerning the Earth as the centre of the planetary system. They conceived of the ecliptic (the apparent orbital circle of the Sun) as being divided into 12 equal parts, or zodiacal signs, each of which consists of 30°; in this they followed the Babylonians. They further regarded each of these 12 signs as the domicile (or house) of a planet and subdivided each into various parts—decans of 10° each, *fines* ("bounds") of varying lengths, and *dōdecatēmoria* of 2°30′ each—each of which is also dominated by a planet. Scattered at various points throughout the ecliptic are the planets' degrees of exaltation (high influence), opposite to which are their degrees of dejection (low influence). Various arcs of the zodiac, then, are either primarily or secondarily subject to each planet, whose strength and influence in a geniture (nativity) depend partially on its position relative to these arcs and to those of its friends and enemies.

The zodiac

Furthermore, each zodiacal sign has a special relation with a part of the human body (the *melothesia*). The 12 signs are further divided into four triplicities, each of which governs one of the four elements. Numerous pairs of opposites (male–female, diurnal–nocturnal, hot–cold, and others), based on the speculations of the followers of Pythagoras, a Greek mystical philosopher of the 6th century BC, are connected with consecutive pairs of signs. Finally, a wide variety of substances in the elemental world and attributes of human character are more or less arbitrarily associated with the different signs. These lists of interrelationships provide the rationale for many of the astrologer's predictions.

The individual planet's influences are related both to its general indications when regarded as ominous in Mesopotamian texts and to the traits of its presiding deity in Greek mythology. But on them are also superimposed the system of the four elements and their four qualities, the Pythagorean opposites (*e.g.*, masculine and feminine, fixed and moving), and lists of sublunar substances. Furthermore, as in the omens, the modes of the planetary motions are carefully considered, since their strengths are partially determined by their phases with respect to the Sun. Also, they exert a mutual influence on each other both by occupying each other's houses and by means of conjunction and aspects—opposition (to the seventh) and quartile (to the fourth or tenth) being generally considered bad, trine (to the fifth or ninth) and sextile (to the third or eleventh) good.

Moreover, as the planetary orbits revolve from west to east, the zodiac rotates daily about the Earth in the opposite sense. From a given spot on the surface of the earth this latter motion—if the ecliptic were a visible circle—would appear as a succession of signs rising one after the other above the eastern horizon. The astrologers regard the one that is momentarily in the ascendant as the first place, the one to follow it as the second, and so on, with the one that rose immediately prior to the ascendant being the 12th. In genethlialogy each place in this *dōdecatropos* determines an aspect of the native's (a person born under a particular sign) life; in other forms of astrology the place determines some appropriate aspect of the sublunar world.

Horoscopes

The astrologer, then, casts a horoscope by first determining for the given moment and locality the boundaries of the 12 places and the longitudes and latitudes of the seven planets. He reads this horoscope by examining the intricate geometric interrelationships of the signs and their parts and of the planets of varying computed strengths with the places and with each other and by associating with each element in the horoscope its list of

Table 1: The Signs and Relationships of the Zodiac

sign	sex, nature	triplicity	house	decan		exaltation
				Greek	Indian	
Aries	masculine, moving	fire	Mars	Mars Sun Venus	Mars Sun Jupiter	Sun (19°)
Taurus	feminine, fixed	earth	Venus	Mercury Moon Saturn	Venus Mercury Saturn	Moon (3°)
Gemini	masculine, common	air	Mercury	Jupiter Mars Sun	Mercury Venus Saturn	
Cancer	feminine, moving	water	Moon	Venus Mercury Moon	Moon Mars Jupiter	Jupiter (15°)
Leo	masculine, fixed	fire	Sun	Saturn Jupiter Mars	Sun Jupiter Mars	
Virgo	feminine, common	earth	Mercury	Sun Venus Mercury	Mercury Saturn Venus	Mercury (15°)
Libra	masculine, moving	air	Venus	Moon Saturn Jupiter	Venus Saturn Mercury	Saturn (21°)
Scorpio	feminine, fixed	water	Mars	Mars Sun Venus	Mars Jupiter Moon	
Sagittarius	masculine, common	fire	Jupiter	Mercury Moon Saturn	Jupiter Mars Sun	
Capricorn	feminine, moving	earth	Saturn	Jupiter Mars Sun	Saturn Venus Mercury	Mars (28°)
Aquarius	masculine, fixed	air	Saturn	Venus Mercury Moon	Saturn Mercury Venus	
Pisces	feminine, common	water	Jupiter	Saturn Jupiter Mars	Jupiter Moon Mars	Venus (27°)

correspondences in the sublunar region. Any horoscopic diagram, of course, will yield an enormous number of predictions, including many that are contradictory or extravagant. The astrologer thus must rely on his knowledge of his client's social, ethnic, and economic background and on his own experience and good sense to guide him in avoiding error and attaining credibility.

Since about 100 BC the above method has been the essential procedure of astrology, though various refinements and additional devices occasionally have been introduced. There is space here only to mention briefly some innovations associated with the Hermetic tradition (associated with Hermes Trismegistus) and with Dorotheus of Sidon, an influential astrological poet of the third quarter of the 1st century AD. The first is the system of lots, which are influential points as distant from some specified points in the horoscopic diagram as two planets are from each other. The second is the prorogator, a point on the ecliptic that, travelling at the rate of one degree of oblique ascension a year toward either the descendant or ascendant, determines a person's length of life. Another is the method of continuous horoscopy, under which anniversary diagrams are compared with the base nativity to provide annual readings. And, finally, certain periods of life are apportioned to their governing planets in a fixed sequence; these period governors in turn share their authority with the other planets by granting them subperiods. All of these complications serve, among other purposes, to provide the astrologer with convenient excuses for his inevitable errors.

Table 2: Relationship of Positions in the Zodiac to Aspects of Life

place	dōdecatropos	
	Greek	Indian
I	life	body
II	wealth	wealth
III	siblings	siblings
IV	parents	relatives
V	children	children
VI	health	enemies
VII	marriage	marriage
VIII	death	death
IX	travel, religion	religion
X	occupation, honors	occupation
XI	benefits, friends	gains
XII	losses, enemies	losses

Astrology after the Hellenistic period. *In India.* Greek astrology was transmitted to India in the 2nd and 3rd centuries AD by means of several Sanskrit translations, of which the one best known is that made in AD 149/150 by Yavaneśvara and versified as the *Yavanajātaka* by Sphujidhvaja in AD 269/270. The techniques of Indian astrology are thus not surprisingly similar to those of its Hellenistic counterpart. But the techniques were transmitted without their philosophical underpinnings (for which the Indians substituted divine revelation), and the Indians modified the predictions, originally intended to be applied to Greek and Roman society, so that they would be meaningful to them. In particular, they took into account the caste system, the doctrine of metempsychosis (transmigration of souls), the Indian theory of five elements (earth, water, air, fire, and space), and the Indian systems of values.

The Indians also found it useful to make more elaborate the already complex methodology of Hellenistic astrology. They added as significant elements: the *nakṣatra*s (or lunar mansions); an elaborate system of three categories of *yoga*s (or planetary combinations); dozens of different varieties of *daśā*s (periods of the planets) and *antardaśā*s (subperiods); and a complex theory of *aṣṭakavarga* based on continuous horoscopy. The number of subdivisions of the zodiacal signs was increased by the addition of the *horās* (15° each), the *saptāṃśas* (4$\frac{2}{7}$° each), and the *navāṃśas* (3°20′ each); the number of planets was increased by the addition of the nodes of the Moon (the points of intersection of the lunar orbit with the ecliptic), and of a series of *upagraha*s, or imaginary planets. Several elements of Hellenistic astrology and its Sāsānian offshoot (see below *In Sāsānian Iran*), however —including the lots, the prorogator, the Lord of the Year, the triplicities, and astrological history—were introduced into India only in the 13th century through the *Tājika* texts. Besides genethlialogy, the Indians particularly cultivated military astrology and a form of catarchic astrology termed *muhūrta-śāstra* and, to a lesser extent, iatromathematics and interrogatory astrology.

In Sāsānian Iran. Shortly after Ardashīr I founded the Sāsānian Empire in AD 226, a substantial transmission of both Greek and Indian astrology to Iran took place. There were Pahlavi (Iranian language) translations of Dorotheus of Sidon, Vettius Valens, Hermes, and an Indian called (in the Arabic sources) Farmasp. Since the Pahlavi originals are all lost, these translations provided the only knowledge of the Sāsānian science. Genethlialogy in Iran was essentially an imitation of the Hellenistic (though without any philosophy), onto which were grafted some Indian features, such as the *navāṃśas* and a Śaivite interpretation of illustrations of the Greco-Egyptian deities of the decans. The most influential and characteristic innovation of the Sāsānian astrologers was the development of the theory of astrological history—that is, the writing of history, both past and future, on the basis of extensions of the techniques of the prorogator, the Lord of the Year, the planetary periods, and the continuous horoscopy employed in Hellenistic genethlialogy. This was done in conjunction with Zoroastrian millennarianism (the division of the finite duration of the material creation into 12 millennia).

In Islām. Astrology entered Islāmic civilization in the 8th and 9th centuries in three simultaneous streams—Hellenistic, Indian, and Sāsānian. Arabic translations from the Greek and Syriac represented the Hellenistic science, from Sanskrit the Indian version, and from Pahlavi the Sāsānian combination of the two. But to these influences Islāmic astrology, through the work of Abū Maʿshar, an astrologer of the 9th century, added the Harranian adaptation of the Neoplatonic definition of

Significance of Arabic astrology

the mode of astral influences in terms of Aristotelian physics. Abū Maʿshar further elaborated Sāsānian astrological history and greatly expanded the number of lots that an astrologer had to take into consideration. Much attention was paid by the Muslims to catarchic and interrogatory astrology, but, after the 9th century, Arabic astrological writings ceased to be original, degenerating into introductory handbooks or vast compendia. Under attack by the theologians for denying divine intervention in the world and man's free will, astrology rapidly declined in its appeal to Muslim intellectuals after the Mongol invasions of the 13th century, though not before its influence had been experienced in India, the Latin West, and Byzantium.

In Byzantium. During the last upsurge of paganism in the 5th and 6th centuries AD, Byzantium (the Eastern Roman Empire) boasted a host of astrologers: Hephaestion, Julian of Laodiceia, "Proclus," Rhetorius, and John Lydus. Though their works are singularly unoriginal compilations, they remain extremely valuable as the main sources for an understanding of earlier Hellenistic astrology. By the end of the 6th century, however, the general decline of the Byzantine Empire's intellectual life and the strong opposition of the church had combined to virtually obliterate astrology, though some practice of reading celestial omens survived in Byzantium as it also did in western Europe. The science was revived only in the late 8th century and the 9th century under the impact of translations from Syriac and Arabic. The period from about 800 to 1200 was the most propitious for Byzantine astrology; numerous texts were translated and many compilations assembled, but nothing was essentially added to astrological theories or techniques. This period was rivalled only by a last flowering of astrology in the late 14th century, when John Abramius and his students revised the older astrological treatises in Greek to provide the Renaissance with vulgate texts.

In western Europe. The astrological texts of the Roman Empire were written almost universally in Greek rather than in Latin; the only surviving exceptions are the poem *Astronomica* of Manilius (*c.* AD 15–20), the *Matheseos libri* ("Books on Astrology") of Firmicus Maternus (c. 335), and the anonymous *Liber Hermetis* ("Book of Hermes") from the 6th century. In the absence of astronomical tables in Latin, however, none of these was of any use, and astrology for all practical purposes disappeared with the knowledge of Greek in western Europe. It was revived only with the numerous translations of Arabic astrological and astronomical treatises executed in Spain and Sicily in the 12th and 13th centuries, supplemented by a few translations directly from the Greek. But the new astrology in the Latin-reading world remained essentially an offshoot of Islāmic astrology, gaining an adequate representation of its Hellenistic originals only in the 15th and 16th centuries. These two centuries also witnessed the fullest flowering of astrology in western Europe, frequently in conjunction with Neoplatonism and Hermeticism. By the 17th century, however—with the displacement of the Earth from the centre of the universe in the new astronomy of Copernicus (1473–1543), Galileo (1564–1642), and Johannes Kepler (1571–1630), and with the rise of the new mechanistic physics of Descartes (1596–1650) and Newton (1642–1727)—astrology lost its intellectual viability and became scientifically untenable. Though Kepler attempted to devise a new method of computing astrological influences in the heliocentric (Sun-centred) universe, he did not succeed, since no astral influences are possible in a Newtonian universe.

Effect of modern science on astrology

Astrology today. Newtonian physics eradicated a belief in astrology among the educated. The practice of the now pseudoscience continued among nonintellectuals in the West, gradually losing contact with its rich tradition and becoming more and more fraudulent, though in countries such as India, where only a small intellectual elite has been trained in Western physics, it manages to retain here and there its position among the sciences. Regardless of its validity, some Indian universities offer advanced degrees in astrology.

Recently in the West, however, astrology has regained a large popular following, though there does not seem to have been any effort made to re-establish a firm theoretical basis for it. The divisions of the year governed by the 12 zodiacal signs (which are derived from Hellenistic astrology) as depicted in newspapers, manuals, and almanacs are as follows:

Aries, the Ram, March 21–April 19
Taurus, the Bull, April 20–May 20
Gemini, the Twins, May 21–June 21
Cancer, the Crab, June 22–July 22
Leo, the Lion, July 23–August 22
Virgo, the Virgin, August 23–September 22
Libra, the Balance, September 23–October 23
Scorpio, the Scorpion, October 24–November 21
Sagittarius, the Archer, November 22–December 21
Capricorn, the Goat, December 22–January 19
Aquarius, the Water Carrier, January 20–February 18
Pisces, the Fish, February 19–March 20

Attempts have been made to incorporate into the general astrological scheme the planets discovered since the Renaissance and to find some sort of statistical relation between planetary positions and human lives. None of these attempts appear to be at all convincing, however, and no serious explanation seems to exist regarding the alleged spheres of influence of the planets, the alleged nature of their influences, or the manner in which they are received. Nor has any modern astrologer proved that arbitrary arcs (houses or zodiacal signs) on a nonexistent circle (the ecliptic) are endowed with existence and with attributes, much less with the power to affect human lives. Moreover, since the phenomena of this world are now largely explainable by the hypotheses of modern science, it is difficult to understand how astrological influences can also be responsible for them. In short, modern Western astrology, though of great interest sociologically and popularly, generally is regarded as devoid of intellectual value.

BIBLIOGRAPHY. The fundamental work on Greek astrology is A. BOUCHE-LECLERCQ, *L'astrologie grecque* (1899). Material concerning ancient astrological literature is assembled in W. and H.G. GUNDEL, *Astrologumena* (1966), though this work must be used with caution. For the older astral omens of Mesopotamia, the most convenient survey is P. HILAIRE DE WYNGHENE, *Les Présages astrologiques* (1932). Aspects of Sāsānian astrology, and in particular its application to history, are covered by D. PINGREE, *The Thousands of Abû Ma'shar* (1968); and E.S. KENNEDY and D. PINGREE, *The Astrological History of Mâshâ'allâh* (1971). The best account of astrology in Islām is by C.A. NALLINO, "Astrologia e astronomia presso i Musalmani. 1. Astrologia," in his *Raccolta di scritti editi e inediti*, vol. 5, pp. 1–41 (1944). The Indian astrologers are dealt with in detail by D. PINGREE in *Census of the Exact Sciences in Sanskrit*, vol. 1 (1970).

The philosophical criticisms of astrology developed in antiquity are discussed by D. AMAND, *Fatalisme et liberté dans l'antiquité grecque* (1945). The position of astrology in the pagan religions is reviewed by F. CUMONT, *Astrology and Religion Among the Greeks and Romans* (1912); and its role in Hermeticism and in theurgy, two of the principle esoteric philosophies of late antiquity, is expounded respectively by A.M.J. FESTUGIERE, *La Révélation d'Hermès Trismégiste*, 4 vol. (1944–54); and by H. LEWY, *Chaldaean Oracles and Theurgy* (1956).

A useful compendium of knowledge about modern astrology is N. DEVORE, *Encyclopedia of Astrology* (1947); an interesting sociological study of the rise of modern astrology in France is J. MAITRE, "La Consommation d'astrologie dans la France contemporaine," in *La Divination*, vol. 2, pp. 429–447 (1968).

(D.E.P.)

Astronomical Maps

The brighter stars and planets are easily recognized by a practiced observer. The much more numerous fainter celestial bodies can be located and identified only with the help of astronomical maps, catalogs, and in some cases almanacs.

The first astronomical charts, globes, and drawings, often decorated with fantastic figures, depicted the constellations, recognizable groupings of bright stars known by imaginatively chosen names that have been for many

centuries both a delight to man and a dependable aid to navigation. Several royal Egyptian tombs of the second millennium BC include paintings of constellation figures, but these cannot be considered accurate maps. Classical Greek astronomers used maps and globes; unfortunately, no examples survive. Numerous small metal celestial globes from Islāmic makers of the 11th century onward remain. The first printed planispheres (representations of the celestial sphere on a flat surface) appeared in 1515, and printed celestial globes made their debut at about the same time.

Telescopic astronomy began in 1609, and by the end of the 17th century the telescope was applied in mapping the stars. The invention of the telescope also made possible maps of the surface details both of the Moon and of the planets. (Maps of individual bodies are discussed in their respective articles.)

In the latter part of the 19th century, photography gave a powerful impetus to precise chart making, culminating in the 1950s in the *National Geographic Society–Palomar Observatory Sky Survey*, a portrayal of the part of the sky visible from Palomar Observatory in California, now part of the Hale Observatories.

Many modern maps used by amateur and professional observers of the sky show stars; dark nebulae of obscuring dust; and bright nebulae, masses of tenuous, glowing matter. Specialized maps show sources of radio radiation, sources of infrared radiation, and quasi-stellar objects having very large red shifts and very small images. Astronomers of the 20th century have divided the entire sky into 88 areas, or constellations; this international system codifies the naming of stars and star patterns that began in prehistoric times. Originally only the brightest stars and most conspicuous patterns were given names, probably based on the actual appearance of the configurations. Beginning in the 16th century, navigators and astronomers have progressively filled in all the areas left undesignated by the ancients.

THE CELESTIAL SPHERE

To any observer, ancient or modern, the night sky appears as a hemisphere resting on the horizon. Consequently, the simplest descriptions of the star patterns and of the motions of heavenly bodies are those presented on a sphere.

The daily eastward rotation of the Earth on its axis produces an apparent diurnal westward rotation of the starry sphere. Thus the stars seem to rotate about a north or south celestial pole, the projection into space of the Earth's own poles. Equidistant from the two poles is the celestial equator; this great circle is the projection into space of the Earth's Equator.

Figure 1 illustrates the celestial sphere as viewed from some middle northern latitude. Part of the sky adjacent to a celestial pole is always visible (the shaded area in the diagram) and an equal area about the opposite pole is always invisible below the horizon; the rest of the celestial sphere appears to rise and set each day. For any other latitude the particular part of sky visible or invisible will be different and the diagram must be redrawn. An observer situated at the Earth's North Pole could observe only the stars of the northern celestial hemisphere. An observer at the Equator, however, would be able to see the entire celestial sphere as the daily motion of the Earth carried him around.

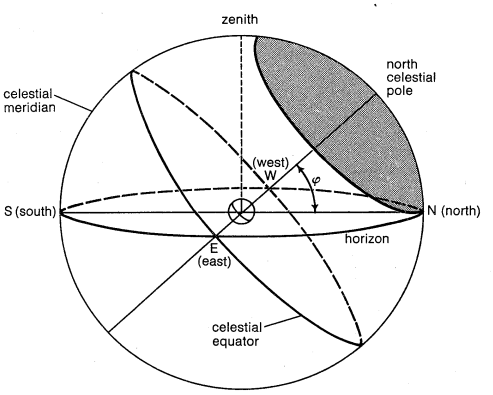


Figure 1: Celestial coordinates seen by an observer in mid-northern latitudes. His celestial meridian is a great circle passing through his zenith and the poles. His astronomical horizon meets the celestial sphere at infinity.

In addition to their apparent daily motion around the Earth, the Sun, Moon, and planets have their own motions with respect to the starry sphere. Since the Sun's brilliance obscures the background stars from view, it took many centuries before men discovered the precise path of the Sun through the constellations that are now called the signs of the zodiac. The great circle of the zodiac traced out by the Sun on its annual circuit is the ecliptic (so called because eclipses can occur when the Moon crosses it.)

The zodiac and the ecliptic

As viewed from space the Earth slowly revolves about the Sun in a fixed plane, the ecliptic plane. A line perpendicular to this plane defines the ecliptic pole, and it makes no difference whether this line is projected into space from the Earth or from the Sun. All that is important is the direction, because the sky is so far away that the ecliptic pole must fall on a unique point on the celestial sphere (Figure 2).

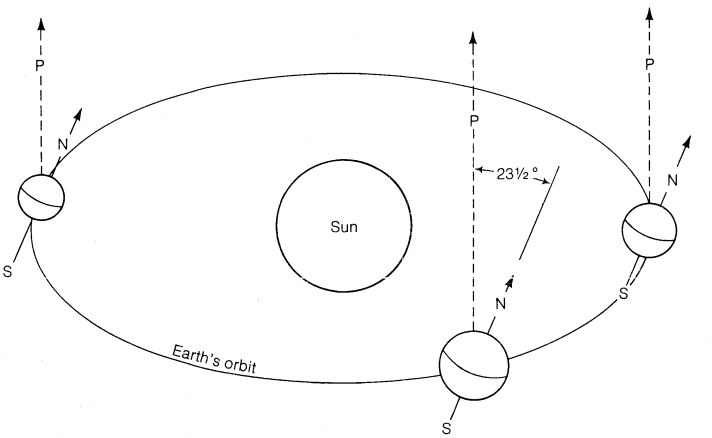


Figure 2: The north celestial (N) and the north ecliptic (P) poles at various positions of the Earth in its annual path around the Sun. Dotted lines are all perpendicular to the plane of Earth's orbit, and all intersect the celestial sphere at the north ecliptic pole. Solid lines drawn through Earth's poles, perpendicular to Earth's Equator, all point to celestial north. The angle of 23½° remains constant.

The principal planets in the solar system revolve about the Sun in nearly the same plane as the Earth's orbit, and their movements will therefore be projected onto the celestial sphere nearly, but seldom exactly, on the ecliptic. The Moon's orbit is tilted by about five degrees from this plane, and hence its position in the sky deviates from the ecliptic more than those of the other planets with the exception of Pluto.

Because the blinding sunlight blocks some stars from view, the particular constellations that can be seen depend on the position of the Earth in its orbit; that is, on the apparent place of the Sun. The stars visible at midnight will shift westward by about one degree each successive midnight as the Sun progresses in its apparent eastward motion. Stars visible at midnight in September will be concealed by the dazzling noontime Sun 180 days later in March.

Why the ecliptic and celestial equator meet at an angle of 23° 26.6′ is an unexplained mystery originating in the past history of Earth. The angle gradually varies by small amounts owing to Moon- and planet-caused gravitational perturbations on the Earth. The ecliptic plane is comparatively stable, but the equatorial plane continually shifts as the Earth's axis of rotation changes its direction in space. The successive positions of the celestial poles

Precession of the equinoxes

trace out large circles on the sky with a period of about 26,000 years. This phenomenon, known as precession of the equinoxes, causes a series of different stars to become pole stars in turn. Polaris, the present pole star, will come nearest to the north celestial pole around the year AD 2100. At the time the pyramids were built, Thuban in the constellation Draco served as the pole star, and in about 12,000 years the first-magnitude star Vega will be near the north celestial pole. Precession also makes the coordinate systems on precise star maps applicable only for a specific epoch.

CELESTIAL COORDINATE SYSTEMS

The horizon system. The simple altazimuth system, which depends on a particular place, specifies positions by altitude (the angular elevation from the horizon plane) and azimuth (the angle clockwise around the horizon, usually starting from the north). Lines of equal altitude around the sky are called almucantars. The horizon system is fundamental in navigation, as well as in terrestrial surveying. For mapping the stars, however, coordinates fixed with respect to the celestial sphere itself (such as the ecliptic or equatorial systems) are far more suitable.

The ecliptic system. Celestial longitude and latitude are defined with respect to the ecliptic and ecliptic poles. Celestial longitude is measured eastward from the ascending intersection of the ecliptic with the equator, a position known as the "first point of Aries," and the place of the Sun at the time of the vernal equinox around March 21. The first point of Aries is symbolized by the ram's horns (♈).

Unlike the celestial equator, the ecliptic is fixed among the stars; however, the ecliptic longitude of a given star increases by 1.396° per century owing to the precessional movement of the equator—similar to the precessional movement of a child's top—which shifts the first point of Aries. The first 30 degrees along the ecliptic is nominally designated as the sign Aries, although this part of the ecliptic has now moved forward into the constellation Pisces. Ecliptic coordinates predominated in Western astronomy until the Renaissance. (In contrast, Chinese astronomers always used an equatorial system.) With the advent of national nautical almanacs, the equatorial system, more suited to observation and navigation, gained ascendancy.

The equatorial system. Based on the celestial equator and poles, the equatorial coordinates, right ascension and declination, are directly analogous to terrestrial longitude and latitude (see Figure 3). Right ascension, measured eastward from the first point of Aries (see above), is customarily divided into 24 hours rather than 360°, thus emphasizing the clocklike behaviour of the sphere. Precise equatorial positions must be specified for a particular year, since the precessional motion continually changes the measured coordinates.

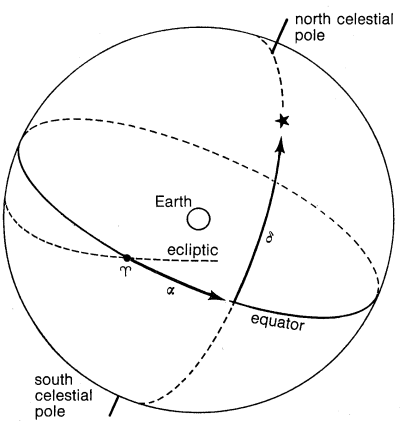


Figure 3: *Equatorial system of coordinates: right ascension (α) and declination (δ) of a star.*
Both are angular measurements, though right ascension is usually given in hours rather than degrees of arc and is measured eastward from the "first point of Aries" (♈), where the Sun crosses the celestial equator in March. Declination is considered positive (+) north of the celestial equator and negative (−) south of it.

Galactic coordinates. For problems relating to the structure of the Galaxy, astronomers have introduced the galactic equator, a great circle girdling the sky and centred in the Milky Way. Galactic longitude is measured from a specified location in Sagittarius in the direction of the nucleus of the Galaxy and is taken as positive in a direction obliquely northward in the sky (increasing declination). Galactic latitude is measured from the galactic equator and is positive toward the north galactic pole in Coma Berenices.

THE CONSTELLATIONS AND OTHER SKY DIVISIONS

The constellations carry man back to the dawn of astronomy. Their erratic star patterns, often whimsically depicted by early man, are in part the fossil remains of a primitive mythology; but as such they have extraordinary interest.

Ancient references to the constellations

The oldest astronomical cuneiform texts, from the second half of the 2nd millennium BC, record the Sumerian names of the constellations still known as the lion, the bull, and the scorpion. Drawings of these astronomical animals appear on Babylonian boundary stones of the same period, and the earlier occurrence of these motifs on prehistoric seals, Sumerian vases, and gaming boards suggests that they may have originated as early as 4000 BC. In China a handful of configurations show similarity to those of the West, including the scorpion, lion, hunter (Orion), and northern dipper, suggesting the possibility of a very old common tradition for a few groups, but, otherwise, almost complete independence.

Greek literature reflects the impact of the stars on the life of an agricultural and seafaring people. Homer (*c.* 9th century BC) records several constellations by the names used today, and the first mention of circumpolar stars is in the *Odyssey*. Odysseus is

> Gazing with fixed eye on the Pleiades,
> Boötes setting late and the Great Bear,
> By others called the Wain, which wheeling round,
> Looks ever toward Orion and alone
> Dips not into the waters of the deep.
>
> *Odyssey*, V

In England the Great Bear (Ursa Major), or Big Dipper, was still called Charles's Wain (or Wagon) in Shakespeare's day:

> An't be not four by
> The day I'll be hanged; Charles' Wain is over
> The new chimney and yet our horse not pack'd.
>
> *King Henry IV*, Part I, Act ii, Scene 1

This form derives from Charlemagne, and according to *The Oxford English Dictionary*, apparently from a verbal association of the name of the bright nearby Arcturus with Arturus, or Arthur, and the legendary association of Arthur and Charlemagne.

The earliest systematic account of the constellations is contained in the *Phaenomena* of Aratus, a poet of the 3rd century BC, who described 43 constellations and named five individual stars. Cicero recorded that

> The first Hellenic globe of the sky was made by Thales of Miletus, having fallen into a ditch or well while star-gazing. Afterwards Eudoxos of Cnidus traced on its surface the stars that appear in the sky; and . . . many years after, borrowing from Eudoxos this beautiful design and representation, Aratos had illustrated it in his verses, not by any science of astronomy, but by the ornament of poetical description.
>
> *De republica*, I, 14

The *Almagest*

By far the most important list of stars and constellations still extant from antiquity appears in the *Almagest* of the great Alexandrian astronomer Ptolemy (Claudius Ptolemaeus; flourished 2nd century AD). It contains ecliptic coordinates and magnitudes (measures of brightness) for 1,022 stars, grouped into 48 constellations. Numerous writers have stated that Ptolemy simply borrowed his material from a now-lost catalog of Hipparchus compiled in 129 BC. A critical analysis of the Hipparchian fragments still extant, including his commentary on the *Phaenomena* of Aratus, indicates that (1) the catalog of Hippar-

Constellations

name	genitive form	abbreviation*	meaning	remarks†
Constellations described by Ptolemy: the zodiac				
Aries	*Arietis*	Ari ♈	Ram	
Taurus	*Tauri*	Tau ♉	Bull	*Aldebaran:* Pleiades; M1 (Crab Nebula)
Gemini	*Geminorum*	Gem ♊	Twins	*Castor: Pollux*
Cancer	*Cancri*	Can ♋	Crab	Praesepe (star cluster)
Leo	*Leonis*	Leo ♌	Lion	*Regulus*
Virgo	*Virginis*	Vir ♍	Virgin	*Spica:* Virgo cluster of galaxies
Libra	*Librae*	Lib ♎	Balance	
Scorpius	*Scorpii*	Sco ♏	Scorpion	*Antares:* many star clusters
Sagittarius	*Sagittarii*	Sgr ♐	Archer	Galactic centre; many star clusters
Capricornus	*Capricorni*	Cap ♑	Sea-goat	
Aquarius	*Aquarii*	Aqr ♒	Water-bearer	
Pisces	*Piscium*	Psc ♓	Fishes	
Other Ptolemaic constellations				
Andromeda	*Andromedae*	And	Andromeda (Princess)	M31 (Great Spiral Galaxy)
Aquila	*Aquilae*	Aql	Eagle	*Altair*
Ara	*Arae*	Ara	Altar	
Argo Navis	*Argus Navis*	Arg	Ship Argo	now divided into Carina, Puppis, Pyxis, and Vela
Auriga	*Aurigae*	Aur	Charioteer	*Capella:* M36, M37, M38 (open star clusters)
Boötes	*Boötis*	Boo	Herdsman	*Arcturus*
Canis Major	*Canis Majoris*	CMi	Greater Dog	*Sirius* (brightest star)
Canis Minor	*Canis Minoris*	CMi	Smaller Dog	*Procyon*
Cassiopeia	*Cassiopeiae*	Cas	Cassiopeia (Queen)	Tycho's nova, 1572 (visible in daytime)
Centaurus	*Centauri*	Cen	Centaur	*Alpha* (nearest star to Sun); *Beta*
Cepheus	*Cephei*	Cep	Cepheus (King)	Delta Cephei (prototype for cepheid variables)
Cetus	*Ceti*	Cet	Whale	Mira Ceti (first recognized variable star)
Corona Austrina	*Coronae Austrinae*	CrA	Southern Crown	
Corona Borealis	*Coronae Borealis*	CrB	Northern Crown	
Corvus	*Corvi*	Crv	Raven	
Crater	*Crateris*	Crt	Cup	
Cygnus	*Cygni*	Cyg	Swan	"Northern Cross"; *Deneb*
Delphinus	*Delphini*	Del	Dolphin	"Job's Coffin"
Draco	*Draconis*	Dra	Dragon	Thuban (polestar in 3000 BC)
Equuleus	*Equulei*	Equ	Little Horse	
Eridanus	*Eridani*	Eri	River Eridanus or river god	*Achernar*
Hercules	*Herculis*	Her	Hercules (Greek hero)	M15 (great globular star cluster)
Hydra	*Hydrae*	Hya	Water Snake	
Lepus	*Leporis*	Lep	Hare	
Lupus	*Lupi*	Lup	Wolf	
Lyra	*Lyrae*	Lyr	Lyre	*Vega:* M57 (Ring Nebula)
Ophiuchus	*Ophiuchi*	Oph	Serpent-bearer	
Orion	*Orionis*	Ori	Hunter	*Rigel: Betelgeuse;* M41 (Great Nebula)
Pegasus	*Pegasi*	Peg	Pegasus (Winged horse)	Great Square (of Pegasus)
Perseus	*Persei*	Per	Perseus (Greek hero)	Algol (eclipsing star); h and χ Persei (double cluster)
Piscis Austrinus	*Piscis Austrini*	PsA	Southern Fish	*Fomalhaut*
Sagitta	*Sagittae*	Sge	Arrow	
Serpens	*Serpentis*	Ser	Serpent	
Triangulum	*Trianguli*	Tri	Triangle	M33 (nearby spiral galaxy)
Ursa Major	*Ursae Majoris*	UMa	Great Bear	seven brightest stars are Big Dipper or Plough
Ursa Minor	*Ursae Minoris*	UMi	Lesser Bear	Polaris (the north pole-star)
Southern constellations, added c. 1600				
Apus	*Apodis*	Aps	Bird of Paradise	
Chamaeleon	*Chamaeleontis*	Cha	Chamaeleon	
Dorado	*Doradus*	Dor	Swordfish	Large Magellanic Cloud
Grus	*Gruis*	Gru	Crane	
Hydrus	*Hydri*	Hyi	Water Snake	
Indus	*Indi*	Ind	Indian	
Musca	*Muscae*	Mus	Fly	
Pavo	*Pavonis*	Pav	Peacock	
Phoenix	*Phoenicis*	Phe	Phoenix (Mythical bird)	
Triangulum Australe	*Trianguli Australis*	TrA	Southern Triangle	
Tucana	*Tucanae*	Tuc	Toucan	Small Magellanic Cloud
Volans	*Volantis*	Vol	Flying Fish	
Constellations of Bartsch, 1624				
Camelopardalis	*Camelopardalis*	Cam	Giraffe	
Columba	*Columbae*	Col	Dove	constellation formed by Plancius, 1605
Monoceros	*Monocerotis*	Mon	Unicorn	
Constellations of Hevelius, 1687				
Canes Venatici	*Canum Venaticorum*	CVn	Hunting Dogs	M51 (Whirlpool Galaxy)
Lacerta	*Lacertae*	Lac	Lizard	
Leo Minor	*Leonis Minoris*	LMi	Lesser Lion	
Lynx	*Lyncis*	Lyn	Lynx	
Scutum	*Scuti*	Sct	Shield	star cloud in Milky Way
Sextans	*Sextantis*	Sex	Sextant	
Vulpecula	*Vulpeculae*	Vul	Fox	M27 (Dumbbell Nebula)
Ancient asterisms now separate constellations				
Carina	*Carinae*	Car	Keel [of Argo]	*Canopus*
Coma Berenices	*Comae Berenices*	Com	Bernice's Hair	Coma (star cluster); north galactic pole
Crux	*Crucis*	Cru	[Southern] Cross	*Alpha; Beta*
Puppis	*Puppis*	Pup	Stern [of Argo]	
Pyxis	*Pyxidis*	Pyx	Compass [of Argo]	
Vela	*Velorum*	Vel	Sails [of Argo]	
Southern Constellations of La Caille, c. 1750				
Antlia	*Antliae*	Ant	Pump	
Caelum	*Caeli*	Cae	[Sculptor's] Chisel	
Circinus	*Circini*	Cir	Drawing Compasses	
Fornax	*Fornacis*	For	[Chemical] Furnace	
Horologium	*Horologii*	Hor	Clock	
Mensa	*Mensae*	Men	Table [Mountain]	
Microscopium	*Microscopii*	Mic	Microscope	
Norma	*Normae*	Nor	Square	
Octans	*Octantis*	Oct	Octant	
Pictor	*Pictoris*	Pic	Painter's [Easel]	
Pyxis	*Pyxidis*	Pyx	Mariner's Compass of Argo	
Reticulum	*Reticuli*	Ret	Reticle	
Sculptor	*Sculptoris*	Scl	Sculptor's [Workshop]	south galactic pole
Telescopium	*Telescopii*	Tel	Telescope	

*The 12 constellations of the Zodiac are accompanied by their symbols. †First magnitude stars are given in italics.

chus did not include more than 850 stars and (2) Ptolemy most likely obtained new coordinates for even those 850 stars. The evidence suggests that Ptolemy, who for over a century has been considered a mere compilator, should be placed among the first-rank observers of all ages.

Nevertheless, Ptolemy's star list presents a curious puzzle. The southernmost heavens, invisible at the latitude of Alexandria, naturally went unobserved. On one side of the sky near this southern horizon, he tabulated the bright stars of the Southern Cross (although not as a separate constellation) and of Centaurus, but on the opposite side a large area including the first magnitude star Achernar has been left unrecorded. Because of precession, before 2000 BC this region would have been invisible from Mesopotamia. Perhaps neither Hipparchus nor Ptolemy considered that part of the heavens unnamed by their ancient predecessors. Ptolemy's catalog of 1,022 stars remained authoritative until the Renaissance.

Ptolemy divided his stars into six brightness, or magnitude, classes. He listed 15 bright stars of the first magnitude but comparatively few of the faint, much more numerous but barely visible sixth magnitude at the other limit of his list. Al-Ṣūfī, a 10th century Islāmic astronomer carried out the principal revision made to these magnitudes during the Middle Ages. Ulugh Beg, grandson of the Mongol conqueror Tamerlane, is the only known Oriental astronomer to reobserve the positions of Ptolemy's stars. His catalog, formed in 1420–37, was not printed until 1665, by which time it had already been surpassed by European observations.

Constellations of the zodiac. The Mesopotamian arrangement of constellations has survived to the present day because it became the basis of a numerical reference scheme—the ecliptic, or zodiacal, system. This happened around 450 BC, when the ecliptic was clearly recognized and divided into twelve equal signs of the zodiac. Most modern scholars take the zodiac as a Babylonian invention; the oldest record of the zodiacal signs as such is a cuneiform horoscope from 419 BC. But since Greek sources attribute the discovery of the ecliptic to Oenopides in the latter part of the fifth century BC, a parallel development in both Greece and Babylon should not be excluded.

At the time the zodiac was established, it was probably necessary to invent at least one new constellation, Libra. Centuries later Ptolemy's *Almagest* still described the stars of Libra with respect to the ancient figure of the scorpion.

The decans. Two other astronomical reference systems developed independently in early antiquity, the lunar mansions and the Egyptian decans. The decans are 36 star configurations circling the sky somewhat to the south of the ecliptic. They make their appearance in drawings and texts inside coffin lids of the 10th dynasty (around 2100 BC) and are shown on the tomb ceilings of Seti I (1318–1304 BC) and of some of the Rameses in Thebes. The decans appear to have provided the basis for the division of the day into 24 hours.

Besides representing star configurations as decans, the Egyptians marked out about 25 constellations, such as crocodile, hippopotamus, lion, and a falcon-headed god. Their constellations can be divided into northern and southern groups, but the various representations are so discordant that only three constellations have been identified with certainty: Orion (depicted as Osiris), Sirius (a recumbent cow), and Ursa Major (foreleg or front part of a bull). The most famous Egyptian star map is a 1st century BC stone chart found in the temple at Dandarah and now in the Louvre. The Zodiac of Dandarah illustrates the Egyptian decans and constellations, but since it incorporates the Babylonian zodiac as well, many stars must be doubly represented, and the stone can hardly be considered an accurate mapping of the heavens.

Hsiu and *nakṣhatra*

Lunar mansions. Called *hsiu* in China and *nakṣhatra* in India, the lunar mansions are 28 divisions of the sky presumably selected as approximate "Moon stations" on successive nights. At least four quadrantal *hsiu* that divided the sky into quarters or quadrants were known in China in the 14th century BC, and 23 are mentioned in the *Yüeh Ling*, which may go back to 850 BC. In India a complete list of *nakṣhatra* are found in the Atharvaveda, giving evidence that the system was organized before 800 BC. The system, however, of lunar mansions may have a common origin even earlier in Mesopotamia.

Relationship of the bright stars and their constellations. Ancient peoples sometimes named individual bright stars rather than groups; sometimes the name of the group and its brightest star were synonymous—as in the case of the constellation Aquila and the star Altair (Alpha Aquilae), both names meaning "flying eagles"—or were used interchangeably as in the case of both the star Arcturus (Alpha Böotis, "bear watcher") and the constellation Boötes ("plowman"). In the star list of the *Almagest*, Ptolemy cites only about a dozen stars by name, describing the others by their positions within the constellation figures. Most star names in current use have Arabic forms, but these are usually simply translations of Ptolemy's descriptions; for example, Deneb, the name of the brightest star in the constellation Cygnus (Swan), means literally "tail" of the bird.

The earliest astronomical globes and ancient maps

Ptolemy's placement of the stars within apparently well-known figures indicates the earlier existence of star maps, probably globes. An example survives in the so-called Farnese Globe at Naples, the most famous astronomical artifact of antiquity. This huge marble globe, supported by a statue of Atlas, is generally considered to be a Roman copy of an earlier Greek original. It shows constellation figures but not individual stars, although the stars may have been painted on the stone.

A unique hemispherical celestial map, which furnishes a remarkable connecting link between the classical representation of the constellations and the later Islāmic forms, is painted in the dome of a bath house at Quṣayr 'Amra, an Arab palace built in Jordan around AD 715. The surviving fragments of the fresco show parts of 37 constellations and about 400 stars.

Circumstantial evidence suggests that a flat representation of the sky, in the form of a planisphere using a stereographic projection, had come into use by the beginning of the present era. This provided the basis for the astrolabe, the earliest remaining examples of which date from the 9th century AD. The open metalwork of the top moving plate (called a spider or rete) of an astrolabe is essentially a star map, and these instruments together with associated manuscript lists provide the basic documentation for Arabic star names.

If astrolabes are excluded, the oldest existing portable star map from any civilization is the Chinese Tunhuang manuscript in the British Museum, dating from about AD 940 (see Figure 4). A Latin document of about the same age, also in the British Museum, shows a planisphere to illustrate the *Phainomena* of Aratus, without, however, indicating individual stars. The oldest illuminated Islāmic astronomical manuscript, an AD 1010 copy of al-Ṣūfī's book on the fixed stars, shows individual constellations, including stars (see Figure 5).

The earliest known western maps of the Northern and Southern Hemispheres with both stars and constellation figures date from 1440; preserved in Vienna, they may have been based on two now-lost charts from 1425 once owned by Regiomontanus. In 1515 Albrecht Dürer drew the first printed star maps, a pair of beautiful planispheres closely patterned on the Vienna manuscripts. Dürer and his collaborators numbered the stars on the charts according to the order in Ptolemy's list, a nomenclature that gained limited currency in the 16th century. The first book of printed star charts, Alessandro Piccolomini's *De le Stelle Fisse* (1540), introduced a lettering system for the stars; although frequently reprinted, application of its nomenclature did not spread.

New constellations: 16th–20th centuries. Star charts contained only the 48 constellations tabulated by Ptolemy until the end of the 16th century. Then Pieter Dircksz Keyser, a navigator who joined the first Dutch expedition to the East Indies in 1595, added twelve new constellations in the southern skies, named in part after exotic birds such as the toucan, peacock, and phoenix.

The southern constellations were introduced in 1601 on

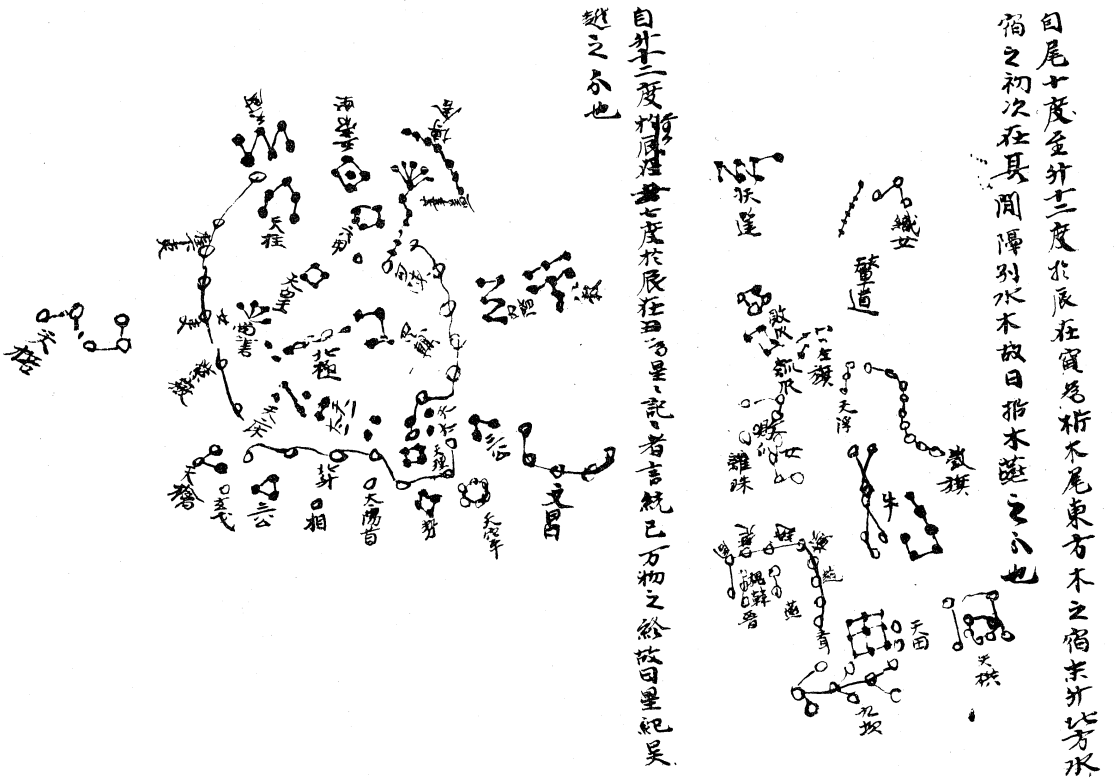


Figure 4: Chinese Tunhuang manuscript, the oldest existing portable star map, excluding astrolabes (*c.* AD 940). In the British Museum (MS. Stein 3326). Actual width of portion shown, about 12¾ inches (32 centimetres).

By courtesy of the trustees of the British Museum

a celestial globe by J. Hondius and again in 1603 on the celestial globe of Willem Blaeu and on a single plate in the *Uranometria* of Johann Bayer. The *Uranometria*, a handsome work that must be considered the first serious star atlas, devotes a large plate to each of the 48 traditional figures. (The constellation Perseus is shown in Figure 5.) Its scientific integrity rests on Tycho Brahe's newly determined stellar positions and magnitudes.

In his *Uranographia* of 1687, the German astronomer Johannes Hevelius devised seven new constellations visible from midnorthern latitudes that are still accepted, including sextans (the sextant), named for one of his own elaborate astronomical instruments. Fourteen additional southern constellations were formed by Nicolas Louis de Lacaille after his visit to the Cape of Good Hope in 1750. They appeared in the *Memoires* of the Académie Royale des Sciences for 1752 (published in 1756). All other attempts to invent constellations have failed to win acceptance.

The classic atlases of Bayer and Hevelius as well as Flamsteed's *Atlas Coelestis* (1729) showed only the brighter naked-eye stars. J.E. Bode's *Uranographia* of 1801 was the first reasonably complete depiction of the stars visible to the unaided eye. It included an early use of constellation boundaries, a concept accepted and refined by 19th-century cartographers (see Figure 6). F.W.A. Argelander's *Uranometria Nova* (1843) and B.A. Gould's *Uranometria Argentina* (1877) served to standardize the list of constellations as they are known today. They divided Ptolemy's largest constellation, Argo Navis (the ship), into three parts: Vela (the sail), Puppis (the stern), and Carina (the keel).

The list of constellations

The definitive list of 88 constellations was established in 1930 under the authority of the International Astronomical Union. Its rectilinear constellation boundaries preserve the traditional arrangements of the naked-eye stars. The smallest of the constellations, Equuleus (the little horse) and Crux (the southern cross), nestle against those over ten times larger, Pegasus and Centaurus respectively. The standard boundaries define an unambiguous constellation for each star.

STAR NAMES AND DESIGNATIONS

Star names. Of approximately 5,000 stars visible to the unaided eye, only a few hundred have proper names, and less than 60 are commonly used by navigators or astronomers. A few names come almost directly from the Greek, such as Procyon, Canopus, and Antares—the latter derived from "anti-Ares" or "rival of Mars" because of its conspicuous red color. The stars Sirius ("scorcher") and Arcturus ("bear watcher") are mentioned both by Homer and Hesiod (8th century BC?). Aratus names those two as well as Procyon ("forerunner of the dog"), Stachys ("ear of corn"?, now Spica), and Protrugater ("herald of the vintage," now Latinized to Vindemiatrix). Protrugater, a relatively faint star in a dull region of the sky, is a rare example from antiquity of a star named for obvious calendarial reasons.

Arabic names

The *Al* that begins numerous star names betrays their Arabic origin, *al* being the Arabic definite article "the": Aldebaran ("the follower"), Algenib ("the side"), Alhague ("the serpent bearer"), Algol ("the demon"). A conspicuous exception is Albireo in Cygnus, possibly a corruption of the words *ab ireo* in the first Latin edition of the *Almagest* in 1515.

Most star names are in fact Arabic and are frequently derived from translations of the Greek descriptions. The stars of Orion illustrate the various derivations: Rigel, from *rijl al-Jawzah*, "leg of Orion," Mintaka, the "belt," and Saiph, the "sword," all follow the Ptolemaic figure; Betelgeuse, from *yad al-Jawzah*, is an alternative non-Ptolemaic description meaning "hand of Orion"; Bellatrix, meaning "female warrior," is either a free Latin translation of an independent Arabic title, *an-najid*, "the conqueror," or is a modification of an alternative name for Orion himself. Only a handful of names have recent origins—for example, Cor Caroli, the brightest star in Canes Venatici, named in 1725 by Edmond Halley.

Other designations. Bayer's *Uranometria* of 1603 introduced a system of Greek letters for designating the principal naked-eye stars. In this scheme, the Greek letter is followed by the genitive form of the constellation name, so that alpha (α) of Canes Venatici is Alpha Ca-

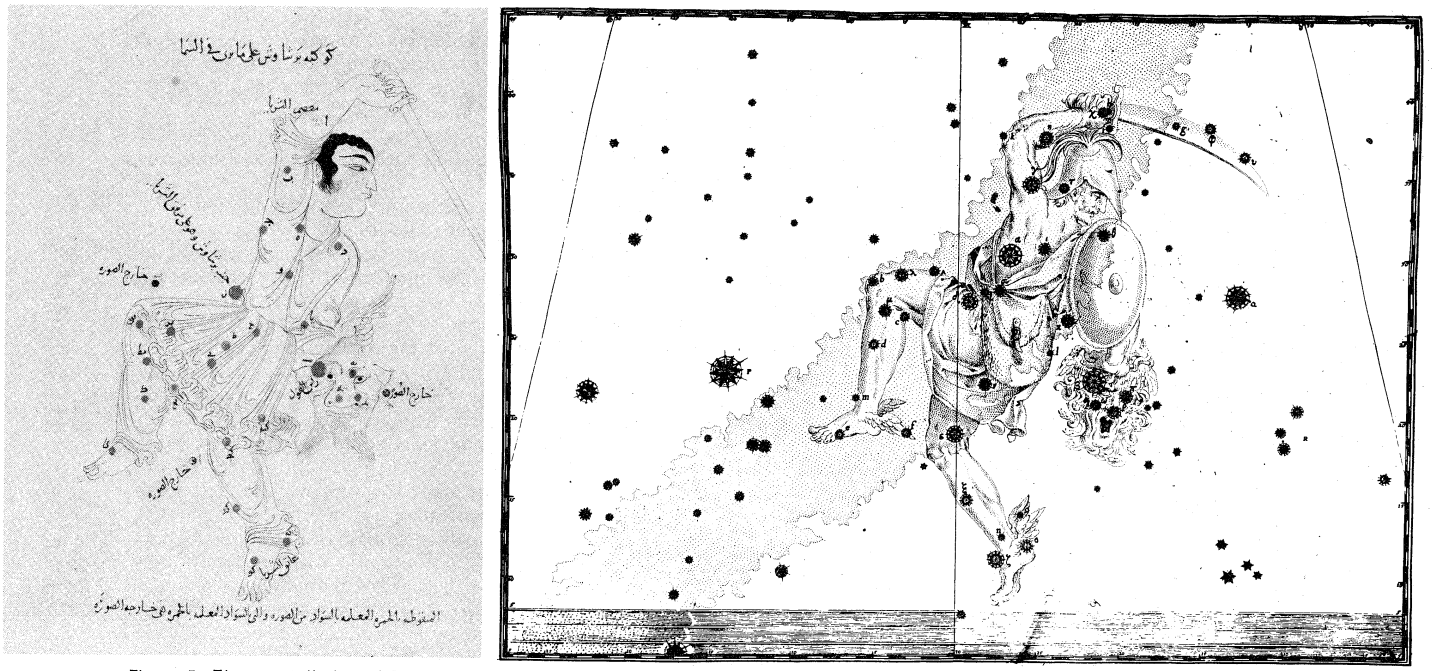


Figure 5: *The constellation of Perseus.*
(Left) From *Suwar al-Kawākil al-thābitah* ("Book of the Fixed Stars"), executed by 'Abd ar-Rahmān Ibn 'Umar as-Sūfī, AH 400 (AD 1009-10). In the Bodleian Library, Oxford (MS Marsh 144, p. 111). (Right) From Johann Bayer's *Uranometria*, 1603, a set of maps in which the use of Greek letter notation for bright stars was first introduced. The Milky Way appears as a dotted band through the figure, and the horizontal stripes at bottom indicate the ecliptic band.
By courtesy of (left) the curators of Bodleian Library, Oxford, (right) Owen Gingerich

num Venaticorum. Bayer's letters and their extension to newer constellations apply to about 1,300 stars. John Flamsteed, in his *Historia Coelestis* (1712), numbered the stars within each of 54 constellations consecutively according to right ascension, and the Flamsteed numbers are customarily used for the fainter naked-eye stars such as 61 Cygni.

Designations of faint and variable stars

An astronomer wishing to specify an even fainter star will usually take recourse to a more extensive or more specialized catalog. Such catalogs generally ignore constellations and list all stars by right ascension. Thus, astronomers learn to recognize that BD +38°3238 refers to a star in the *Bonner Durchmusterung* and that HD 172167 designates one in the *Henry Draper Catalogue* of spectral classifications; in this case, both numbers refer to the same bright star, Vega (Alpha Lyrae). Vega can also be specified as GC 25466, from Benjamin Boss' *General Catalogue of 33,342 Stars* (1937), or as ADS 11510, from Robert Grant Aitken's *New General Catalogue of Double Stars* (1932). These are the most widely used numbering systems. For more obscure names, such as Ross 614 or Lalande 21185, most astronomers would have to consult a bibliographical aid to discover the original listing.

Variable stars have their own nomenclature, which takes precedence over designations from more specialized catalogs. Variable stars are named in order of discovery within each constellation by the letter R to Z (providing they do not already have a Greek letter). After Z the double form RR to RZ, SS to SZ, . . . is used; after ZZ come the letters AA to AZ, BB to BZ, and so on, the letter J being omitted. After the letters QX, QY and QZ, the names V335, V336, etc., are assigned. Hence the first lettered variable in Cygnus is R Cygni, and the most recently named one is V1252 Cygni. The names were assigned by the Soviet authors of the *General Catalogue of Variable Stars* (3rd edition, 1969), with the approval of the Commission on Variable Stars of the International Astronomical Union.

Two catalogs are frequently used for designating clusters, nebulae, or galaxies. The shorter list of these, which includes about 100 of the brighter objects, was compiled in three installments by a French astronomer, Charles Messier, in the latter part of the 18th century; M1 and M31 are examples of this system, being respectively the Crab Nebula and the great galaxy in Andromeda. A much more extensive tabulation in order of right ascension is the *New General Catalogue* (NGC; 1890), followed by the *Index Catalogue* (IC; 1895, 1908); examples are NGC 7009 or IC 1613.

By courtesy of the Observatories, University of Cambridge

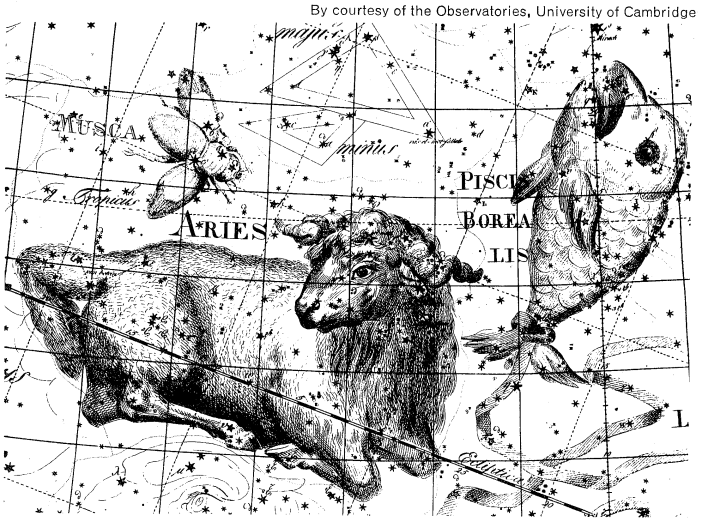


Figure 6: The constellation Aries (the Ram) and others, from J.E. Bode's *Uranographia*, 1801. The constellation Musca (the Fly) shown here is obsolete; the modern constellation Musca is in the southern skies.

MODERN STAR MAPS AND CATALOGS

Near the end of the 16th century, Tycho Brahe, a Danish nobleman, resolved to provide an observational basis for the renovation of astronomy. With his large and sturdy (but pretelescopic) quadrants and sextants, he carefully measured the positions of 777 stars, to which he later added enough hastily observed stars to bring the catalog up to exactly 1,000. A comparable catalog of southern stars was not available until 1678, when the young Edmond Halley published positions of 350 stars measured during a British expedition to St. Helena.

The first Astronomer Royal, John Flamsteed, pioneered the use of telescopic sights for measuring stars' positions. His *Historia Coelestis Britannica*, published posthumously in 1725, listed 3,000 stars, greatly exceeding all former catalogs in number and accuracy. These observations provided the basis for his great *Atlas Coelestis* (1729). The measurements of the third Astronomer Royal, James Bradley, achieved a precision within a few seconds of arc; as reduced by the German astronomer F.W. Bessel in 1818, his positions are the oldest still considered useful in modern astronomy.

Survey and zone maps and catalogs. The German word *Durchmusterung*, literally a "scanning through," was introduced by F.W.A. Argelander, an astronomer who undertook to list all the stars visible in the three-inch Bonn refractor. Keeping the telescope fixed, he recorded the stars, zone by zone, as the Earth's rotation carried the stars past the field of view. The resulting *Bonner Durchmusterung* (1859–62), or *BD* catalog, contains 324,189 stars to about the 9th magnitude between declinations $+90°$ and $-2°$. The accompanying charts, published in 1863, far surpassed all former maps in completeness and reliability. In spite of the lapse of more than a century, these maps are of continual great value today. The Bonn survey was extended to $-23°$ in 1886, and at Córdoba, Argentina, it was carried to the parallel of $-62°$ by 1908, and to the south pole by 1930; because observing conditions changed over the many years required, the resulting *Córdoba Durchmusterung*, or *CD*, lacks the homogeneity of its northern counterpart.

The *Bonner Durchmusterung*

In 1867 Argelander proposed to the German Astronomical Society (Astronomische Gesellschaft) a massive project to document stellar positions with far greater precision. Although the observing of selected star positions with meridian circle telescopes had become well established by observers in the 18th and early 19th centuries, the new plan called for meridian observations of all stars down to the ninth magnitude. A score of observatories on four continents, each responsible for a specific zone of declination, cooperated to complete the catalog and its southern supplements. The northern sections, known as the *AGK1*, were published by zones; not until 1912 was it complete to $-18°$.

Meanwhile, in quite another way, the Dutch astronomer J.C. Kapteyn completed an inventory of the southern sky by the measurement of the positions and magnitudes of about 454,000 stars from a set of photographic plates, taken in Cape Town. Known as the *Cape Photographic Durchmusterung* (1896–1900), or *CPD*, the result covers the sky from declination $-19°$ to the South Pole, down to the 11th magnitude.

Beginning in 1924, the Astronomische Gesellschaft catalog was repeated photographically by the Bonn and Hamburg-Bergedorf observatories; published in 1951–58, the new catalog is called the *AGK2*. Neither the *AGK1* nor the *AGK2* provided information on proper motions—that is, the small but perceptible individual movements of the stars in the plane of the sky. Therefore, another set of photographic plates were obtained in Hamburg during the 1950s in order to obtain the motions; the resulting *AGK3* was distributed on magnetic tape in 1969.

In 1966 the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts, issued a reference star catalog to be used as an aid in finding artificial-satellite positions from photographs. Although the *SAO Star Catalog* of 258,997 stars contains no new basic data, it does present the information in a particularly useful form. An accompanying computer-plotted atlas (1968), which includes over 260,000 stars plus galaxies and nebulae, achieves an unprecedented accuracy for celestial cartography.

Fundamental catalogs. The measurements of accurate places for vast numbers of stars rests on painstakingly and independently determined positions of a few selected stars. A list of positions and proper motions for such selected stars well distributed over the sky is called a fundamental catalog, and its coordinate system is a close approximation to a fixed frame of reference. When the German astronomers began the *AGK2* in the 1920's, they first required a fundamental reference system that by the following decade was defined in the *Dritter Fundamental-Katalog des Berliner Astronomischen Jahrbuchs*, or *FK3*. The *Fourth Fundamental Catalogue* (1963), or *FK4*, published by the Astronomisches Rechen-Institut in Heidelberg, contains data for 1,535 stars and has now superseded the *FK3*.

Photometric catalogs. A complete mapping of the sky includes magnitudes (and colours) as well as positions and motions. The great survey catalogs included magnitude estimates, but since photometric procedures are quite different from astrometric ones, a separate family of photometric catalogs has developed. Visual observations provided the basis for major tabulations published at Oxford, Harvard, and Potsdam around the turn of the century, but these were soon superseded by photographic work. Studies of galactic structure, which required accurate magnitudes for at least some very faint stars as well as the bright ones, led to the establishment of the plan of 206 selected areas. These were well-defined areas of sky containing stars of many representative kinds that could be used as standards of comparison, and the *Mount Wilson Catalogue of Photographic Magnitudes in Selected Areas* (1930), which was made about 20 years later, was for many years a leading reference for celestial photometry. Today, several catalogs of photoelectric measurements in three or more colours set the standards for precision magnitudes.

The *Mount Wilson Catalogue*

Another important physical quantity that can be measured is a star's spectral type (that is, a class assigned according to the details seen in the star's spectrum; it turns out to be an indicator of both temperature and size of the star). One of the greatest collections of astronomical data is the *Henry Draper Catalogue* (1918–24), formed at Harvard by Miss A.J. Cannon and E.C. Pickering. The *HD* lists spectra of 225,300 stars distributed over the entire sky, and the *Henry Draper Extension* (1925–36, 1949) records 133,782 additional spectra.

Photographic star atlases. Astronomical photography was scarcely past its infancy when an international conference in Paris in 1887 all too hastily resolved to construct a photographic atlas of the entire sky down to the 14th magnitude, the so-called *Carte du Ciel*, and an associated *Astrographic Catalogue*, with measured star places down to the 12th magnitude. The original stimulus had come in 1882 with the successful construction of a 13-inch astrographic objective lens at Paris. For decades the immense *Carte du Ciel* enterprise sapped the energies of observatories around the world, especially in France, and even now is incomplete in the form originally planned. Within a few years of its inception, E.C. Pickering at Harvard showed that highly corrected and efficient multiple-component lenses could considerably reduce the effort involved in mapping the sky photographically. Nowadays such a program could be speedily completed with the aid of the automatic measuring machines.

The first photographic atlas of the entire sky (if a set of 55 glass plates offered by Harvard in 1903 be excepted) was initiated by an energetic British amateur. Issued in 1914, the (John) *Franklin-Adams Charts* comprise 206 prints with a limiting magnitude of 15.

The monumental *National Geographic Society-Palomar Observatory Sky Survey*, released in 1954–58, reaches a limiting photographic magnitude of 21, far fainter than any other atlas. (The southernmost band has a slightly brighter limiting magnitude of 20.) Each field was photographed twice with a 48-inch Schmidt telescope at Mount Palomar to produce an atlas consisting of 935 pairs of prints made from the original blue sensitive and red-sensitive plates, each about 6° square. The atlas proper extends to a declination of $-33°$, but 100 additional prints from red-sensitive plates now carry the coverage to $-45°$ (see Figure 7).

Atlases for stargazing. Three modern atlases have gained special popularity among amateur as well as professional observers. *Norton's Star Atlas*, perfected through numerous editions, plots all naked-eye stars on 8 convenient charts measuring 10×17 inches and in-

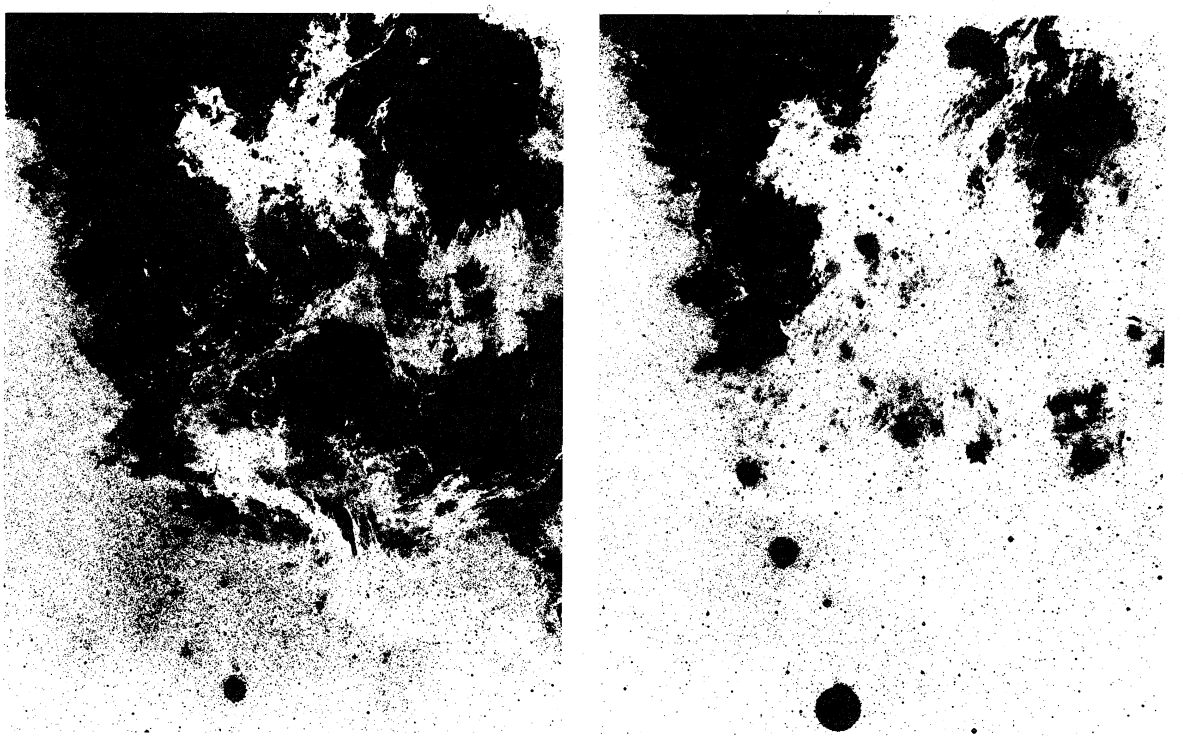


Figure 7: Negative plates showing part of the North America Nebula in (left) red light and (right) blue light. Note the much greater extent of the nebulosity in the dark areas (left).
By courtesy of Hale Observatories; © by the National Geographic Society—Palomar Observatory Sky Survey

cludes an excellent lunar map. The *Atlas Coeli 1950.0* (4th ed., 1962), commonly called the *Skalnate Pleso Atlas of the Heavens 1950.0*, shows stars to magnitude 7.75, nebulae, clusters, galaxies, and radio sources. Its 16 charts, measuring 16 by 22 inches, are famed for their cartographic clarity. An associated reference catalog gives additional information on the celestial objects plotted in the atlas. The German astronomer Hans Vehrenberg's *Photographischer Stern-Atlas* (1962–64), covering the entire sky in 464 sheets, each 12° square, has probably reached wider use than any other photographic atlas because of its quality and comparatively modest cost.

CATALOGS AND MAPS OF THE GALAXIES

Photograhic studies. The *Reference Catalogue of Bright Galaxies* (1964) gives positions, classifications, magnitudes, and other data for 2,599 galaxies brighter than about magnitude 13; it replaces the list of 1,249 galaxies tabulated in 1932 in the Shapley–Ames survey at Harvard. The *Catalogue of Galaxies and of Clusters of Galaxies* (1961–68) by the California astronomer Fritz Zwicky and his collaborators records 31,350 galaxies and 9,700 clusters found on the Palomar Sky Survey and is intended to be complete to photographic magnitude 15.5; the *Morphological Catalogue of Galaxies* (1962–68) by a group of Soviet astronomers contains data for about 29,000 galaxies on the basis of the Palomar Sky Survey, as well as data for about 5,000 additional fainter galaxies. *The Hubble Atlas of Galaxies* (1961) illustrates 176 galaxies in 50 plates and serves as a guide to a revised version of Hubble's classification scheme. The *Atlas of Peculiar Galaxies* (1966) shows 338 objects mainly on photographs taken with the 200-inch telescope on Palomar.

The sky in nonvisual wavelengths. Space-age explorations of the heavens in radio, infrared, ultraviolet, and X-ray radiations open the possibility of mapping the sky as it is seen in other wavelengths. Such maps are at best rudimentary, although the basic data are becoming increasingly available. Of particular note are the Cambridge catalogs for discrete radio sources; the successive surveys have been designated 1C, 2C, 3C, 4C, and 5C. A catalog of 2,270 radio sources at Sydney (1958–61) has been repeated at shorter wavelengths and extended at Parkes, Australia (1965–66). In radio wavelengths it is also possible to map the general level of radiation, which increases toward the Milky Way and especially toward the galactic centre. Maps of this sort have enabled astronomers to define much more precisely the system of galactic coordinates.

LUNAR MAPS

Early maps. The study of lunar topography became possible with the beginning of telescope astronomy in 1609. Galileo's *Sidereus Nuncius* (1610) contained five drawings of the Moon, all so rough that the features can be only approximately identified today. At about the same time the pioneering English scientist Thomas Harriot and his pupils examined the Moon with a telescope; but these observations were not described in print until 1788, and his map of the Moon remained unpublished until 1965.

The first lunar map to give names to features was prepared by Michel Florent van Langren, and appeared as a small broadsheet at Brussels in 1645. His map for the first time assigned proper names to 322 markings, mostly craters, but virtually none of his nomenclature has survived. The *Selenographia* (1647) of Johannes Hevelius, containing three maps and 40 drawings of the Moon, became a fundamental work in this field; nevertheless, of all the names he proposed only a few of those for lunar mountains remain in current use.

Continuing use of Riccioli's names

Much of today's basic nomenclature for the visible face of the Moon derives from a map in the *Almagestum Novum* (1651) of the Italian lunar observer G.B. Riccioli, though the actual work was done primarily by Riccioli's collaborator, F.M. Grimaldi. In this scheme, craters along the central line were named after famous philosophers and astronomers in historical order, with Plato, Archimedes, and Eratosthenes in the north and a chain leading from Ptolemy to Tycho in the south. The Jesuit authors relegated Copernicus, Kepler, and Galileo to the western quadrant, outside the mainstream of astronomy. Riccioli and Grimaldi commemorated themselves by large lunar craters even farther to the west, near the edge of the visible surface.

The finest achievement in 17th-century lunar cartography was a great, and now very rare, lunar map by the director of the Paris Observatory, Giovanni Domenico

Cassini; 53 centimetres across, this superbly engraved work prepared between 1671 and 1679 shows each feature in relief as if illuminated from the same angle. Considering that Cassini obtained his observations with a small but very long and clumsy refractor, his accomplishment is indeed noteworthy.

A new stage in selenography was inaugurated in mid-18th century when improved telescopes equipped for micrometer measurements came into use. The German astronomer Johann Tobias Mayer pioneered the new technique at Göttingen, but except for the coordinates of 23 reference points, his measurements remained unpublished for more than a century. In 1822, W.G. Lohrmann, a professional surveyor, began measuring the positions of several score craters; though completed in 1836, his topographic map was not fully published until 1878. Meanwhile, in 1836 a Berlin banker, Wilhelm Beer, and the astronomer J.H. Mädler published the *Mappa Selenographia*, as observed with a 4-inch refractor and based on 105 fundamental points related with those of Lohrmann. This landmark of lunar cartography, nearly a metre in diameter, remained unsurpassed in its wealth of information for several decades, until the appearance in 1878 of the vast lunar atlas compiled in Athens by the then director of the Athens Observatory, J.F. Julius Schmidt. Schmidt's map, divided into 25 sections, records the positions of 32,856 features.

Photographic maps. Rapid advances in photography were beginning to reduce visual selenography to secondary importance by the end of the 19th century. By 1909 three major photographic lunar atlases had appeared. The *Lick Observatory Atlas of the Moon* (1896–97) reproduced photographs taken with their 36-inch telescope, then the largest refractor in the world. The *Photographischer Mond-Atlas* (1899), prepared by Ladislaus Weinek of Prague, was based on 200 plates taken at Lick and at Paris. The *Atlas Photographique de la Lune* (1896–1909) by M. Loewy and P. Puiseux, containing enlargements of 80 plates made at the Paris Observatory, became the standard work for five decades.

The *Photographic Lunar Atlas* (1960), edited by G.P. Kuiper and his collaborators, includes 281 enlargements of plates obtained at the Lick, Mt. Wilson, Yerkes, McDonald and Pic du Midi observatories. Forty-four lunar regions are shown under four or five different angles of solar illumination. In an associated Rectified Lunar Atlas (1963), pictures of the Moon were projected onto a sphere and rephotographed in order to diminish the foreshortening caused by the curvature of the Moon's surface. The *Consolidated Lunar Atlas* (1967) consists of a better selection of photographs published in the form of positive prints. Although these will probably remain the ultimate photographic atlases resting on Earth-based observation, they have already been superseded by results from space missions. Even the far side of the Moon, unknown until the Soviet Luna-3 in 1959 and the Zond-3 in 1965, has now been recorded in more detail than the front side was in the *Photographic Lunar Atlas*.

Modern maps. A major lunar mapping project, far larger and more detailed than any previous effort, was initiated in 1960 by the Aeronautical Chart and Information Center (ACIC) of the United States Air Force. Their principal cartographic series, the 1:1,000,000 lunar charts, are based on more than 90,000 photographs secured at the Pic du Midi Observatory and supplemented by visual observations with the 24-inch refractor of the Lowell Observatory in Flagstaff, Arizona. Unlike previous lunar charts, these have been rectified to remove the perspective effects of the Moon's curvature. This ACIC series has become conveniently available in *The Times Atlas of the Moon* (London, 1969).

Mosaics of photographs from the Lunar Orbiter program provide the basis for ACIC charts of a still larger scale, 1:100,000, for possible landing sites (see MOON).

BIBLIOGRAPHY. R.H. ALLEN, *Star Names and Their Meanings* (1899; reprinted as *Star Names: Their Lore and Meaning*, 1963), a classic compendium; for a more trustworthy treatment of Arabic names, see PAUL KUNITZSCH, *Arabische Sternnamen in Europa* (1959). See also E.J. WEBB, *The Names of the Stars* (1952). JOSEPH NEEDHAM, *Science and Civilisation in China*, vol. 3, *Mathematics and the Sciences of the Heavens and the Earth* (1959); and K. LUNDMARK, "Luminosities, Colours, Diameters, Densities, Masses of the Stars," in *Handbuch der Astrophysik*, vol. 5, ch. 4 (1932–33), are both far more germane to astronomical maps than their titles imply. Current information may be found in H. EICHHORN, "Star Catalogues," in I.I. MUELLER, *Spherical and Practical Astronomy*, ch. 6 (1969); and ZDENEK KOPAL, *The Moon*, esp. ch. 15 (1969).

(O.G.)

Astronomical Spectroscopy, Principles of

Almost the only means of learning about celestial objects is through the analysis of electromagnetic radiation emitted by or reflected from them. Spectroscopy—*i.e.*, the analysis of such radiation into individual frequencies—can give a remarkable amount of information. Individual atoms and molecules absorb or emit light at characteristic frequencies, and can, therefore, be selectively observed.

The spectrum of a star is a series of overlapping images in its constituent colours, arranged in order of frequency (or wavelength). An example is the display of colours produced by sunlight when it is refracted and dispersed by a prism. Since the spectrum covers a much larger area than a single image it is much fainter than if the star would appear in integrated light.

Information on the physical state of an object can be obtained through the spectrographic comparison of different frequencies of its radiation (emission) or lack of it (absorption). Line-of-sight velocities of stars, planets, regions of the Sun, and so on, can be obtained from the Doppler-frequency shift (*i.e.*, an alteration in the frequency of the radiation from a moving source in proportion to the change in line-of-sight velocity). The spectral region that has been studied most is that of visible light with small extensions into the infrared and ultraviolet: the range of frequencies is about two to one. A much wider range of frequencies may be included, from microwave radio to X-rays, a ratio of 200,000,000 to 1. Although frequency is the more fundamental property of a wave, it is conventional to speak of wavelength, which is easier to visualize and which is related to frequency in that the frequency times the wavelength equals the speed of light. The spectroscopist is concerned with wavelengths from 20 centimetres to ten angstroms (100,000,000 Å = one centimetre). The wavelength limits of visible light are roughly 4,000 Å (blue) and 7,000 Å (red).

I. The nature of astronomical spectra
  Spectral lines and the continuum
  Molecular lines and bands
  Formation of spectra
II. The development of astronomical spectroscopy
  Early work on the solar spectrum
  Early work on the spectra of stars and nebulae
III. Principles of design for spectroscopic instruments
  The formation of spectra
  Spectral dispersion
  Spectrometry
  Spectroscopy at short wavelengths
IV. Programs in spectroscopy
  Solar spectroscopy
  Telluric and interstellar absorption line
  Stellar spectroscopy
  Planetary spectroscopy
  Spectroscopy of comets
  Spectroscopy of meteors
  Spectroscopy of nebulae
  Spectroscopy of galaxies
  Spectroscopy of quasars
  Radio spectroscopy

THE NATURE OF ASTRONOMICAL SPECTRA

Spectral lines and the continuum. The methods of spectrum analysis vary according to the wavelength region being studied. A familiar example of a spectroscope is a radio receiver, which can be adjusted to any wavelength between 550 and 1,870 metres, or to 30 metres for a shortwave set. Each broadcast station may be said to produce a "line" in the spectrum; the use of the word



Spectral image of star group Hyades, viewed through a prism placed before the telescope lens, revealing their temperature and basic composition.
By courtesy of the Department of Astronomy, University of Michigan

"line" arises from the appearance of a single visible frequency when the light is passed through a narrow slit; in a spectroscope, the series of coloured images of the slit are focussed in orderly array. The best known natural line that is produced in interstellar space by hydrogen atoms is in the radio region at 21 centimetres. Radio spectroscopists usually work at a single wavelength region. They must use not only a very sensitive receiver as their spectroscope but also a large antenna or radio telescope, which like other telescopes, selects a small area of the sky and collects from it as much energy as possible. Optical spectroscopy is done with the aid of prisms or, more commonly now, diffraction gratings, which disperse light into its component wavelengths. With suitable optical arrangements, the spectrum can be recorded on a photographic plate or measured by a photoelectric detector.

Continuous radiation: the so-called continuum

A common spectral feature is the so-called continuum, in which energy is radiated over a broad band of wavelengths, or colours. A familiar example of continuous radiation is the radiation from the glowing filament in a light bulb, or from hot coals in a fire. Because of the lack of discrete features comparable to the spectral lines, continua tend to contain less information than line spectra, but they are still important. In particular, the change of colour from red through white to blue indicates increasing temperature for a solid body and for most stars. Most astronomical continua contain absorption lines, which are closely related to the emission lines mentioned above. Whether a line from a given element will appear as an absorption or an emission line depends on physical conditions in the vicinity of the source and in the background.

Molecular lines and bands. Dust, gas, and solid bodies near a star will reflect or scatter its light, imposing in the process additional spectral features. The spectra of the planets are quite similar to the solar spectrum because they shine by reflected light from the Sun. Most contain additional absorption lines and molecular bands, however, because of gases in the planetary atmospheres. If solid surfaces can be seen, they modify the reflected continuous spectrum, the colour of which is changed by absorptions in the rock or other material. The Earth's own atmosphere adds its absorption to all astronomical spectra viewed from below it. No X-rays or ultraviolet light beyond 3,000 Å can reach the surface; oxygen and ozone absorb everything at the longer wavelengths between 1,000 Å and 3,000 Å, all gases absorb them below 1,000 Å. Discrete bands (groups of closely spaced lines) of molecular oxygen, O_2, are found at red and near-infrared wavelengths, particularly near 7,600 and 12,700 Å. Starting in the yellow wavelengths and becoming stronger and stronger in the infrared, water vapour produces many complicated bands and is opaque over large regions. Other important infrared absorbers are carbon dioxide and ozone; absorptions by extremely small (trace)

amounts of other gases can also be observed. Out to 14 microns (140,000 Å), only about half the spectrum can be observed through the Earth's atmosphere; from 14 to 1,000 microns the atmosphere absorbs everything and the opacity is complete. Because much of the infrared absorption is due to water vapour, a very large improvement can be obtained by observing from a jet aircraft; a balloon, though it can go higher, is little better because it generally floats in a cloud of its own vapour. For the ultraviolet and X-ray regions a rocket or satellite is necessary, and even for the infrared these vehicles give a clearer view.

Formation of spectra. The spectrum of a star is influenced by a number of competing effects, which may be seen most clearly in simple situations. It is useful to consider what happens to light from a strong source with a continuous spectrum as it passes through a cool gas: the case approximated by sunlight passing through the Earth's atmosphere, or by the light from a distant star passing through the interstellar medium. Radiation at certain frequencies (written as f) is absorbed by molecules in the atmosphere as they make a transition from their normal "ground state," or state of lowest energy, to some "excited state" at a higher energy (E) according to the relation $E = hf$ (in which h is a fundamental physical constant called Planck's constant). Dark lines (gaps) in the spectrum correspond to the absorbed frequencies. The excited state may be one of increased rotation or vibration of the atoms within the molecules, and the absorbed wavelengths are then in the infrared spectral region. A change in electronic motion influences frequencies in and near the visible region. Removal of an electron (ionization) or separation of the component atoms (dissociation) of a molecule requires still more energy, and results in the strong continuous absorptions of the ultraviolet spectral region. The frequencies absorbed by a given atom or molecule are unique to it, and give an absolute identification. Analysis of the closely spaced lines in a molecular band gives an excellent measurement of the temperature of a gas; this method can be applied to the atmospheres of several of the planets and the coolest stars. The width of individual lines depends mainly on the gas pressure or on the temperature if the pressure is very low.

The second simple case of influences on spectra is that of a low-pressure gas in which some of the atoms or molecules are continually receiving energy; that is, being excited by the impact of fast electrons, as in a neon sign. Relaxation of the excited particles to lower energy states, and eventually to the ground state (lowest of all), is accompanied by the emission of radiation at characteristic frequencies. Transitions that end at the ground state give bright lines at the wavelengths that are absorbed under other conditions. Perhaps the purest natural example of such radiance at discrete wavelengths (bright lines) is the aurora borealis, the northern lights produced in the upper atmosphere by fast electrons associated with the interaction of the solar wind and the Earth's magnetic field. The spectra of many gaseous nebulae in the Galaxy beyond the solar system show similar effects.

The concept of optical depth in gases. In the above examples it has been implicitly assumed that the gas is "optically thin"; *i.e.*, it is at least somewhat transparent even at the wavelengths it absorbs most strongly. The opposite extreme is a gas that is "optically thick" at all wavelengths, or at least over a wide range; a prominent example is a stellar atmosphere such as that of the Sun (though it is not equally opaque at all wavelengths). In describing depths within any gas such as an atmosphere of a star or planet, it is convenient to speak in terms of unit optical depth—the depth from which radiation has a probability of about 37 percent of escaping freely. (This percentage is used because it is numerically equal to 1/e, e being the symbol for the number 2.71818 . . ., the base of natural logarithms; with this definition, calculations are greatly simplified.) As a practical example, unit optical depth (or distance) in a fog—the point at which about one-third of the details can be seen—is quite close; on a clearer day it could be almost at the horizon.

Unit optical depth

The corresponding level in the medium depends strongly on wavelength, especially near an absorption line. It is deepest for wavelengths between lines, but even here the gas is remarkably tenuous. For the Sun, unit optical depth in the continuum occurs at pressure and density one-tenth and a few ten-thousandths respectively of the values at the Earth's surface. It was long a mystery how such a thin gas, consisting mostly of atomic hydrogen and helium, could be so opaque, even at a temperature greater than 6,000° K (about 10,300° F). The identity of the most important absorber was recognized only in 1938: it is the negative hydrogen ion, a rather loosely bound combination of a hydrogen atom with an extra electron, which can be dissociated (with corresponding absorption of energy) by any photon (the smallest radiation unit, sometimes called a "quantum" of light) of wavelength shorter than 16,500 Å.

Stellar spectra. Even when the opacity is known at each wavelength, it is a complicated problem to predict the spectrum of a star. A feeling for the processes at work may be obtained by regarding the radiation at each wavelength as arising from the level of the corresponding unit optical depth. The continuum is produced at the lowest visible level, and the centre of a strong line at a higher level. If the lines are to appear dark, this higher level must be cooler, and indeed the temperature is expected to decrease outward from any star. For the Sun, the temperature at heights somewhat above unit optical depth begins to climb again in the chromosphere, the extremely tenuous layers of the outer solar atmosphere; any lines formed there should therefore appear bright instead of dark, as is indeed observed, particularly in the far ultraviolet. A similar explanation applies to the bright lines at visible wavelengths for certain stars.

The "shape" of the continuum is a graph or plot of the amount of its intensity at different wavelengths (or spectrum). Over a considerable wavelength range, this resembles the spectrum of a "black body" (the physicist's name for a perfect radiator), which has a definite shape for each appropriate temperature. In general, the apparent temperature deduced from one region may differ from that measured from another. The reason for this is that the continuous absorption coefficient; that is, the ratio of absorption to emission at a given temperature is changing. The radiation, consequently, comes from different depths and, therefore, from different temperature regions. Temperature found by measuring the shape of the continuum (apparent temperature) is only approximately equal to the actual temperature at the "surface" (unit optical depth).

Planetary spectra. For planets and satellites, problems arise from the presence and amount of atmosphere. Such bodies can have a spectrum that combines contributions from two, almost independent, sources—reflected solar radiation and their own thermal emission at much longer wavelengths. The thermal component is similar to black-body radiation for airless bodies such as Mercury and the Moon; the reflected light, however, does show coloration and some broad absorptions because of surface constituents. For planets with appreciable atmosphere, the physics of the thermal emission resembles that for a star, discussed above. For Mars, the reflected solar radiation can be regarded as coming entirely from the surface, modified by two passages through the thin atmosphere. This approximation breaks down in the ultraviolet and is similarly poor for the Earth, especially in hazy or cloudy regions. Deep, cloudy atmospheres are present on Venus, Jupiter, Saturn, Uranus, and Neptune; the returned solar radiation must be regarded as having been scattered by the same medium that does the absorbing. Again the situation resembles that of a stellar atmosphere, except that the radiation is imposed from outside instead of flowing up from the interior.

Two sources of details in planetary spectra

THE DEVELOPMENT OF ASTRONOMICAL SPECTROSCOPY

Early work on the solar spectrum. In 1666, the English physicist Isaac Newton discovered how to produce a spectrum of the Sun, by means of a glass prism and a

hole in a window blind. Much later, 1800–02, several important discoveries were made with similar equipment. The English astronomer William Herschel probed the solar spectrum with a thermometer and found the measurable heat produced to continue beyond the red end of the spectrum; he had discovered the infrared. In 1802, the ultraviolet was observed by its ability to darken silver chloride. And at this time, an English scientist noted a few dark lines in the solar spectrum, which was studied in much more detail by the German physicist Joseph von Fraunhofer in 1814. Fraunhofer was able to map 754 lines in the solar spectrum, the most prominent of which are still called "Fraunhofer lines." He also observed several bright stars and found that their lines differed from those of the Sun. Fraunhofer made the first crude diffraction grating (an instrument that uses different deviations at different wavelengths to separate light of different colours) by stretching wires between the threads of two screws; with it he was able to measure actual wavelengths in centimetres.

The work of Fraunhofer

Interpretation of the Fraunhofer lines awaited further developments in physics, which in turn depended on further laboratory work. Similar delays in understanding of discoveries recurred until well into the 20th century; only since then has immediate physical interpretation been available for the majority of astronomical discoveries. Earlier, however, astronomers often found many uses for empirical classifications (particularly of stellar spectra) long before a detailed understanding was reached.

Another point to be noted is the importance of advances in technology, as illustrated by Fraunhofer's work and that of his successors; in his day, it was difficult to obtain glass in large, homogeneous pieces. Telescope lenses were small, and reflecting telescopes did not become practical until methods were developed, initially in 1856, for the deposition of a reflecting film of silver on the front surface of a glass mirror. Without large reflecting telescopes, spectroscopy would still be limited to the study of the Sun and a few bright stars.

In 1859, the German physicist Gustav Robert Kirchhoff formulated his laws connecting the absorption and emission of light and explained that the Fraunhofer lines were due to absorption by familiar elements present on the Sun. He pointed out that a perfect absorber (described by physicists as a black body) should also be an ideal emitter of light, and suggested its realization in the form of a small hole leading into a large cavity. Two years later, the Swedish physicist Anders Jonas Ångström found the lines of hydrogen on the Sun and initiated a program of wavelength measurements in the unit that now bears his name.

In 1868, the relatively faint solar prominences at the edge of the Sun's disk, the limb, were observed in full daylight. Subsequently, in Sweden the rotation of the Sun was determined from the Doppler effect at its limbs and found to be different in different latitudes of the Sun. A United States physicist, Henry A. Rowland, devised methods of ruling diffraction gratings of very high quality; some Rowland gratings are still in use. In the years 1887–96 he produced a solar spectrum of very high resolution and a catalog of wavelengths and intensities. The spectroheliograph, which enabled observation of the whole Sun or large areas of it simultaneously at a particular wavelength, was invented independently in the United States and France. In 1908 the large magnetic fields in sunspots were discovered from observation of the Zeeman effect; that is, the shift in the positions, shapes, and structure of the sunspots' spectral lines that results from the interaction of the atoms with the magnetic field. In recent years the solar magnetograph has been used to extend this work to the production of detailed maps of the solar magnetic field.

Early work on the spectra of stars and nebulae. In 1855, an English astronomer, William Huggins, began the study of the spectra of the Sun, stars, nebulae, comets, and planets. He reaped a harvest of astronomical discoveries, mainly by visual observation of spectra, though he succeeded in applying photography in 1875. He found many different types of stellar spectra. He showed from the nature of its emission spectrum that the Orion Nebula is gaseous, but found the Andromeda Nebula (now known to be a galaxy) to have a stellar-type spectrum and therefore concluded that it was composed of stars. He measured some of the first stellar Doppler shifts and obtained the corresponding line-of-sight, or radial, velocities.

Line-of-sight velocities. The German astronomer Hermann Karl Vogel was the first to introduce the photographic method into spectroscopy, in 1873, and succeeded in determining accurate radial velocities of many stars by measuring the positions of the lines on the photographic spectra. Vogel's work, especially his measurements of radial velocities, stimulated astronomers elsewhere near the turn of the century to undertake systematic studies in this field. At Victoria, British Columbia, for example, many spectroscopic double stars were discovered and many of their orbits have since been determined from measurements of the spectral lines. At Mount Wilson Observatory (now part of Hale Observatories), California, the velocities of hundreds of faint stars were determined with an increase of precision; in the case of the bright star Arcturus, the velocity was measured to about ± 0.01 kilometre per second or 0.006 miles per second. Soviet scientists working at Pulkovo, near Leningrad, and, more recently, at Simeiz, in the Crimea, also made contributions in this field. Numerous measures of stellar motions in the southern sky have been published.

The first measurements from photographs

Chemical composition. A striking result of astronomical spectroscopy is the discovery of the uniformity of chemical composition throughout the universe. Belief in the universal occurrence of the same chemical elements was greatly strengthened when the lines of helium, originally found in the spectrum of the chromosphere of the Sun in 1868, were first produced in terrestrial laboratories about 25 years later. In 1941 the last great enigma of line identification was solved when previously unidentified lines of the solar corona were found to originate in the atoms of the common elements iron, calcium, nickel, and argon excited to a degree of ionization not even dreamed of previously. Earlier, when the so-called nebulium lines in certain nebulae were disposed of in a similar manner, it was remarked that the mysterious substances of the astrophysicists one after another disappeared into "thin air"; the nebulae had, in fact, been found to consist of oxygen, nitrogen, hydrogen, and a few other gases—not very different in composition from air.

The principle of the uniformity of chemical elements has now been extended to include even the distant galaxies, whose light travels for hundreds of millions of years at the rate of 186,000 miles per second before it reaches the eye of the observer. This uniformity means that the atomic building blocks of the universe are the same throughout space. It does not mean that the proportions of these elements are the same in all astronomical objects. Certain stars differ strikingly from others in their composition, and among the most interesting spectroscopic problems of today is the measurement of these differences in the abundance of the elements. Such measurements permit the nuclear processes responsible for the formation of the elements to be inferred. The occasional observation of the unstable element technetium shows that it must have been produced within the last few million years, indicating that element building is still proceeding in stars.

PRINCIPLES OF DESIGN FOR SPECTROSCOPIC INSTRUMENTS

The formation of spectra. Astronomical spectroscopes operate on the same principles as their laboratory counterparts, though they have special features of their own. Spectroscope is used here as a general term for any kind of spectroscopic instrument, such as a spectrograph, which uses photographic detection, or a spectrometer (also called a scanner by astronomers), which uses photoelectric detection.

Spectrographs are used with ground-based (reflecting or refracting) telescopes. They operate at wavelengths from

3,000 Å in the ultraviolet (limited by atmospheric ozone absorption) to about 10,000 Å in the infrared (limited by the sensitivity of available photographic emulsions); most work is done at wavelengths in the range 3,500–7,000 Å.

Optical limitations

Further limitations are encountered if refractors are used because glass refracts different colours by different amounts (the refractive index across the spectrum varies) and only a short wavelength region can be brought to a good focus. The same effect in cheap or defective field or opera glasses can cause coloured halos around objects in view.

High photographic speed, which is necessary when faint objects are to be recorded, requires a camera lens, or mirror system, of short focal (f) length and large diameter; that is, a small f-number just as in an ordinary camera. Schmidt cameras (named after their inventor, Bernard Schmidt, the German optical worker) can be made at $f/1$ or slightly less, in sizes large enough to accept the light beam from any available grating. In small sizes, semisolid cameras, in which the photographic film is pressed against the glass of the mirror system, can be as fast as $f/0.35$. The limits of faintness that can be recorded have been greatly extended as a result of such increases in speed.

The collimator, a lens introduced to make the light beam parallel, and grating must be matched to the telescope, as illustrated in Figure 1. The telescope admits

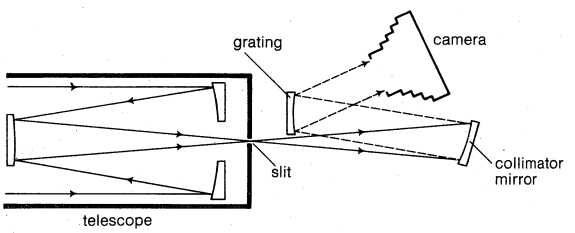


Figure 1: Astronomical spectrograph mounted at the focus of a Cassegrain reflecting telescope. The cones of light converging from the telescope and diverging to the collimating mirror must be matched (see text).

parallel light that is refracted or reflected into a cone that should just fill the grating. A small grating is used with a short-focus collimator; and the longest useful focal length is limited by the size of available gratings. The advantage of a long focal length is that a wider slit can be used for a given resolution (*i.e.*, the degree of spectral purity). The choice of resolution depends on the nature of the investigation and on the brightness of the source; higher resolution always requires longer exposure. For extremely faint stars, or galaxies, the spectrum must be kept short—that is, the dispersion must be kept low, if there is to be any blackening at all in the image; a small grating and a wide slit can be used to prevent any cutoff of light at the slit jaws. Overlapping images of nebulae of galaxies are produced by instruments, called nebular spectrographs, which have such arrangements. On large telescopes, such instruments are placed at the focus of the primary mirror (the prime focus) to minimize the loss of light by extra reflections. Medium-dispersion spectra can be obtained with instruments of moderate size that can conveniently be used as shown in Figure 1. High dispersion spectra are produced by instruments of large size that must be fixed and are therefore mounted at the *coudé* ("elbowed") focus, to which the light is brought by means of two or three additional plane mirrors. In the *coudé* spectrograph many combinations of grating size and camera focal length (but the same camera diameter) are possible. The longest cameras are rather slow photographically; moreover, it is not unusual to lose 90 percent of the light from the telescope behind the slit edges (called jaws), because even a star has a finite size and when its diameter is increased, or blown up, by atmospheric turbulence ("poor seeing"), it can be wider than the slit and only part of the light can pass through. Under such conditions, much of the light gathered by a large telescope is wasted. Only the development of much larger gratings, or a different approach to spectroscopy, can change this situation.

Comparison spectrum

For accurate wavelength identification and measurement, it is essential to include a comparison spectrum from a laboratory source on the same plate. The usual practice is to illuminate both ends of the slit with a spectrum produced by an electric arc, so that the astronomical spectrum is flanked by two comparison spectra. Neon and iron arcs are most commonly used. Use may also be made of natural lines from the Earth's airglow, frequently accompanied by mercury lines from the lights of nearby or distant cities.

Since about 1965, there has been increasing use of image-tube spectrographs for recording very faint objects. In this method, the spectrum is focussed on a photoelectric screen, the cathode, which emits a pattern of electrons. These electrons are focussed by electric and magnetic fields on a fluorescent screen, or sometimes directly on a photographic plate in the vacuum of the tube. If a fluorescent screen is used, the light from it is then transferred to a photographic plate.

Spectral dispersion. *Low-dispersion spectra.* Because the light from a star reaches the Earth in a parallel beam, it is possible in some arrangements to dispense entirely with the preliminary telescope, optical system, slit, and collimator. The essential parts of a spectrograph then remain: the grating (or prism) and the camera (usually a wide-field telescope). When photographed, the star field becomes a field of spectra; these spectra are usually widened for more convenient viewing by a slow motion of the camera at right angles to the dispersion. The ability to take many spectra at once is the chief virtue of the method of "slitless," or "objective prism," spectroscopy. The enormous *Henry Draper Catalogue*, published 1918–24, in which 225,300 bright stars were classified, was produced from objective-prism photographs with small telescopes; more recently, very good quality spectra, over a large angular area of sky, have been obtained with prisms or gratings with angles of a few degrees placed in front of Schmidt-type (wide-field) objectives (see below, spectroscopy at short wavelengths). The slitless method is also used in space spectroscopy, in which compactness and a small number of reflections are advantages. Objective-grating spectrometers and spectrographs have been used in rockets in several experiments and also on the first "Orbiting Astronomical Observatory" in 1970.

Radial velocities (velocities in the line of sight from object to observer) can be measured by the Fehrenbach method developed in the 1940s and named after the French astronomer Charles Fehrenbach. A special prism is used, containing two elements of different glasses, somewhat as in an achromatic lens. The angles are chosen so that a central known wavelength is not deviated; red goes one way and blue the other. Two exposures are made on one plate, with the prism rotated 180 degrees between them. The stellar spectrum can thus be made to serve as its own reference. With suitable precautions, a precision of three to five kilometres per second is achieved.

High-resolution spectra. Spectroscopy of solar, planetary, and interplanetary lines usually requires high spectral resolution. Modern solar work is concentrated on obtaining accurate line profiles; that is, measures of the distribution of absorption or emission with wavelength, which can be obtained only at high resolution. Planetary and interstellar lines are very narrow, because temperatures and pressures are low; such lines may not even be detectable at low and medium resolution because the overlap from neighbouring wavelengths may completely mask them. A further difficulty with planetary atmospheres is that nearly all their absorption lines are in the infrared, for which sensitive detectors do not exist.

The diffraction grating, described above, though valued for its versatility, is only one of several means for selecting wavelengths.

The Fabry-Perót interferometer consists of a pair of closely and accurately spaced plane-parallel glass plates, with special reflecting coatings on the inner surfaces. Used with photoelectric detection, it has become popular

since the introduction of efficient evaporated coatings for the two reflecting surfaces. The plates are located in a chamber in which the pressure can be changed, which makes it possible to measure radiation over a very narrow range of the spectrum. Its greatest virtue is the large angular acceptance; high resolution can be obtained, with no loss of light at a slit, when it is used with the largest telescopes. It is inconvenient when used to scan more than a small range of wavelengths; the Fabry-Perót interferometer is most useful for detailed exploration of a single line for profiles or Doppler shifts.

High resolution spectra in the infrared region

For the infrared above 10,000-angstroms wavelength, an old device, the Michelson interferometer has been perfected into a valuable spectroscopic tool. The principle of Fourier spectroscopy, in its modern form, was suggested by the English astrophysicist Peter Fellgett in 1950, and was subsequently applied successfully to astronomical spectroscopy. The angular acceptance of this instrument is large, like that of the Fabry-Perót interferometer, and, in addition, there is the multiplex advantage, or Fellgett advantage: all wavelengths of the spectrum are measured simultaneously and recorded in a coded form, called an interferogram. The spectrum is obtained from the interferogram by the mathematical process of Fourier transformation (see also OPTICS) in an electronic computer. This method has been used to produce spectral atlases of the one to three micron region for the Sun, the four brightest planets, and several bright stars. The spectral atlases of the inner planets are of better quality than solar atlases of the early 1950s in spite of the enormous differences in light level. The improvement in resolution, by a factor of ten to 100 over the best previous work, has resulted in the identification of several new gases in the atmospheres of Mars and Venus. At shorter wavelengths, for which photomultipliers are available, the Fellgett advantage is cancelled by the corresponding increase of the noise in the detector, and the method is little better than older conventional methods.

Still further in the infrared, medium-resolution spectroscopy is possible with a small Fourier spectrometer, or with a filter disk. An interference filter (a version of the Fabry-Perót interferometer: see above) is deposited around the rim of a transparent disk so that the transmitted wavelength varies smoothly with angle. The spectrum is scanned by the slow rotation of the disk past an aperture, through which the radiation passes to a detector.

At radio wavelengths, an enormous increase of sensitivity is possible because the electromagnetic waves can be treated as electrical signals rather than crude energy fluxes. The difference may be compared to that of using a radio receiver rather than a sensitive thermometer in trying to detect the presence of broadcast radio signals. As mentioned in the introduction, a standard radio receiver is a spectrometer, and the methods used for radio astronomy differ only in technical detail from those of astronomical spectroscopy. Nevertheless, some of the details are important. It is common to obtain a multiplex advantage by the use of multiple channels—essentially, many receivers tuned to adjacent frequencies. Or, a correlation technique may be used; in principle it resembles an electrical form of the Fourier spectrometer, and like it uses an electronic computer.

Radio spectroscopy

Spectrometry. In a spectrometer, a selected wavelength or band of wavelengths from the spectrum passes through an exit slit to a detector. At wavelengths in the ultraviolet, visible, and near infrared ranges, the most sensitive detector is a photomultiplier. At longer wavelengths the preferred detectors are photoconductive cells of lead sulfide, or bolometers (heat detectors) of germanium. From the visible onward, the detectors must be refrigerated at lower and lower temperatures by means of ordinary refrigeration systems, solidified carbon dioxide, liquid nitrogen, and finally liquid helium.

A spectrum can be produced by rotation of the grating, which progressively changes the wavelength that falls on the slit; the successive readings of the detector are recorded. The advantages over photography are the possibility of more accurate measurement, a strictly linear intensity scale, and greater sensitivity at longer wavelengths. The spectrum must, however, be recorded one wavelength at a time while the light of all the rest of the spectrum is wasted. Even so, photoelectric recording may be superior, especially if only a short region of the spectrum is required and if the external conditions do not change during the exposure. Also, various schemes of multiplex spectroscopy are possible, by which many wavelengths can be observed simultaneously. Most of them show their advantage best in the infrared; with photomultipliers, the most practical scheme is through the use of many detectors, each with its own amplifying system.

Spectroscopy at short wavelengths. Spectroscopes designed for the shortwave end of the spectrum, the ultraviolet, do not differ in principle from those used for visible wavelengths. An important change of detail occurs at 1,200 Å: shorter wavelengths are absorbed by all known materials, and, therefore, prisms, lenses, and windows cannot be used. In addition, unnecessary reflections must be avoided because efficient reflecting surfaces do not exist. Concave gratings and windowless photomultipliers are, therefore, used. All instruments must be carried above the atmosphere by rockets or satellites; consequently, sizes are limited and photoelectric detection is favoured, so that the data can be transmitted to the ground by radio. Below 100 Å, X-ray techniques become applicable, and windows can again be used. To improve their efficiency, concave gratings may be used at very small angles of incidence ("grazing" incidence), at which the reflection is almost total, or they may be replaced by Bragg crystals, which operate on the same principles as gratings, but use layers of atoms instead of ruled lines. At the shortest wavelengths, below 10 Å, low-resolution spectroscopy can be done by measurement of the photon energy; one suitable device for this measurement is the proportional counter.

X-ray spectroscopy

Developments in instrumentation have continued to be important in extending the range of spectroscopy. In 1931 reflecting surfaces with evaporated coatings of aluminum were found to be superior to those of silver for resisting tarnish and for achieving higher reflectivity in the ultraviolet. At about the same time, Schmidt developed his wide-field, short-focus camera, which after a few years came to be widely applied in astronomical spectrographs. In 1936, improvements in ruled gratings made possible the concentration of light to an efficiency of 60 percent or better (such a concentration is called a "blaze"). The use of large plane gratings and Schmidt cameras (instead of one or more prisms with lens cameras, which absorbed a significant proportion of the incoming light) totally altered the appearance and the efficiency of astronomical spectrographs. Although photoelectric cells had been used for stellar photometry for several decades, their application to spectroscopy was not practical until the development of the photomultiplier between 1935 and 1939. Later, a great interest developed in image tubes and television techniques.

PROGRAMS IN SPECTROSCOPY

Solar spectroscopy. The Sun is a typical star, near the middle of the known range in both temperature and luminosity. It offers a unique opportunity for detailed study because of the enormous amount of available light and because light from different regions of it can be studied separately. More distant stars are probably at least as complicated as the Sun, and present models of them may omit essential features.

Length of solar spectrum at high dispersion

The solar spectrum can be observed in great detail; the available light can be spread out into a spectrum many metres, or yards, long. The dispersion can be made so great that many centimetres may be needed to record the lines in a single angstrom. It is clear, therefore, that with high resolution, a spectrum of very great purity can be recorded and many details of solar behaviour can be accurately studied.

By using a very long focus telescope, a large image of the Sun can be produced, even when the light-gathering power (determined by the mirror or lens diameter) of the telescope is quite small. Quite small areas such as sunspots, the chromosphere, a flare, or granules (see SUN)

can be studied in detail. A major difficulty in most solar observations is the heating of the instrument and its environment by the Sun, although many sophisticated methods have been tried to overcome it.

Because of the problem of heating, solar telescopes are usually of small to moderate aperture but of long focal length, so that a relatively small amount of light is collected even when a large image is produced. A solar telescope usually has a fixed combination of mirrors, and the light is reflected into it by an arrangement of one (heliostat) or two (coelostat) auxiliary moving plane mirrors. To minimize the effect of the heating of the ground, these mirrors may be mounted on a high tower; or the telescope may be placed in the middle of a small lake. The associated spectroscopes usually feature very high dispersion and spectral resolution, and, therefore, also normally have a long focal length.

A special feature of solar spectroscopy is the study of limited areas that may be much fainter than the disk, or photosphere, such as sunspots on the disk, and the chromosphere and corona near and above the limb. Detailed study of such features requires extreme precautions to avoid inclusion of stray light from the photosphere; telescopes designed for observation of the chromosphere and corona are often called coronagraphs. The chromosphere and corona are optically thin, and the temperature increases with height—that is, distance outward from the Sun; for both reasons, their line spectra appear in emission, not absorption. The classical way of studying them is during a total eclipse, when the Moon just hides the photosphere; the great disadvantage of this method is the infrequency of solar eclipses that occur in convenient locations. The name "flash spectrum" refers to the spectrum of the chromosphere at eclipse (see also ECLIPSE, OCCULTATION AND TRANSIT). It can also be observed, with care, outside eclipse by setting a solar image tangent to a spectroscope slit. Chromospheric features on the disk are best studied in the far ultraviolet because relatively little such light is produced from the underlying photosphere. As the Earth's atmosphere absorbs ultraviolet light, it is necessary to make such observations from outside it. The series of space probes called Orbiting Solar Observatories has been particularly successful, and good results have also come from many rocket flights. Another method is that of surveying the Sun at a single wavelength by a spectroheliograph in which a spectrograph functions as a very narrow-band filter. An image either of the whole disk or of part of the Sun can be built up by a scanning process. Some of the same results can be obtained by a Lyot filter, which is built up of elements of quartz crystals and polarizers. The two commonest wavelengths studied belong to hydrogen (6,563 Å) and ionized calcium (3,933 Å); here, the light from the photosphere is reduced by the presence of deep Fraunhofer lines, and the chromospheric emission can be observed.

Solar characteristics

Spectroscopy can be expected to give information of several kinds on the chemical and physical state of the solar atmosphere: the composition, the temperature, the density, and the state of motion. The best results are obtained from detailed descriptions of the atmosphere (called models) that include all these factors at once. In such a model there is no single temperature for the atmosphere, but a profile relating the temperature to some quantity such as height, density, or optical depth. The only unique "temperature of the Sun" that can be assigned is the effective temperature, about 5,800° Kelvin (10,000° F). This is the temperature at which a black body, or perfect radiator, would emit the same total radiant energy. The temperature of the deepest "visible" region (unit optical depth) is a few hundred degrees greater; about 300 kilometres (200 miles) higher, the temperature of the Sun's disk has its smallest value, about 4,600° K (7,800° F). (This cool region is still occasionally called the "reversing layer," from the obsolete picture of a "photosphere" in which the continuum is produced and a cooler layer above it in which dark—or reversed—lines are formed. The hotter and cooler regions are realities, but the separation of the roles is not.) At still greater heights the temperature rises rapidly. First, in the chromosphere, the temperature rises to 30,000° K (about 54,000° F); farther out, the very extended corona is at about 2,000,000° K (3,600,000° F). The effective temperature of a large sunspot is only about 4,400° K (7,500° F).

The main results of the chemical analysis are as follows: molecules are rare, because of the high temperature; atoms and some positive ions are dominant; nearly all the atoms are hydrogen and helium, in a ratio 14:1; and the remaining 1 percent of the atoms have relative abundances much like those of the Earth and meteorites, except that the three lightest elements after helium (lithium, beryllium, and boron) are much rarer. Some elements are easier to detect and measure in sunspots (low temperature) or in the chromosphere (high temperature). In addition, simple molecules are rather prominent in sunspots.

The rotation of the Sun is most readily determined by following spots at the latitudes at which they occur. The Doppler method works at latitudes far enough from the poles: at the solar equator, one limb (or edge) of the Sun's disk is found to be receding and one approaching, at nearly two kilometres (more than one mile) per second, giving a period of 26.9 days for a complete rotation as seen from the Earth; the period with respect to the stars is shorter by about a day, because the Earth has moved in its orbit around the Sun by 26 degrees in 27 days. At higher solar latitudes the period increases steadily, and at 45° a complete rotation takes 29.5 days.

Close inspection of an image of the Sun reveals that it is covered by bright "granules", separated by a network of slightly darker material. The supposition that the granules represent rising convection cells is supported by the observed radial velocity, about 900 metres per second upward.

The bright lines of the corona were a mystery for many years and were at one time suspected to be caused by an unknown element "coronium." In 1941, a nearly complete explanation was found in terms of iron lacking 9 to 14 electrons, nickel lacking 11 to 15 electrons, and similar ions of calcium and argon. Several other elements have since been included, to account for weaker lines. The energy required to remove as many as 15 electrons from an atom corresponds to a temperature of over 1,000,000° K. The breadth of the lines also implies a high temperature, which produces large line-of-sight random velocities. Many details concerning the properties and behaviour of the outer layers of the Sun that can be obtained from a study of the spectrum at high or low dispersion are described in the article SUN.

Telluric and interstellar absorption lines. Absorptions in the Earth's atmosphere, called telluric absorption lines and bands, are prevalent at nearly all wavelengths except in the blue and near-ultraviolet parts of the spectrum. The features in the solar spectrum called A and B by Fraunhofer are due to atmospheric molecular oxygen, O_2. As described above the atmosphere is opaque below 3,000 Å and through much of the infrared. In the photographic spectral range the principal absorbers, which produce dark bands in the spectrum, are molecular oxygen, O_2, and water, H_2O. In addition there are a number of diffuse bands due to the double molecule, $(O_2)_2$; some are near the O_2 bands, but others are also found at other wavelengths.

It is not always easy to distinguish a telluric line from one coming from the light source, particularly since water has a very complicated spectrum, which is not fully tabulated, and since it is variable in amount. There are two principal tests: telluric lines, unless they are very strong, are narrower than solar and stellar lines; and they vary in strength according to the zenith angle of the source, being much stronger when the source is near the horizon. A third test can be applied if the radial velocity of the source varies with time; the telluric lines are of course "stationary"; that is, they do not share the velocity of the source.

So-called stationary lines

There is another class of stationary line, formed in the interstellar medium. In the spectra of bright distant stars such lines are particularly prominent because their

strength depends on the number of the interstellar particles that produce them between the observer and the source, especially those that are reddened by intervening clouds of interstellar dust: these clouds contain gas as well. Sometimes several lines are seen that correspond to different radial velocities, implying that two or more clouds are moving with distinct speeds along the path. Such patterns are found in the two sodium D lines (5,890, 5,896 Å) and the H and K lines of the calcium ion (3,933, 3,968 Å) as would be expected in clouds containing both sodium and calcium. (Here D, H, and K are the Fraunhofer designations, still in use.) Any other explanation of the correspondence is difficult to imagine. Lines are also known that are produced by simple molecules—CH, CH^+, and CN. The molecular bands are much simpler than those produced at normal temperatures, showing only two or three lines instead of the normal 20 or 30. The temperature of the molecules seems to be in equilibrium with the universal 3° K (−454° F) background radiation that has recently been discovered; and the line intensities for the interstellar compound cyanogen have been used to verify this temperature.

Other absorptions in the ultraviolet that are attributed to interstellar gas have been detected by rocket-borne spectrographs and by the spectrometers aboard the first "Orbiting Astronomical Observatory." Hydrogen, the most abundant atom in nature, is prominent and produces an enormously strong line (called Lyman alpha) at 1,216 Å. Molecular hydrogen has been detected in the difficult-to-observe region 1,000–1,120 Å.

Spectroscopy at radio wavelengths has been used with regard to a number of molecules, in addition to the famous 21-centimetre emission line of atomic hydrogen. This subject is discussed below in a separate section. (See also INTERSTELLAR MEDIUM.)

Stellar spectroscopy. Spectroscopy of faint objects, such as even the brightest stars, poses problems of a quite different nature from those associated with the Sun.

Stellar spectroscopy necessarily must be concerned with the problem of the concentration of radiation, particularly of light. The light from the brightest stars falls on the Earth with an intensity 10^{10} times less than that of sunlight. Since any possible increase in light-gathering power is limited to a few powers of ten, improvements in stellar spectroscopy can be obtained mainly by an increase in exposure time, by a decrease in the dispersion, and by efficiency of design in the spectrograph. The exposure time must be increased to minutes or hours, the length of the spectrum cut down to a few centimetres (at high dispersion) or less (at medium or low), and the sensitivity of the receiver, usually a photocell or photographic plate, made as great as possible. The three examples of classes of spectral dispersion—high, about 2 angstroms per millimetre; medium, about 40–100 angstroms per millimetre; low, from about 200 angstroms per millimetre—can be compared with the dispersion possible in solar spectroscopy, in which, because of the enormous amount of light available, as has been noted above, several centimetres may be used to record the lines in a single angstrom.

Classification of stellar spectra. Spectral classification of stars was initiated in the middle of the 19th century by the astronomer Angelo Secchi in Italy.

At the beginning of the 20th century, a group at Harvard undertook the classification of several hundred thousand stars over the entire sky, an effort that culminated in the *Henry Draper Catalogue* published 1918–24. It provides the basis for all modern astrophysical work.

The system used in the catalog has been refined many times, by means of physical theory and empirical methods. Some terminology has survived from the days of pure empiricism, in particular, "early" and "late" spectral class for "hot" and "cool" stars. Most stars can be placed in the sequence O-B-A-F-G-K-M which is controlled by the temperature of the surface. The sequence is further subdivided by adding the figures 0 to 9 to the letters; *e.g.*, the type B9 precedes type A0. The range is from about 40,000° K (O type, 71,500° F) to 2,000° K (3,100° F) or even 1,500° K (M type); the Sun is intermediate—type G1 at 5,800° K (10,000° F). Parallel sequences exist at the low-temperature, or late, end; stars of class C (formerly class R–N) are rich in carbon, and class S stars are rich in metals like zirconium. The effect on the spectrum of differences in composition is less marked for hotter stars, but such differences are readily resolvable on medium- and high-dispersion spectrograms. Much effort has also been expended in luminosity (or brightness) classification, because the distance of a star can immediately be found if its absolute brightness is known (see STAR). For two stars of equal temperature, the luminosity is proportional to the surface area, or to the square of the radius. Thus, luminous stars are called giants and less-luminous ones are called dwarfs. The pressure at the visible surface of a giant is relatively small, and spectral information about pressure, therefore, indicates the size of the star. Surface pressure may be indicated by the width of hydrogen lines (broad for dwarfs, narrow for giants), or by a comparison between the line strengths or intensities from neutral atoms and those from ions (neutrals favoured for dwarfs, ions for giants). Such work was at first done with high-dispersion spectra and extended to medium dispersion in the 1940s.

By courtesy of Hale Observatories

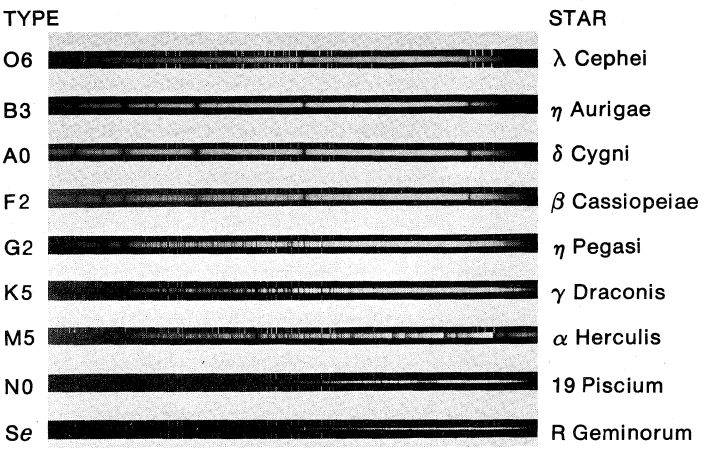


Figure 2: Principal types of stellar spectra, flanked by comparison lines from a laboratory source. Hot stars, at the top, are brightest in the blue; and cool ones, at the bottom, are red. The spectrum at the bottom contains emission lines, denoted by the *e* attached to the spectral class.

The effect of temperature on a stellar spectrum was first quantitatively explained by the Indian physicist M.N. Saha in 1920. At the low temperatures of M-stars, many simple molecules are present, and their bands dominate the spectrum. Increasing thermal agitation at higher and higher temperatures has several effects: molecules dissociate into atoms; many atoms are maintained in excited states in which they can absorb new lines; and atoms are ionized by the loss of an electron. The solar spectrum (class G) illustrates an intermediate temperature stage in which neutral atoms and ions (such as those of calcium and iron) both are seen. Hydrogen lines are moderately strong: they originate from the first excited state of the atom. At Class A, these hydrogen lines are very strong, but most other atoms are ionized and show only a few weak lines. In Class B, even hydrogen begins to be ionized, and helium lines are found; the hotter O-stars show ionized helium.

Many stars of Class F and hotter have broad, dish-shaped lines, though the areas (that is, the total amounts of absorption) are normal. The spread of the line across the spectrum has been shown to be due to rapid rotation of the star; the resulting line shape is characteristic of the rotation speed and from the shape it is found that speeds of rotation at the equator may be as much as 400 kilometres per second (250 miles per second) or more. Such speed is almost enough to throw material out from the equator. Rapid rotation is much less common in cooler stars. It is surmised that most stars are created with high angular momentum from the motions of the collapsing nebula that forms them. For the cooler stars, some brak-

ing mechanism must operate to slow them to velocities as modest as that of the Sun.

Measurement of radial velocities

Velocities in the line of sight have been determined for many stars from the Doppler effect, which shifts all lines by an amount proportional to wavelength. To make such a determination, a spectrum of medium to high dispersion is taken, with a comparison spectrum from a laboratory source on each side. Many precautions are necessary for accurate results; with high dispersion and sharp lines the probable error can be as small as 0.1 km/sec (0.06 mi/sec), but more often, it is 1–2 km/sec (0.6–1.2 mi/sec). A photoelectric method has been developed in which a mask resembling a negative of the stellar spectrum is placed over it. Wherever the transparent lines of the mask are in register with the dark lines of the star, the transmitted light is weakest. This minimum light can be detected by a photomultiplier; the corresponding position of the mask gives the radial velocity in terms of the measured shift from its expected position. This method is much faster than photography, but a given mask can be used only for a narrow range of spectral class. All radial velocities must be corrected for the motion of the observer that results from the rotation and orbital motion of the Earth.

Among the results obtained with the aid of radial velocities are the following: (1) the determination of the velocity (about 20 km/sec or 12 mi/sec) and the direction in space of the motion of the solar system with respect to the nearer stars; (2) the estimation of the random motions of the stars, which for the massive B-stars averages as little as 15 km/sec (9 mi/sec), increasing to nearly 60 km/sec (37 mi/sec) for typical M-dwarfs; (3) the recognition that the entire Galaxy rotates around a distant centre located in the direction of the constellation Sagittarius, with a velocity of between 200 and 300 km/sec (125 and 200 mi/sec) in the region of the Sun; (4) the comparison of radial velocities and proper motions (the rate of movement across the sky is called proper motion) to determine mean distances of groups of stars of any one kind; (5) the discovery of a group of high-velocity stars which are moving with respect to the Sun with speeds of the order of 100 km/sec (62 mi/sec) and in a direction opposite to the normal direction of galactic rotation; (6) the measurement of red shifts in very dense stars (white dwarfs), which are probably the gravitational red shifts predicted by Einstein (see GRAVITATION); (7) the law of linear increase with distance in the velocity of recession of external galaxies; (8) the discovery of numerous stars showing periodic variations in radial velocity—resulting from orbital motions of close double stars or from pulsations in the radii of individual stars (such as Cepheid variables); (9) the study of eruptions in novae and in other expanding shells of gas which surround such stars as P Cygni; (10) the discovery of stationary lines, which do not show the Doppler shifts characteristic of the stars but are interpreted as absorption by clouds of interstellar gas.

Stellar temperatures. The temperature of a stellar "surface," like that of the Sun, is not a well-defined quantity, because the atmosphere is not isothermal (at the same temperature everywhere) and different wavelengths originate at different levels. In a spectral region for which the continuous opacity changes slowly with wavelength, however, the spectrum resembles that of a perfect radiator (black body) at the temperature of the layer at unit optical depth. This condition obtains both for stars that are cooler and for those that are hotter than the Sun, in which the negative hydrogen ion is the dominant cause of the opacity. Measurement of the continuous spectrum, the continuum, of such stars thus yields realistic temperatures, and the colours of the stars obey the usual rule of going from red to yellow, white, and blue as the temperature increases. In the coolest stars (Class M, also C and S), however, so many lines and molecular bands are present that they dominate the spectrum. As a result, the colours, though very red, do not give a good measure of the temperature. Similar reservations apply in the ultraviolet range for almost all stars.

Colour-temperature relationships

Another measure, called "excitation temperature," can be obtained from the relative intensities of suitable lines of an atom, of an atom and its ion, or even of lines of different atoms such as hydrogen and helium. In essence, such intensities form the basis of spectral classification.

Line broadening. The lines of spectra are not perfectly sharp; indeed, a perfectly sharp absorption line would be nearly unobservable. Several broadening mechanisms exist, and observations of the actual shapes of lines (density plotted against wavelength—the line profiles) can be used to infer something about physical conditions in the star. (1) Natural broadening is intrinsic to the atom (or molecule); it is intimately related to the Heisenberg uncertainty principle of quantum mechanics and is usually a very small quantity. (2) Pressure, collisional, or Lorentz broadening is caused by collisions or near-misses with other atoms while the atom in question is radiating or absorbing and becomes more important at high pressures. The shape or profile is similar in the two cases: the opacity falls off as the square of the wavelength difference from the line centre, but at a rate that depends on pressure. (3) Doppler broadening is due to random motions in the line of sight; such motions can be produced in any of several ways. First is thermal agitation; the characteristic speeds are proportional to the square root of the ratio of absolute temperature to atomic or molecular weight and are, therefore, largest for light atoms at high temperature. Mass motions (turbulence) can also be important and are classified according to the optical thickness of the individual moving elements, measured at the centre of the line. Turbulence on a small scale, called microturbulence, for which the elements are thin, affects the spectral lines in the same way as thermal agitation: generally speaking, the lines become strong as well as broad. Turbulent motion on a large scale, called macroturbulence, like stellar rotation, spreads the lines out but does not strengthen them: they become shallow and broad.

Electric and magnetic fields

The shapes and intensities of spectral lines are also affected by the presence of electric and magnetic fields (Stark and Zeeman effects). If a line is produced in the presence of a magnetic field and if the magnetic field is uniform, the line is split into several components; this phenomenon is observed in the spectra of so-called magnetic stars and in sunspots (Zeeman effect). The Zeeman effect produces a broadened line if the spectral purity is not good enough to show the separate components. Strong, uniform electric fields are not present in most astronomical objects, for they are short-circuited by the flow of electrons and ions. Stark broadening, caused by local electric fields around abundant ions, is important, however, especially for hydrogen; collisions with ions are particularly effective in perturbing this atom.

Unusual stellar spectra. A variety of stellar objects show emission lines along with, or instead of, the usual absorptions. According to the general principles discussed above, the regions of the star's atmosphere where these lines are formed are optically thin or, if opaque, must be hottest near the surface. In the most commonly observed case, a star is surrounded by an extended envelope or shell of gas. If this envelope can be observed directly, it is called a nebula.

Spectra of novae

Not all the processes operating in the formation of stellar spectra, especially the unusual or peculiar ones, are understood, but several can be mentioned here. (1) Rapidly rotating stars, most of which are hot, can throw off gas from their equatorial regions. (2) Processes resembling the emission of the solar wind seem to operate in cool stars, especially giants; here the driving force is probably energy from the violent convection in the outer part of the star. (3) Extremely hot stars may be able to evaporate gas at a substantial rate. (4) Finally, a rather common configuration is that of a close pair of stars. Particularly if one of them has evolved into a giant, a star with an extremely large radius, gas can spill out towards the other star or go into orbit about both of them. Novae, or exploding stars, seem to be of this sort: the gas accumulates on the small, hot star, and is periodically blown off to form an expanding shell. This ex-

pansion may be measured by the Doppler effect. At first the shell is optically thick, and its spectrum resembles that of a star but it is abnormally bright because of its large area. In time bright lines appear, as the gas becomes optically thin; finally, "forbidden lines" characteristic of very low densities are seen (the term "forbidden" should be interpreted as meaning "highly improbable"). All these details are noted by careful observations and measurement of details of the spectrum and, in some cases, of their charges.

The rather small class of stars with unusual spectral characteristics called Wolf-Rayet stars are extremely hot; they show emission lines of several-times ionized elements and clear evidence of expanding shells. Many are known to be close binaries. Even normal O- and B-stars, when observed in the far ultraviolet from space, are seen to have expanding shells; the behaviour of the spectral lines resembles that of the P Cygni stars, another class of peculiar stars with expanding shells of gas. Another interesting class contains the cooler T Tauri stars, which are probably accreting material from an associated nebula, and may be in the last stages of formation. (The P Cygni and T Tauri classes are named after the prototype stars).

White dwarfs are stars that are believed to have used up all their nuclear fuel. They concentrate a mass comparable to that of the Sun into a volume comparable to that of the Earth; hence the name dwarf. The colour of their radiation is white. As would be expected, the surface pressure is high and the few spectral lines observed are greatly smeared. White dwarfs cannot have a much greater mass than the Sun; more massive stars must presumably shed the excess. It is thought that planetary nebulae, discussed below, represent the mass ejected in this way. Other stars, perhaps more massive originally, are believed to eject their outer layers more violently, while the core implodes into a neutron star, releasing a vast amount of gravitational energy. This train of events is identified with the rare and unpredictable phenomenon of a supernova. The spectrum of the early stages of a supernova is almost impossible to record. Spectroscopic study of supernovae in nearby galaxies does suggest that very violent events take place but the spectra are difficult to decipher. There is little doubt that the Crab Nebula, discussed below, is the product of a supernova explosion observed by Chinese and Japanese astronomers in 1054.

Planetary spectroscopy. The solar radiation reflected from a planet contains information of two kinds about the planet. (1) In addition to the normal solar spectrum, this radiation produces atomic lines (which are rare) and molecular bands characteristic of the constituents of the atmosphere during the passage of the light through it. (2) Colorations and broad absorptions are characteristic of solid surfaces or cloud material from which the light is reflected; the interpretation of such data is seldom unique, but with care, useful conclusions can be obtained.

Spectral features added by planetary material

Atmospheric absorption lines are usually narrow and can be observed best with high spectral resolution. With the development of *coudé* spectrographs of long focus, with which high-dispersion spectra could be obtained, and infrared-sensitive emulsions in 1933, astronomers succeeded in observing carbon dioxide on Venus and complex absorptions on Jupiter that were later shown to be due to ammonia and methane. Methane is also found on Saturn, Uranus, and Neptune. The lack of ammonia absorption (though there are some reports of it for Saturn) is reasonable at the low temperatures of the outer planets; and the clouds of Jupiter and Saturn are probably ammonia ice.

Observations in the infrared and ultraviolet wavelength regions. Other substances were sought in planetary atmospheres without success, and progress had to await extension of the observable wavelength band further into the infrared region. In 1947 use of a lead-sulfide detector extended the observable infrared out to 2.5 microns, enough to detect carbon dioxide on Mars. The spectral resolution was not high, but the carbon dioxide bands in this region are extremely strong. By 1963, improvements in infrared photography and in *coudé* spectrographs produced a single plate showing a weak band of carbon dioxide at 8,689 Å. The actual purpose of the exposure was to search for evidence of water vapour on Mars at a time of large planetary motion in the line of sight and, therefore, large Doppler shift, which would separate the contributions to the water vapour bands from the Earth's atmosphere and that of the planet; weak absorptions were indeed found. It soon became clear, however, that the atmosphere of Mars is nearly pure carbon dioxide; the only other gases likely to be present in substantial amounts are nitrogen and argon, which are invisible to this type of spectroscopy.

Still another advance, the perfection of high-resolution Fourier spectroscopy, permitted the detection of a trace of carbon monoxide in spectra of Mars obtained by Pierre and Janine Connes, two French astronomers. The resolution and consequent purity of this method is such that many new identifications of features, hitherto hopelessly blended with neighbouring ones, are now possible. Oxygen absorption has been suggested, but may not be real.

Space observations suggest a trace of ozone absorption around 2,500 Å. The emission spectrum of the upper atmosphere of Mars (the daytime airglow) was observed by spectrometers aboard United States flyby spacecraft in 1969. The presence of carbon monoxide and carbon dioxide were confirmed, and traces of hydrogen, oxygen, and carbon atoms, presumably the result of dissociation by solar ultraviolet, were found. No emissions of nitrogen were seen, in striking contrast to the dayglow of the Earth.

Observations of small planetary areas. In 1969, the carbon dioxide (CO_2) absorption and, since it must be proportional, the pressure of the atmosphere of Mars at the surface, above a large number of small areas of Mars were measured. On the assumption that higher ground levels had less atmosphere above them, these data were then converted to a topographic map showing large-scale relief of more than 10 kilometres (6 miles). A similiar method was applied to data from spacecraft flying past Mars in 1969. Both sets of data agree, in most respects, with topography deduced from Earth-based radar.

Venus

Quantitative analysis of the Venus atmosphere is hampered by the ubiquitous haze. The path taken by the radiation as it is scattered from point to point in the atmosphere is unknown, but is presumably similar at nearby wavelengths; thus, the relative abundances of various gases can be obtained. The most abundant gas is known to be carbon dioxide again, from direct chemical analysis by three Soviet Union entry probes. With this knowledge, the amounts of the remaining gases can be obtained. One hundred and ninety bands of carbon dioxide have been identified in the Connes spectra. Rare isotopes of carbon and oxygen also have been identified from peaks in the spectra that are shifted from the positions of the main carbon dioxide lines because of the mass differences between the isotopes. This technique has been possible for the first time because of the very large improvement in the resolution provided by the Connes spectra. The isotopic abundances are found to be the same as on Earth. Carbon monoxide is also present, in three isotopic forms. A most unexpected result was the detection of trace amounts of hydrogen chloride (two isotopic forms) and hydrogen fluoride. Oxygen absorption has not been detected. Water vapour is definitely present, but the amount is variable and probably too small to permit the cloud particles to be ice. Much of the work on water vapour has been done from high-altitude platforms to avoid the telluric absorption. One United States scientist has worked with unmanned balloons, and another has used a jet aircraft. Ground-based observations are limited to dry periods at times of large Doppler shift. Far-ultraviolet emission by hydrogen and oxygen atoms has been observed by means of a rocket-borne telescope and spectrometer; the hydrogen has been observed in detail from Soviet Union and United States spacecraft.

The outer planets

It has long been realized that the low density of Jupiter (and also of the other three giant planets, Saturn, Uranus, and Neptune) implies the presence of large amounts of

hydrogen, and the detection of methane (CH_4) and ammonia (NH_3) offered support for this reasoning. But direct detection of H_2 became possible only in 1949, when laboratory work showed the existence of two classes of infrared absorption that could be sought by direct observation. Broad pressure-induced lines were soon found in the spectra of Uranus and Neptune. A decade later, the narrow quadruple lines of hydrogen produced in relatively rare transitions were observed on Jupiter, and they have been studied in detail since; the expected dominance of hydrogen is confirmed. In 1963, use of a large balloon-borne telescope made possible the observation of very strong pressure-induced absorption on Jupiter in the 2.4-micron region, with similar results. Spectroscopy from rockets has also demonstrated the presence of hydrogen atoms in the upper atmosphere of Jupiter.

Spectroscopy of comets. Special difficulties in comet observation include the sporadic nature of cometary appearances; also, as all cometary radiations are either reflected or induced by the effect of sunlight, changes occur constantly in the spectrum as the comet approaches and recedes from the Sun. In the past, most comet spectra have been made at low dispersion, partly because a short exposure or the use of a slow or small telescope was necessary. Such spectra show the gross features well, but fine details can best be studied only on the rare occasions when a really bright comet can be observed at high dispersion.

The nature of comets has been revealed in some detail by spectroscopy. A nucleus, heated by the Sun, gives off a cloud of gases, many of which are unstable free radicals. Ions are also produced and are blown radially away from the Sun by the solar wind. Spectra of the head, the brightest and densest part of the comet, show the presence of C_2, C_3, CH, CN, NH, NH_2, OH, H, and O; sometimes, near the Sun, sodium and iron lines also appear. The tail ions are CO_2^+, CO^+, and N_2^+; the presence of dust is revealed by the solar spectrum it scatters. The nucleus is composed of compounds like H_2O, NH_3, and CH_4 along with heavier hydrocarbons and dust. These molecules are released by the heat of the Sun and dissociated and ionized mainly by solar ultraviolet radiation. The detection of hydrogen atoms was accomplished only in 1970 by two spacecraft, the first "Orbiting Astronomical Observatory" and an "Orbiting Geophysical Observatory." The hydrogen cloud is enormous, even on the scale of the visible comet, the tail of which can be millions of kilometres long.

Detailed study of high-dispersion spectra reveals additional information. Motions are found from Doppler shifts, and intensities of molecular bands give "temperatures," which relate to the equilibrium of the molecules of the tenuous gas with solar radiation. (See also COMETS.)

Spectroscopy of meteors. Though meteorites can be recovered and analyzed in the laboratory, there is no assurance of a similar composition for the much smaller objects, the meteors, that vaporize high in the atmosphere. Meteor spectroscopy is worthwhile not only for chemical analysis, but also for clues to the physical conditions in the meteor trail. Meteors form short-lived bright tails, lasting for a few seconds; they appear unpredictably in any part of the sky, but more frequently during meteor showers. The only practical method is to use an objective prism or, more often, a grating with a wide-field camera. Long exposures at random or during meteor showers have yielded a total of some 1,500 spectra, most at very low dispersion, enough for a detailed study. The chemical elements found do indeed suggest a composition similar to the larger meteorites; they include iron, magnesium, sodium, calcium, manganese, chromium, aluminium, nickel, hydrogen, and silicon, some of them ionized. A curious fact is the appearance of the auroral green line of atomic oxygen in some meteor spectra. Its presence is not fully understood, but presumably it is due to the excitation of atmospheric oxygen. Bands and lines of nitrogen are also seen. One existing spectrum shows, in addition to the usual metallic lines, many molecular bands suggesting cometary material. Apparently present are the oxides of iron, magnesium, calcium, and carbon, as well as CN and C_2. For the other lines, the degree of excitation depends on the velocity and consequent heating of the meteoric particle: slow ones show temperatures around 1,700° Kelvin (2,600° F), and faster ones may reach 3,200° Kelvin (5,300° F). (See also METEORS.)

Analyzing meteor spectra

Spectroscopy of nebulae. The word nebula in Latin means a cloud, and has been applied in the past to many kinds of cloudy objects. Current usage favours restriction of the term to literal clouds of gas, dust, or both; aggregations of stars are called clusters or galaxies, depending on their size and structure. Usually a distinction can readily be made with the spectroscope: clusters and galaxies show a stellar spectrum, whereas gaseous nebulae show a bright-line spectrum. Borderline cases exist; some galaxies have emission lines, and dust clouds near a suitable star (effective temperature 20,000° Kelvin or cooler) can behave as reflection nebulae showing essentially the same spectrum as the exciting star. A well-known example is found in the Pleiades star cluster.

Spectra of gaseous irregular nebulae. Gaseous nebulae are usually excited by intense ultraviolet light from one or more hot stars in the neighbourhood. A very hot star emits huge amounts of far-ultraviolet radiation—the maximum of the black-body spectrum is at 1,000 Å for a temperature of 29,000° K (about 52,000° F). All the gas near it is, therefore, ionized and constitutes an "HII region." The term HII, or H^+, refers to ionized hydrogen. The outer boundary of an HII region is rather sharp; here the ionizing radiation is quickly used up, and the hydrogen in the surrounding HI region is mostly neutral. If the gas is uniform around the star, the HII region is called a Strömgren sphere after the Danish astronomer Bengt Strömgren, who first studied their structure in detail. The observed nebula may appear irregular because of patchy absorption by dust clouds, or because the density of hydrogen varies in different directions.

Emission processes

The emission lines of an HII region, or emission nebula, are produced by two main processes. Electrons and positive ions are continually recombining; in particular, electrons and protons recombine into hydrogen atoms. The result is a continual production of excited atoms and of ions in lower stages of ionization; a rich recombination spectrum results. Because hydrogen is very abundant, its lines are particularly prominent. In addition to the familiar lines in the visible and near ultraviolet ranges, there are many more in the radio region. Recombination also produces continuous spectra extending over several hundred angstroms in various spectral regions, and through much of the radio region. These spectra are produced by electrons approaching ions with considerable energy and combining with them into a specific bound energy level. Thus, the width of each continuum gives, in principle, a direct measure of the energy distribution of the electrons. In practice this information is difficult to obtain because of uncertainties in photometry over the required wavelength range.

Subsequently, the electrons are ejected again, having absorbed energy from the radiation of the central star, and most of them have considerable kinetic energy. This energy is quickly shared with the other electrons and ions, which typically attain a temperature of 10,000° K (17,500° F). The more energetic electrons can collisionally bring the atoms and ions of the medium to their lowest excited states. Many of the lines in such an electron-impact spectrum are unobservable in the laboratory; they were for many years half-seriously ascribed to an unknown element "nebulium." But in 1927 the U.S. physicist and astronomer Ira S. Bowen found that they were due to forbidden transitions in the ions of common elements, especially oxygen and nitrogen. In the laboratory the excited states are unable to radiate before they are quenched by a collision with another atom or the wall of the vessel. But in the tenuous environment of a nebula such collisions are so rare as to be unimportant. The excited states in question are of low energy and are very efficiently populated by the electrons in the nebula; indeed, the production of forbidden lines is the process that limits the temperature of the electrons. In 1925 the green

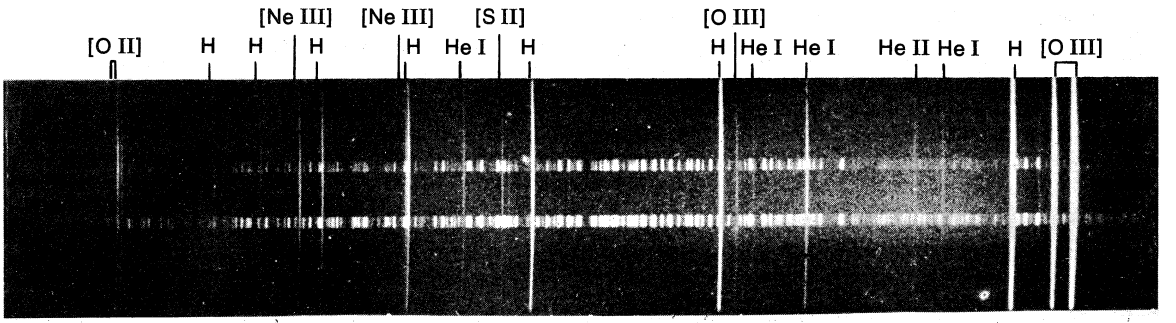


Figure 3: The emission spectrum of the Orion Nebula from 3700 to 5100 Å. showing the Balmer lines of Hydrogen, due to recombination, and the forbidden lines of ionized oxygen, sulfur and neon, due to electron impact.

By courtesy of Hale Observatories; from C. Payne-Gapocshkin and K. L. Haramundanis, *Introduction to Astronomy* (© 1970); Prentice-Hall

auroral line had been similarly identified as a forbidden transition in neutral oxygen, and several other forbidden transitions are now known in the spectrum of the Earth's upper atmosphere.

Bowen fluorescence

A third excitation process was worked out by Bowen in 1934: it is known as Bowen fluorescence. The strongest line of the helium ion, at 304 Å, is emitted by the nebula, and this wavelength happens to coincide with a line of doubly ionized oxygen (OIII). Ions raised to this state may cascade down, through several intermediate energy levels, emitting the observed lines. The last photon radiated can initiate a similar process in nitrogen ions (NIII).

The best-known nebula, visible to the naked eye as the middle "star" in the sword, is the Orion Nebula. Well over 200 lines are identified in its spectrum, representing a dozen of the common elements in several stages of ionization. Several lines are known in the radio region. The composition is similar to that of the Sun. A complicated field of radial velocities is found. The general expansion is about 12 km/sec (7 mi/sec) relative to the central stars. Expanding bubbles are shown in many places by doubled lines—both the front and the back are seen, one approaching and the other receding. It seems that the whole region is in rapid turbulent motion, with discontinuities suggesting shock fronts. Similar data are available for one or two other diffuse nebulae, but in most cases only a general velocity of approach or recession can be obtained.

HII regions can readily be observed in nearby galaxies; in more distant ones, the brightest lines sometimes show above the stellar continuum. The forbidden doublet (a spectral line with two close components) of O^+ at 3,727 Å is found this way in about half the galaxies observed, and even more frequently in spiral and irregular galaxies that tend to be rich in gas. (See also NEBULA.)

Spectra of planetary nebulae. Somewhat over a thousand planetary nebulae are known. The term planetary has nothing to do with the nature of the objects: it refers to the telescopic appearance of the brighter ones—a greenish disk resembling Uranus or Neptune. Photographs usually reveal considerable structure in addition to the basic circular shape. Usually a central star can be seen. Spectral lines near the centre are double and indicate expansion velocities of 10–30 km/sec (6–18 mi/sec). Comparison of this velocity with a typical diameter suggests a lifetime of only 30,000 years. It is generally believed that such nebulae represent the outer envelope of a very old star, whose inner parts are in the process of collapsing into a star of very small diameter called a white dwarf. It is not surprising that this stellar core has an exceedingly high temperature, typically 50,000° K (90,000° F) and sometimes as much as 100,000° K (180,000° F). Though such an object would be expected to be deficient in hydrogen (which is consumed by the processes that cause a star to shine), the abundances of it and of the other chemical elements in both nebulae and central stars seem to be normal. Perhaps the hydrogen is not used up in the outer parts of the star; or perhaps the picture of the central star as a white dwarf is in error.

It is uncertain whether the outer boundary of a planetary nebulae represents the boundary of the ejected gas or a Strömgren sphere embedded in this gas. It is difficult to tell directly from the spectra, because any surrounding HI region would be very faint. Indirect evidence suggests that the smaller nebulae, which are also the youngest and densest, may have HI regions around them. Stratification may be seen directly for ions like He^+ and Ne^{4+} which tend to be found closer to the centre than the less-ionized species.

Supernova remnants. The third principal species of nebula is the supernova remnant. The most famous and best-understood example is the Crab Nebula; the supernova associated with it was observed by Chinese and Japanese astronomers in 1054. This remarkable nebula still presents major new insights; in 1968 the supernova remnant was found to be a pulsar (see PULSAR).

Explanation of supernova

It is thought that a supernova represents the explosion of a massive star, during which a core, about the mass of the Sun, is imploded into a neutron star about 12 kilometres (7 miles) in radius. Such a collapse releases an enormous amount of gravitational energy, comparable to the nuclear energy radiated during the lifetime of the normal star; and a large fraction is stored in the rapid rotation of the neutron star. At the same time, the stellar magnetic field is enormously compressed and strengthened. As the star rotates, its magnetic field acts as a powerful dynamo, accelerating ions, and especially electrons, to high velocities: 900 years after the explosion, the rotation rate of the whole star is still 30 revolutions per second, very close to that of a standard electric motor.

Spectroscopically, the star appears blue but shows no absorptions. Photometry and time-resolved photography show that the light is emitted in flashes, twice per revolution, as in a rotating beacon. The source of the light is not understood; no other pulsar is detectable optically, but they are all older than the one in the Crab. The surrounding nebula consists of two parts: an amorphous mass with a continuous spectrum and a high degree of polarization; and a complicated network of ropy filaments with a normal nebular spectrum. In 1953 it was suggested that the continuum is due to synchrotron emission: light given off by electrons moving at almost the speed of light when deflected by a magnetic field. The subsequent observation of polarization confirmed this idea, but the energy source for the electrons was a mystery until the discovery of the pulsar. The synchrotron continuum extends all the way from the radio, through the optical, into the X-ray region. The ultraviolet part of this spectrum seems adequate to explain the excitation of the gas in the filaments. The gas is distinctly deficient in hydrogen and rich in helium.

Expansion velocities

The radial velocities of the filaments have been explored in some detail. There are large irregularities, but the general speed of expansion is about 1,150 km/sec (710 mi/sec). The corresponding transverse expansion can be measured by comparing photographs taken many years apart; the results strongly suggest that the motion has accelerated somewhat since the explosion, presum-

ably under the influence of the energy emitted by the pulsar.

The Crab, though it has received most study, is not a typical supernova remnant. Several others are known; they show much greater expansion velocities, up to 20,000 km/sec (12,000 mi/sec); naturally, they dissipate much faster and are usually observed as rings or hollow spheres. Much of the optical excitation may be from interaction with the interstellar gas at the advancing front.

Spectroscopy of galaxies. One of the most striking of all spectroscopic results was the discovery of the apparent expansion of the universe, by the U.S. astronomer Edwin P. Hubble. He found the remarkably regular rule that distant galaxies appear to recede faster than those that are near our own system: their spectral lines are shifted towards the red. Hubble's Law states that the velocity of recession is proportional to the distance; it has continued to be confirmed by later work. On such faint objects as galaxies, for most of which only low- or medium-dispersion spectra can be obtained even with very long exposure and large telescopes, Doppler shifts can be measured only because they are so large; a typical spectrum on a blue-sensitive plate is only three millimetres long. The nebular spectrograph is used at the prime focus of the telescope to conserve light.

For the brighter galaxies, it is possible to classify the spectra in much the same way as the spectra of individual, relatively close stars and also to obtain some idea of which kinds of star contribute most to the light. Galactic spectra resemble the spectra of the median type of their component stars and can be simulated by combining the spectra of a number of stellar types. The results depend on the wavelength region used: hot stars have a more important effect on the combined spectrum in the blue, and cool ones in the red. A correlation of the spectrum with the appearance of the galaxy exists. For blue wavelengths, elliptical galaxies tend to have predominantly G- or K-type spectra, indicating that most stars in them are somewhat cooler than the Sun. Irregular galaxies have A-type spectra, considerably hotter than the Sun. Spirals are intermediate, the tightly wound ones having cooler spectra than the loosely wound ones.

The presence of gas in galaxies

The presence of gas is indicated by emission lines, especially the nebular doublet (explained above) of ionized oxygen at 3,727 Å, and lines of hydrogen and ionized nitrogen in the red. The correlation with the type of galaxy resembles that for spectral class: prominent in irregulars and improbable in ellipticals. Gas is associated with young, hot stars because copious amounts of gas are necessary for star formation, and because hot stars make the gas more visible. The situation in the nuclei of galaxies appears to be more complicated, but gas is quite common there.

About two dozen of the type called Seyfert galaxies are known, named for the U.S. astronomer, Carl Seyfert, who first studied them in detail; they exhibit starlike, very bright nuclei, with a high-excitation spectrum of broad emission lines. Turbulent motions of several thousand kilometres per second are implied. Somewhat similar are the N galaxies (nuclear galaxies), some of which have considerable redshifts, and that are usually strong radio sources. The relationship of the two groups to each other is obscure, as is the relation of N galaxies to quasars.

Spectroscopy of quasars. Quasi-stellar objects (QSO's), or quasars, are objects that appear to be point sources of light, like stars, but have large redshifts appropriate to distant galaxies. (See QUASI-STELLAR SOURCES.) In their spectra, emission lines are prominent, as is a continuum with absorption lines. Spectra of these objects can only be obtained with very large telescopes because of their extreme faintness.

The nature of the line spectrum was a mystery for several years, in 1963 when a redshift of 0.158 (the increase of wavelength divided by the original wavelength) was found. Within two years a redshift of 2 had been found, and by 1970 one as large as 2.877. In this last case, the far-ultraviolet Lyman-alpha line of hydrogen is shifted from 1,216 Å to 4,714 Å, in the blue. If interpreted as Doppler shifts, redshifts of 2 and 2.877 correspond to velocities of 80 percent and 87.5 percent of the speed of light. Altogether, lines of ten common elements have been identified, and as far as can be determined the composition is like that of the Sun. Relative intensities differ from quasar to quasar, but these differences could be due to different conditions of excitation. Also, since forbidden lines are seen, densities must sometimes be very low in the regions where the spectrum, or part of it, is formed.

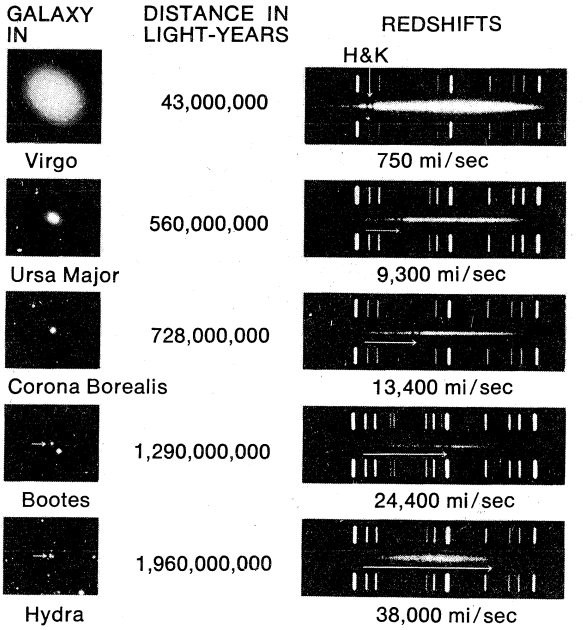


Figure 4: The relation between redshift and distance for five distant galaxies. Redshifts are expressed as velocities, $cd\lambda/\lambda$. Arrows on right indicate shift for calcium lines H and K. One light-year equals about 6 trillion miles, or 6×10^{12} miles.

courtesy of Hale Observatories

Absorption lines are common only in those objects with redshifts approaching or exceeding 2. Most often, the absorption redshift is less than that for the emissions, and in some cases there are two or more redshifts. Unidentified absorptions short of the Lyman-alpha emission are common and may be due to hydrogen at still other shifts, which cannot be verified by correlation with other lines.

Alternative interpretations of redshift

The cause of the redshifts is still undetermined (see QUASI-STELLAR SOURCES). Its interpretation remains one of the fundamental problems of spectroscopy. Among other possibilities, it has been suggested that they are caused by gravitational redshifts, but it is difficult to imagine any so large; also, the narrowness of the spectral lines presents a serious problem because it indicates that they must originate in a uniform region of the gravitational field.

Radio spectroscopy. The most famous spectral line of the radio region is the 21-centimetre (1,420-megahertz—a hertz is a cycle per second) emission of the hydrogen atom. There are two reasons for its importance: hydrogen is the most abundant constituent of the interstellar gas; and the Galaxy is transparent in this wavelength region. Moreover, Doppler shifts and the corresponding radial velocities can be found by a procedure as simple as tuning a radio receiver. The existence of the 21-centimetre line was predicted during World War II and was detected almost simultaneously by three different groups a few years later.

The states involved in the 21-centimetre transition are the so-called hyperfine levels of the ground state. In a hydrogen atom, the magnetic moments of the electron and proton can point either parallel or antiparallel; the latter situation has the lower energy, just as it would for

a pair of bar magnets. But the number of electrons with higher energy—that is, at the upper level—is kept up by thermal collisions; even in the extremely tenuous gases of the interstellar medium such collisions occur once every 400 years. On the average, a radio-frequency photon is emitted by an undisturbed, excited atom every 11,000,000 years: the transition is very infrequent; that is, it is highly "forbidden." Nevertheless, in practice, the line is readily detected in any direction, because there is so much hydrogen and because radio techniques with large antennas are so sensitive.

Mapping the Galaxy

In surveying the Galaxy, the different Doppler shifts of various spiral arms are significant. With the aid of a model of galactic rotation, a detailed map was built up by 1958. Because the Doppler shifts are zero in the directions toward and away from the centre, there are two gaps in the map. The spiral arms can be seen rather well and traced for long distances. In the confused region towards the centre, rapid motions can be discerned; it appears that a ring of gas is moving outwards at about 50 km/sec (30 mi/sec). It was found that about 2 percent of the mass of the Galaxy is in the form of gas. The velocity distribution that was once assumed has been shown to be too simple to fit the real Galaxy. Despite progress in refining the map, a truly satisfactory result has yet to be achieved.

Other galaxies can also be observed, and their general Doppler shifts measured. For the nearest objects, fairly detailed maps have been produced.

Atomic hydrogen also has a rich recombination spectrum in the radio region, as first predicted in 1959. These lines are unobservable in the disturbing environment of the laboratory because they are highly susceptible to atomic collisions. The electronic states of hydrogen are described by a quantum number n, which is 1 for the ground state. Lyman alpha, a well-known ultraviolet hydrogen line, is emitted in a transition from $n = 2$ to $n = 1$; Balmer alpha, or hydrogen alpha ($H\alpha$), in the red, is the 3–2 transition. These lines belong to well-known hydrogen series of lines named after prominent physicists. Radio lines have been observed from 8.2 millimetres (56α, 57–56) to 75 centimetres (253α, 254–253). Jumps of two, three, and four are also found, as are lines of helium and (probably) carbon. The best sources are HII regions, as would be expected for a recombination spectrum. The physical information that can be obtained is similar to that found from optical recombination lines. But the radio lines have the important advantage of needing no corrections for interstellar absorption. A radio continuum is also produced, as in the optical region, by the electrons as they are captured into the high orbits.

Molecular lines are known in the range of wavelengths 2.6 millimetres to 36 centimetres, from ten different molecules; they include water vapour, ammonia, formaldehyde, methyl alcohol (CH_3OH), and some sulfur compounds. For three of them, isotopically substituted forms are also observed. A line at 3.36 millimetres has not been identified. It is common to find emission lines, as well as absorptions against strong sources. The emission lines of OH are notable for their high intensity, small size, and unexpected intensity ratios. The only likely explanation is that stimulated emission is amplifying the radio waves; in other words, an interstellar maser (an acronym for the process: microwave amplification by stimulated emission of radiation) is operating.

BIBLIOGRAPHY. Stellar, nebular, and galactic spectroscopy are thoroughly covered in 5 volumes published in the compendium *Stars and Stellar Systems*, ed. by G.P. KUIPER and B.M. MIDDLEHURST; most relevant volumes are 1, *Telescopes;* 2, *Astronomical Techniques;* 5, *Galactic Structure;* 6, *Stellar Atmospheres;* and 7, *Nebulae and Interstellar Matter*. Articles in the *Annual Review of Astronomy and Astrophysics* are: C. ARPIGNY, "Spectra of Comets and Their Interpretation," 3:351–376 (1965); E.M. BURBIDGE and M. SCHMIDT respectively, "Quasi-Stellar Objects," 5:399–452 (1967) and 7:527–552 (1969); F.J. KERR, "The Large-scale Distribution of Hydrogen in the Galaxy," 7:39–66 (1969); and A.K. DUPREE and L. GOLDBERG, "Radiofrequency Recombination Lines," 8: 231–264 (1970).

(D.M.H.)

Astronomy and Astrophysics

Astronomy (Greek *astron*, "star"; *nomos*, "law") is the scientific study of all objects outside the Earth and its immediate environment, including the Moon, Sun, planets, stars, the Galaxy and similar external star systems, interplanetary and interstellar matter, and the universe as a whole. It must be distinguished from astrology, which has no scientific basis. Astrophysics, the study of physical and chemical properties of astronomical objects, is a relatively recent development of 19th- and 20th-century astronomy. The term astronomy is often used to include astrophysics.

Until the 17th century, astronomy was concerned largely with the measurement of the positions and motions of the Sun, Moon, planets, and apparently fixed stars visible to the unaided eye. Then the laws of planetary motion were discovered, the telescope was invented and applied to astronomy, and the laws underlying motion and gravitation were formulated. In the 18th century the first ideas based on extensive observations of the structure of the Galaxy that contains the Earth and of the universe were put forward. The 19th century brought the introduction of two basic techniques, spectroscopy and photography, which led to new and quantitative methods for measuring the quantity and quality of light and enabled physical studies to be made of brightnesses, temperatures, and chemical nature of stars and nebulae. Theoretical analysis of their constitution followed. Such studies, enormously advanced in the 20th century through the development of quantum theory and other branches of physics, are called astrophysics. Astronomy now flourishes as never before; unexpected objects such as quasars and pulsars have been discovered, and there are hopes that answers to problems of the origins of the universe, of chemical elements, of the Earth, and of life may be found.

Limitations of astronomy. Compared with other experimental sciences, astronomy has certain limitations. First, apart from meteorites, the Moon, and the nearer planets, the objects of study are inaccessible and cannot be manipulated, although nature sometimes provides special conditions, such as eclipses and other temporary effects. The astronomer must usually content himself with studying radiation emitted or reflected from celestial bodies.

Second, from the Earth's surface these are viewed through a thick atmosphere that completely absorbs most radiation except within certain "windows," wavelength regions in which the radiation can pass through the atmosphere relatively freely in the optical, near-infrared, and radio bands of the electromagnetic spectrum; and even in these windows the atmosphere has considerable effects.

Effects of the atmosphere on observation

For light these atmospheric effects are as follows: (1) some absorption that dims the radiation somewhat, even in a clear sky; (2) refraction, which causes slight shifts in direction so that the object appears in a slightly different place; (3) scintillation (twinkling); *i.e.*, fluctuations in brightness of effectively pointlike sources such as stars, fluctuations that are, however, averaged out for objects with larger images, such as planets (the ionosphere, an ionized layer high in the atmosphere, and interplanetary medium have similar effects on radio sources); (4) image movement because of atmospheric turbulence ("bad seeing") spreads the image of a tiny point over an angle of nearly one arc second or more on the celestial sphere (one arc second equals $\frac{1}{3,600}°$); and (5) background light from the night sky. The obscuring effects of the atmosphere and its clouds are reduced by placing observing stations on mountains, preferably in desert regions (*e.g.*, southern California and Chile), and away from city lights. The effects are largely eliminated by observing from high-altitude aircraft, balloons, rockets, space probes, and artificial satellites. From stations outside all or most of the atmosphere, gamma rays and X-rays—that is, high-energy radiation at extremely short wavelengths—and far-ultraviolet and far-infrared radiation, all completely absorbed by the atmosphere at ground-level observatories, can be measured. At radio wavelengths between about one centimetre and 20 metres,

the atmosphere (even when cloudy) has little effect, and man-made radio signals are the chief interference.

Third, the Earth is a spinning, shifting, and wobbling platform. Spin on its axis causes alternation of day and night and an apparent rotation of the celestial sphere with stars moving from east to west. Ground-based telescopes use a mounting that makes it possible to neutralize the rotation of the Earth relative to the stars; with an equatorial mounting driven at a proper speed, the direction of the telescope tube can be kept constant for hours while the Earth turns under the mounting. Large radio telescopes usually have vertical and horizontal axes (alt-azimuth mounting), with their pointing continuously controlled by a computer.

Precession and nutation

In addition to the daily spin, there are much more gradual effects, called precession and nutation. Gravitational action of the Sun and Moon on the Earth's equatorial bulge causes the Earth's axis to precess like that of a top or gyroscope, gradually tracing out a circle on the celestial sphere in about 26,000 years, and also to nutate or wobble slightly in a period of 18.6 years. The Earth's rotation and orbital motion provide the basic standard of directions of stars, so that uncertainties in the rate of these motions can lead to quite small but important uncertainties in measurements of stellar movements. Precession and nutation are not important in driving the telescope.

The Earth's movements until recently also provided the basis of time measurement from a fixed meridian (Universal Time, or Greenwich Mean Time). In the mid-20th century extremely accurate time measurements using quartz-crystal and atomic clocks have revealed that the length of the day fluctuates by about a millisecond because of tides, winds, and seasonal variations. Furthermore, records of ancient eclipses and modern observations of the Moon and planets (defining, over a long period, Ephemeris Time based on Newton's laws of motion) have revealed that the average length of the day is slowly increasing by about two milliseconds per century because of tidal friction. Thus, Universal Time (based on the Earth's daily rotation) does not flow quite uniformly, and periodical corrections must be made to relate it to Ephemeris Time (needed for positions of the Sun, Moon, and planets) and to atomic time (adopted as the international standard in 1967). Also, the surface of the Earth is not exactly fixed relative to its axis, so that any one place undergoes variations in latitude up to about 0.5″ of arc (equivalent to 15 metres [50 feet] on the surface) in a period of 428 days.

This article is divided into the following sections:

I. Astronomy
  Methods of study
  Component disciplines and their relationship to other sciences
  Investigation of the scale of the universe and of the distribution of objects within it
  Orbit theory
II. Astrophysics
  The study of the stars
  The study of galaxies and the universe
III. Trends in modern astronomical investigations

I. Astronomy

Patronage of astronomical investigation

Astronomy has been carried on from the earliest times by amateurs using their spare time and resources and by professionals working in universities and institutions financed by governments or privately (*e.g.*, by charitable foundations). The governmental tradition goes back to antiquity, when priests and other high officials already were engaged in astronomy to fix the seasons and calendar and to study celestial omens. More recently, kings, noblemen, and popes have patronized the subject for its intellectual interest; the Vatican Observatory (founded 1576; refounded 1888) still makes important contributions to astrophysics. From the 17th century, governmental support for astronomy in many countries, justified by the needs of navigation and timekeeping, led to the establishment of national observatories and nautical almanac offices in several countries. In the second half of the 20th century, governmental involvement has greatly increased, partly in line with a general trend toward the promotion of science and partly because of national interest in the exploration of space: government support in some cases takes the form of research contracts to industrial firms. Despite these developments, amateurs using simple equipment still play an important role; *e.g.*, in observations of variable stars, of the Moon and planets, and of occultations (eclipsing) of stars by the Moon and in tracking artificial satellites by optical and radio methods. In some cases, amateurs working together also use ham-radio communications. One of the most famous amateur contributions was the pioneer radio work by Grote Reber in Illinois in detecting radio signals from the Milky Way, work that was fundamental in establishing the science of radio astronomy in the 1940s.

An important part in the development of astronomy has been played since the early 19th century by national and regional astronomical societies, such as the Royal Astronomical Society in London, the American Astronomical Society, and the German Astronomische Gesellschaft, which sponsor meetings and publications and sometimes organize collaboration in research. The British Astronomical Association, the American Association of Variable Star Observers, and other groups all over the world perform a similar function for amateurs. Astronomy also has a strong tradition of international cooperation, motivated partly by the need to combine observations from different parts of the globe and partly by the sheer magnitude of the amount of data to be collected and processed. In the second half of the 20th century also, groups of countries have combined resources to provide such international facilities as the European Southern Observatory and the Inter-American Observatory, both in Chile. Most international cooperation takes place under the auspices of the International Astronomical Union (IAU), founded in 1919. The IAU has members in 45 countries and organizes meetings, publications, and a telegram service for urgent information relating, for example, to appearances of comets, supernovae (violently exploding stars), and other transient phenomena. Other organizations dispense information on the behaviour of the Sun (sunspot numbers, flares), on the fall of meteorites, and on the appearance of meteoric fireballs. Communications may be by telegram, postcard, or any other convenient means.

International co-operation

International cooperation in space exploration is also beginning. Lunar rock samples were distributed to scientists of many countries following the return of Apollo 11 and later lunar landings. Tracking of the earliest Soviet lunar probes was done at Jodrell Bank, England, in cooperation with the Soviets. When Apollo 13 ran into difficulties, prompt offers of tracking help were received from the U.S.S.R.; and many international payloads have gone up with the space laboratories known as Orbiting Astronomical Observatories. In 1971, discussions on docking between Soviet and United States spacecraft were begun.

METHODS OF STUDY

Equipment. Celestial objects are studied through observations (usually involving measurement) or theoretically (mathematically) and occasionally experimentally (*e.g.*, by radar). Apart from special investigations (*e.g.*, of cosmic rays, lunar rocks, and meteorites), astronomical observations involve the study of electromagnetic radiation; *i.e.*, not only light but longer wavelengths in the infrared, microwave, and radio bands of the spectrum and shorter wavelengths in the ultraviolet, X-ray, and gamma-ray bands. Observations usually involve three basic elements: a telescope acting as a camera or as a collector of radiation; a detector such as the eye, a photographic plate, a photomultiplier, radio receiver, ionization chamber; and measuring machines, chart recorders, counters to translate the signal received into numbers. Also there is generally some means of selecting a definite colour of light, or spectral band (*e.g.*, a light filter or tuned amplifier); or radiation may be spread out into a spectrum and recorded or measured at selected places along it.

For some purposes, it is desirable to use telescopes of

Advantages of large telescopes

large diameter or aperture: (1) to collect as much radiation as possible and (2) to improve the fineness of detail that can be distinguished on the sky. The ability of a single telescope to resolve fine detail is limited, especially at radio wavelengths much longer than those of visible light. Better resolution can be provided by interferometers. In an interferometer two or more often widely separated apertures are used together; this arrangement can give the effect of a much larger telescope.

Types of observational information. The main types of information provided by astronomical observation are the following:

Positional information. This allows location and identification of the object and the study of its motions across the line of sight through gradual changes in its position.

Structural information. The Sun, Moon, planets, and nebulae appear as disks, the angular sizes and structural detail of which can be studied through visual observation, photography, or scanning with a telescope and detector. Also, special interferometric techniques enable apparent angular sizes of some stars and radio sources to be measured down to 0.001″ of arc.

Photometric or radiometric information. The amount, or brightness, of radiation in a particular spectral band is measured by comparing it with that received from other objects or from a laboratory standard. In optical astronomy the brightness of the source is usually expressed in magnitudes. Originally (*e.g.*, in Ptolemy's catalog *c.* AD 140), the brightest stars were said to be of the first magnitude, and the faintest visible to the unaided eye of the sixth. Nowadays, magnitudes (including negative and fractional values) are on an exact logarithmic scale such that an increase of 5 in the magnitude represents a decrease in intensity by a factor of 100. There are now many magnitude systems that describe the brightnesses in different colour bands. To relate each magnitude system to the total energy radiated at all wavelengths requires a bolometric (or total-energy) correction, which depends mainly on temperature.

Spectral information. The spectrum of an object shows how its energy is radiated in the form of electromagnetic waves, the different wavelengths being separated for study—*e.g.*, into the different colours of visible light (X-ray, gamma, and other wavelengths of radiation also produce observable spectra). A star radiates over a wide range of wavelengths, so that a stellar spectrum generally has a continuous background, with the wavelength of maximum energy roughly inversely related to temperature, on which narrow, darker absorption features (or sometimes bright emission features) are superposed that correspond to characteristic radiations of gases; gaseous nebulae and radio and X-ray sources have different types of continuous spectra. By studying spectra and comparing them with those of a laboratory source, it is possible to (1) identify elements and compounds; (2) find the velocity of approach or recession in the line of sight (radial velocity); (3) estimate temperatures, densities, and velocity variations over the surface; and (4) find relative amounts of different elements (see ASTRONOMICAL SPECTROSCOPY, PRINCIPLES OF).

Polarimetric information. Electromagnetic radiation can be thought of as waves of varying lengths, all carrying oscillating electric and magnetic fields in a plane at right angles to the direction of travel. After scattering or when magnetic fields are present, the electric field of the radiation may have a preferential direction (plane polarization) or sweep out a circle (circular polarization), in contrast to ordinary radiation, in which its direction varies at random. Polarization measurements, often combined with spectral analysis, thus give information on the scattering (reflecting) properties of planetary surfaces and on magnetic fields in radio sources, the interstellar medium, the Sun, and certain stars.

COMPONENT DISCIPLINES AND THEIR RELATIONSHIP TO OTHER SCIENCES

Planetary sciences. The Earth, accompanied by a relatively large satellite—the Moon—is one of nine known planets, five others of which are also known to have satellites, which along with innumerable minor bodies orbit around the Sun, mostly close to the plane of the Earth's orbit (the ecliptic). Planets and satellites shine by reflecting sunlight, but they also emit infrared and radio waves because their surface temperatures are above absolute zero (0° K; −273° C; −460° F); Jupiter also emits strong radio bursts from a Jovian magnetosphere; *i.e.*, the region of space surrounding the planet and strongly influenced by its magnetic field. Interplanetary space is filled by an extremely tenuous gas of ions (mostly hydrogen nuclei or protons) and electrons, which is called the solar wind, flowing out at high speed from the solar corona; there is also dust near the ecliptic, which scatters sunlight, causing a faint glow in the sky known as the zodiacal light. Other, relatively small bodies orbiting around the Sun are minor planets (asteroids) and comets. The whole collection of the Sun and matter moving under its gravitational control is called the solar system.

Objects in the solar system

Solar-system objects are studied by (1) measurement of changing positions and orbital analysis giving distances and masses (from gravitational action on satellites and other planets); (2) telescopic measurements to find size, polar flattening, surface features, rotation, and optical properties of surface material; (3) spectroscopy revealing molecules in planetary atmospheres (which impose their own absorption lines on reflected sunlight), radial velocities, and rotation; (4) infrared and radio studies for surface temperatures; (5) radar, giving accurate distances and rotational velocities and revealing solid surface features otherwise hidden by clouds; (6) closeup photographs and physical measurements from space probes (*e.g.*, the U.S. Mariner series for Venus and Mars, and the Soviet Luna and U.S. Ranger, Surveyor, and Lunar Orbiter series for the Moon); (7) landings by automatic-instrumented probes, such as the Soviet Venera series for Venus and Luna series and Lunokhod on the Moon and U.S. Ranger and Surveyor landings on the Moon; and (8) manned landings in the U.S. Apollo series of Moon shots (see SOLAR SYSTEM; MOON; and articles on individual planets).

Planetology. Planetology, the study of the solid parts of the planets of the solar system, as geology is that of the Earth, is a relatively new branch of astronomy. Only in the 1960s did sufficient data become available for any planet other than Earth to raise such studies to the level of an independent discipline. Now, information relayed back from space vehicles and obtained by radar is available for Mercury, Venus, Mars, and the Moon. From this information, temperatures and pressures at the base of the atmosphere (Venus) can be found and the relief of the surface plotted; improved values for the radii, periods of revolution, and other characteristics are also available. Many uncertainties remain, and, in this respect, a projected space journey to be called the Grand Tour, possibly during the late 1970s, would provide many more details, particularly about the outer planets, Jupiter, Saturn, Uranus, and Neptune.

Comparisons between Earth sciences and planetary sciences

The planet best known and most studied is, of course, the Earth. Since about 1960 enough information on the nearer planets and the Moon has become available to make possible some detailed comparisons between Earth and these other bodies. Techniques developed and experience gained in the Earth sciences are being brought to bear on studies of the other planets.

The study of planetary atmospheres. The study of planetary atmospheres is concerned with their wind and circulation patterns (governed by the temperature difference from Pole to Equator and the planet's rotation); their temperature and pressure structure (of which something has been learned by observing occasional occultations of stars when a planet such as Venus, Mars, or Jupiter passes in front of them and more by means of space probes to Mars and Venus); and the chemical nature of the atmosphere and its origin; *i.e.*, whether primary (as may be the case with Jupiter) or secondary. Secondary atmospheres are exuded from rocks and afterward modified by solar radiation and, in the case of the Earth, photosynthesis; the atmospheres of the inner or

terrestrial planets seem to be secondary. The smaller, warmer bodies such as Mercury, the Moon, and the minor planets have lost any atmosphere they had, because the average speeds of atmospheric molecules exceed that required to escape from the planet's gravitational attraction. Atmospheres enhance planetary temperatures by the greenhouse effect, whereby the atmosphere like a glass roof lets in sunlight but prevents heat energy created by the sunlight from escaping. This seems to be especially strong on Venus, where carbon dioxide lets in visible sunlight but holds back infrared radiation. Some forms of life could be present on Venus and Mars, but there is no firm evidence for it.

Meteoritics. During its journey around the Sun, the Earth may encounter bodies ranging from tiny specks of dust to minor planets the size of mountains. The smaller bodies, with masses up to about one gram (0.04 ounce), are called meteoroids and are believed to be debris of extinct comets. They are vaporized by friction at about 100 kilometres (60 miles) altitude, leaving behind them shining trails of glowing gas that are seen as meteors (shooting stars or falling stars) and reflect radar waves. About 20 percent of them belong to streams or showers approaching from a definite direction (through perspective effects they appear to diverge from a point of the sky called the radiant) and are encountered annually on orbits that can be identified with those of former comets. Much smaller specks (micrometeorites) are not vaporized and simply fall to the ground. More rarely, larger bodies weighing tens of kilograms come down to about ten kilometres (six miles) height and break up, sometimes with an audible explosion, producing a very bright fireball, or bolide, while still larger bodies crash to the ground as one or more meteorites. These are not associated with meteor streams but come in on orbits suggesting an origin in the belt of minor planets. Round craters up to 40 kilometres (25 miles) across are attributed to falls of meteorites weighing more than 1,000 tons. Special kinds of rounded glassy objects found in certain areas and known as tektites may or may not be of extraterrestrial origin.

The study of meteorites. Meteorites are samples of extraterrestrial material containing vital clues to the chemical composition of the universe and the early history of the solar system. Most are stones or irons. Iron meteorites (alloys of iron with nickel and other elements) may have been formed in the interior of a former planet about 300 kilometres (about 200 miles) in diameter. Stony meteorites are subdivided into two main classes: chondrites (containing droplike inclusions called chondrules) and achondrites (like terrestrial igneous rocks). Chondrites seem to be a fair sample of the raw material of the solar system, in contrast to terrestrial or lunar rocks, which have been more differentiated by melting. Like other pieces of rock or iron, meteorites can be studied using the methods, principles, and techniques of petrology, radioactive dating, and other branches of geophysics and geochemistry. It is of interest to find evidence of shock (shattering), and the texture and composition of a meteorite give clues to its history (see METEORITE).

Petrology and other branches of geosciences as part of meteoritics

The study of meteors. Meteors can be best studied by positional measurements based on both optical and radar observations, by statistics, and, where possible, from their spectra. The last measurements are difficult to obtain because of the sporadic nature of meteors, but a large number of good spectra are now available (see METEOR).

The study of comets and minor planets. Comets are members of the solar system. About 20 per year approach close enough to the Sun to become visible for some months before receding again along highly elongated orbits. Typically, a comet has a small nucleus (not always visible) surrounded by a diffuse coma (Latin: "hair") of dust and gas merging into an immense tail or tails pointing away from the Sun.

Comets, like stars, can be studied for their position and, through their spectra, for their chemistry. Using three position measurements, a preliminary orbit (path in space) can be found; this can then be improved by further observations, if available. Many comets are now known to be periodic; that is, they return and are seen again on roughly the same path a second time or more. The recovery of periodic comets is an important field of cometary study. The origin of comets is rather obscure, but most have such immense periods that they seem to be coming in for the first time. Comets may be occasionally pulled in toward the Sun by gravitational action of passing stars. They could then be frequently captured into the inner part of the solar system by gravitational action of planets, especially Jupiter, whereafter they gradually disintegrate.

The distances of successive planets from the Sun form a fairly regular progression, but until 1800 there was an apparent gap between Mars and Jupiter that is now known to be occupied by a large number of tiny bodies called minor planets, or asteroids, distinguished from comets by their small sharp, starlike images. The study of asteroids occupies astronomers at several large observatories full time. Positions, brightness, and light polarization are all monitored regularly; recovery and identification programs are important, and, because there are now more than 1,700 numbered minor planets, statistical programs are important also (see COMET; PLANET, MINOR).

Lunar sciences. Generally thought of as a satellite of the Earth, the Moon is more like a sister planet because it has a far larger mass (about 1/81 the Earth's mass) in comparison with Earth than other satellites in comparison with their central planet (usually less than 1/1,000). The Moon was observed in the 17th century through one of the first telescopes by Galileo, who named the darker, often circular areas maria ("seas"), although there is no surface water on the Moon.

The name of the Greek moon goddess was Selene. Hence, properly, studies of lunar formations, structure, composition, and so on are called selenography, selenology, and so on. By popular usage, especially in the United States, terms such as astrogeology have become accepted as alternatives.

Lunar photographs and rock samples

Because the Moon is the nearest planetary body in space, it has been the target of many space probes. Many high-quality photographs have been taken from vehicles in orbit around the Moon and by men standing on its surface. Lunar-rock samples have been distributed to hundreds of investigators since 1969, and the laboratory techniques of many disciplines have been brought to bear on them. Many of the traditional beliefs have been swept away in consequence, and lunar science has become a fast-moving field for many scientists other than astronomers. This area can use to real advantage the pooling of many branches of knowledge—chemistry, petrology, geology, and statistics, for example (see MOON).

The amount of new lunar data of the geochemical and geophysical type obtainable from the lunar rocks is immense. Tidally triggered moonquakes have been recorded, and the apparent escapes of gas and dust from within the Moon previously recorded by observers and ignored for years now seem to be real.

Lunar-motion studies

The motion of the Moon is not a simple ellipse around the common Earth–Moon gravity centre, because many factors continuously alter the shape of the orbit slightly; the most important of these are the presence of the Sun and the slight bulges (departures from sphericity) on both the Earth and the Moon. The resulting complexities provided 19th- and early-20th-century astronomer-mathematicians with difficult problems in celestial mechanics. These problems are approached by a method in which a first, very approximate solution is improved upon by taking account successively of as many small differences as seem necessary. Modern work of this kind is done by computer, the work that formerly would have taken a lifetime requiring a few seconds or minutes of computer operation. The results, however, are still approximations; and, though they can now be brought to within seconds and feet of the true (instantaneous) value, they cannot be relied on beyond a certain calculable length of time and, as before, will need continual revision.

The Moon raises tides on the Earth by attracting the nearer parts more strongly than the farther parts (see TIDES). An indirect effect is a very gradual increase in the average lengths of the day and lunar month.

Special types of information may be obtained during times of eclipse. Eclipses of the Moon occur when the Full Moon passes through the Earth's shadow, and eclipses of the Sun when the shadow of the New Moon covers part of the Earth's surface. Eclipses occur only a few times per year, when the Moon (which moves in a plane that is inclined at an angle to that of the Earth's orbit around the Sun) happens to be close enough to the Earth's orbital plane (hence called the ecliptic) at its new or full phase so that all three bodies are in line. By a fortunate coincidence, the Moon has the right size and distance to cover the brilliant photospheric disk of the Sun almost exactly, so that a total solar eclipse, in good weather, may provide many excellent opportunities to obtain data on the outer layers of the Sun, otherwise very difficult to get. These and other eclipse projects are more fully discussed in ECLIPSE, OCCULTATION, AND TRANSIT.

The study of the origin of the solar system. The origin of the solar system is bound up with the origin of stars, galaxies, and the universe itself and thus forms part of cosmogony, the study of the origin and development of the universe, and any theory concerning it must be compatible with all the evidence from the individual members.

Many factors do not fit comfortably into such schemes as are at present under consideration, and there are numerous uncertainties. Many single stars, like the Sun, may have planetary systems, though observational evidence is scanty; sinuous motions of a few nearby stars imply that they have Jupiter-sized companions but with more elliptical orbits.

Evidence from very young, still contracting, variable stars of the T Tauri type (named for their prototype in the constellation Taurus) is thought to be important in understanding the early stages. Theories of the origin of the solar system, however, still contain a large element of speculation.

INVESTIGATION OF THE SCALE OF THE UNIVERSE AND OF THE DISTRIBUTION OF OBJECTS WITHIN IT

The determination of positions. Space stretches indefinitely around the observer in all directions, so that he seems to be in the centre of a sphere—the celestial sphere—on which the stars are seen. Brighter objects have been given names as members of groups called constellations; otherwise they are designated by numbers in catalogs or by coordinates. Three important reference planes are projected as great circles on the celestial sphere; those of the equator (defined by the Earth's Equator), the ecliptic (in which Earth orbits around the Sun), and the Milky Way (or galactic equator), near which most remote stars and clouds of interstellar matter are situated. To each of these planes there corresponds a north and a south pole, along the axis or line perpendicular to that plane. Positional astronomy is concerned with determining the directions of celestial bodies, which can also be looked upon as coordinates on the celestial sphere (similar to latitude and longitude on the Earth's surface).

Reference planes

Most stellar positions are measured on photographs using offsets from "reference" stars observed at the meridian. Radio source positions are found using interferometers or by observing occultations by the Moon. Positions of X-ray sources observed from rockets and satellites raise special problems both of orientation and of achieving fine resolution on the sky because of the problem of relating the rocket position to those of known stars, the X-ray images of which may bear little relation to the familiar optical ones. Space vehicles either roll, bringing different portions of the sky successively within the field of view of instruments on board, or are stabilized by inertial guidance from gyroscopes that provide a platform against which instruments can be moved to sweep around the sky or to slew to a given object through a preset angle; directions are monitored using instruments for measuring magnetic fields, star cameras, and photoelectric star trackers.

Applications in navigation

One of the oldest applications of positional astronomy is in navigation. The sailor, airman, or astronaut measures the altitude (angle) of a celestial object above the horizon with a sextant, thus establishing (with the aid of tables and chronometers checked by radio time signals) one position line that is a small circle on the Earth's surface. Observation of a second object (or of the same one after travelling at known velocity for a known time) defines a second circle intersecting the first. Large ships now carry automatic star trackers and computers giving a continuous record of position, and star-tracking methods are widely used, even in space travel, despite the increasing development of radio beacons. Manned and unmanned space vehicles are continuously tracked from the ground by optical, radio, and radar methods aided by highly sophisticated dead reckoning, with orbital data continuously updated in a computer.

Stars have proper motions—that is, they move across the sky at a slow, individual rate; over years their relative positions gradually change, occasionally by several seconds of arc per year. Stellar motions, studied statistically, are of great value in studying stellar distances and orbits around the centre of the Galaxy. Remote stars have proper motions that are small and hard to measure, but their very smallness can be a useful sign of great distance. Superimposed on the proper motions, which carry the stars uniformly in straight lines, each of the very nearest stars has a small measurable annual oscillation caused by parallax (*i.e.*, by being seen from different points in the Earth's orbit during the year) and sometimes an additional oscillation caused by orbital motion around the centre of gravity of a binary system (two stars in orbit around each other), giving information on stellar masses.

The measurement of distances. Just as distances on the Earth can be related from local measurements (inches and feet or centimetres and metres) through steps of progressively larger dimensions to continental and larger lengths, so have the dimensions of the Earth and planets to be related to the solar system, the nearer stars, the Galaxy, and beyond. At each stage there are yardsticks that can be used to measure distances over a certain range. Among the problems facing astronomers in dealing with the structure of the universe (cosmology) is that of relating the various scales of distance in a way that is reasonably reliable; this is discussed below. A second problem of concern to cosmologists is whether the characteristics of space-time, formerly taken for granted as Euclidean (*i.e.*, conforming to commonsense notions of geometry), are unchanged at the enormous distances of the fainter external galaxies.

Problem of space-time

The measurement of distances within the solar system. The oldest method of measuring distances is that of triangulation—constructing a triangle from a measured base line and two measured angles. The size of the Earth in terms of the current units of distance and the distance of the Moon were thus found with reasonable accuracy already in ancient times. Once the laws of planetary motion were known, it became an easy matter to find relative distances of planets in astronomical units, the astronomical unit being essentially the mean distance between the Earth and the Sun; the problem that then remained was to measure the astronomical unit itself in terms of local units, such as miles or kilometres. This was achieved to an accuracy of about one part in 1,000 by observing suitable planets (*e.g.*, Venus in transit across the Sun and asteroid Eros at its passage within 26,000,000 kilometres [16,000,000 miles] of the Earth in 1931) from widely separated points on the Earth's surface. Similar accuracy was achieved by exploiting two effects of the Earth's orbital motion combined with the finite speed of light—aberration (an annual shift in stellar positions caused by the fact that the Earth's motion tilts the direction from which light is received slightly forward, like raindrops falling on the windshield of a moving car) and an annual variation of up to 60 kilometres per second (134,000

miles per hour) in the radial velocities of suitably placed stars—and by calculating planetary perturbations of the orbits of Eros and certain space probes.

Use of radar for distance measurements

Guidance of planetary space probes demands much higher accuracy, which is provided by radar measurements. Pulses from a powerful transmitter are beamed to a planet, reflected waves are collected by a large radio "dish," and the time delay (light time for the double journey) is very accurately measured. The beaming of continuous waves of well-defined frequency gives an alternative answer (and other information, such as rotation speed) by measuring relative velocities deduced from the change in frequency that is measured. Radar measurements have been carried out on targets including the Moon, the Sun, and the planets out to Saturn. From measurements with Venus since 1961 the astronomical unit has been found to a precision of about 100 kilometres (limited by knowledge of the speed of light) and is close to 149,597,900 kilometres (92,975,699 miles). Time-delay measurements also give a very accurate test of Einstein's general theory of relativity. An analogous experiment with laser beams, an optical analogue of radar, reflected from the Moon is expected to give additional precise information, this time about the Earth also; *e.g.*, on minute shifts of the continents.

The measurement of distances beyond the solar system. Relatively few stars are sufficiently nearby to show a small annual shift in direction relative to more distant stars because of the different positions taken up by the Earth in its orbit around the Sun. Measurements on suitably timed photographs give the annual parallax, defined as the angle subtended at the star by the mean radius of the Earth's orbit. The usual unit of distance is the parsec (pc), the distance corresponding to a parallax of 1″ of arc and equal to 206,265 astronomical units, 3.086×10^{13} kilometres, or 3.26 light-years; one light-year is the distance covered in a year by light travelling at about 300,000 kilometres (186,000 miles) per second. The nearest star (Proxima Centauri) is 1.3 parsecs, or 4.26 light-years, away. Another direct method of finding distances is based on moving clusters of stars, such as the Hyades (40 parsecs away). Methods of determining even greater distances are described in the article PARALLAX, ASTRONOMICAL.

Very large astronomical distances can be found by measuring magnitudes of suitable "beacons" with characteristics that enable them to be identified with nearer objects whose luminosities have been determined from distances found directly. In this way one can fit together overlapping distance scales for progressively more and more distant objects, such as stars, gas in the spiral arms of the Galaxy, distant star systems known as external galaxies, and even the strange objects known as quasars, though there the distances are not universally accepted as reliable. In general, however, considerable confidence can be placed in distance estimates, especially where more than one method is available for comparison of results.

ORBIT THEORY

Mathematical basis of celestial mechanics

The mathematical basis of celestial mechanics is the solution of equations of motion for bodies (planets, stars, galaxies, and so on) moving under the gravitational attraction of others. It was developed in classical form by the English mathematician, physicist, and astronomer Sir Isaac Newton in the 17th century and his successors, starting from the empirical laws of planetary motion that had been discovered by the German astronomer Johannes Kepler. The classical theory is still sufficient for most purposes, but occasionally modifications are required to explain the observations and can be worked out according to Einstein's theory of general relativity.

The two-body problem. The simplest case is that of the two-body problem, in which two effectively point-like masses move in orbits around their mutual centre of mass, or barycentre. Newton showed that the curve traced out is always a conic section: a circle or an ellipse if the relative speed is too small for one body to escape; otherwise a parabola or hyperbola, both of which are smooth, open-ended curves.

Practical considerations. The mutual orbits of two bodies are seldom quite as simple as this because they can be represented as points only if they are symmetrical in all respects, and real planets and other bodies are not. Small corrections must be introduced, and, from the observations of slight differences, planetary characteristics, such as the equatorial bulge of the Earth, can be worked out.

Elements of an orbit

An orbit is defined by six constants, or elements: the period, or major axis, of the ellipse, the eccentricity (degree of departure from a circle), the direction of the major axis in the plane of the orbit, the time of perihelion or periastron passage (when the two bodies are closest together), and two numbers giving the orientation of the orbital plane in space. Elements are constant in the simple two-body case, but they change with time when more masses are present, when one has a slightly nonspherical shape, or when there is a resisting medium. Resistance by the atmosphere and the effect of polar flattening of the Earth change the paths of artificial satellites: polar flattening causes precession of the orbital plane and of the orbit within it, and air resistance causes the satellite to lose energy and height but to gain speed. Such problems cannot be solved, as the two-body case can, in an exact formula; but solutions can be found to a high degree of approximation using modern computers to solve equations of motion by numerical methods. These numerical solutions are valid for only a limited time, say about 1,000,000 years for the planets, because even tiny errors introduced by approximations build up with time; and so it cannot yet be said definitely that the solar system is secularly stable—that is, will remain in much the same form for as long as can be imagined—or, put another way, that planetary-orbital elements will remain the same apart from small oscillations, though it is very likely that this is so.

Another application of orbit theory is to the rotation of galaxies (including the Earth's) in which every star orbits in the collective gravitational field of other stars. Spectral observations of stars and gas clouds in the Earth's and other spiral galaxies show how the speed of rotation varies with distance from the centre and enable the distribution of density and an estimate of the total mass to be deduced. Elliptical galaxies and clusters of galaxies present a more complicated problem, but masses for them can still be found by statistical arguments. Application of present methods to clusters of galaxies gives surprisingly large values of mass compared with the observed luminosity (see GALAXIES, EXTERNAL).

II. Astrophysics

THE STUDY OF THE STARS

Astrophysics, literally the "physics of the stars," is concerned with the processes of physical nature and with evolution in all cosmic objects and is closely related to atomic and nuclear physics, thermodynamics, spectroscopy, plasma physics, solid-state physics, and other fields. Theoretical astrophysics interprets the observations and also uses mathematical methods to solve the equations that describe the physical conditions. It is therefore necessary to study the evidence from stars, interstellar gas, galaxies, and other bodies with the aim of applying laws discovered in the laboratory to the interpretation of the evidence. The most detailed physical evidence from distant bright objects is obtained when the light can be spread out into a spectrum.

The study of stellar structure. Most observable stars are giants (of large radius) or dwarfs, such as the Sun, with average densities of about 0.01 and one gram per cubic centimetre (about that of water), respectively; white dwarfs, such as the companion of Sirius, with about the mass of the Sun, have higher surface temperature but very low luminosity and, correspondingly, a radius like that of the Earth and a density of a ton per cubic inch—about 60 kilograms per cubic centimetre. All these stars are nevertheless gaseous throughout (because atoms in them are stripped of almost all their electrons and are reduced to nuclei taking up very little space) and, indeed, except for white dwarfs, behave like perfect gases. White

dwarfs show a special kind of behaviour (explained by quantum theory) called electron degeneracy, with much higher pressures than in a perfect gas. Neutron stars (probably observed as pulsars) represent a still higher state of compression in which electrons have been crushed onto protons to form a degenerate neutron gas.

The problem of the internal structures of stars

A star is a more or less spherical globe of gas kept in a steady state over long times by the interplay of two opposing forces: gravitational pull of its separate parts on one another and an outward force caused by the increase of pressure inward. Its internal structure is thus a problem in thermodynamics and atomic physics. To maintain the central pressure, the gas must have a high central temperature, about 15,000,000° C (27,000,000° F) for a star with the mass (2×10^{33} grams) and radius (700,000 kilometres [435,000 miles]) of the Sun. Such hot matter necessarily radiates energy, which trickles outward through successive layers of the star and is eventually radiated away as light and heat from the surface. This process leads to a relationship between masses and luminosities of stars (radius having only a minor effect), the luminosity increasing very rapidly with the mass; most stars have masses between 0.1 and ten times that of the Sun, but their luminosities differ over a range of more than 1,000,000.

Radiation causes a steady loss of energy. Degenerate stars can slowly cool down, but ordinary stars cannot, because loss of heat energy leads to contraction and a rise in temperature. Stars must either gradually contract or derive energy from some other source now identified as thermonuclear reactions, causing fusion of light atomic nuclei (mainly hydrogen nuclei or protons) into heavier ones (mainly helium nuclei or α-particles). At the high temperatures in central regions of stars, a few protons move fast enough to circumvent the electrical repulsion exerted on them by other protons or heavier nuclei and stick to them by virtue of short-range nuclear forces. Thus, when hydrogen is gradually transformed into helium, enough energy is made available to enable the Sun to shine at its present rate for about 10,000,000,000 years.

From nuclear physics one may predict how the rate of nuclear reactions depends on physical conditions (chiefly temperature), and atomic theory predicts the rate at which radiation will be absorbed by matter on its way out through the star. Using the appropriate initial chemical composition (typically about 73 percent by mass of hydrogen, 25 percent helium, and 2 percent heavier elements), it is possible to calculate a theoretical model of the internal structure of a stable homogeneous star in which all features are fixed by choosing the total mass.

Stellar evolution

The theory of stellar evolution is concerned with how stars reach this phase by gravitational contraction and what happens afterward as hydrogen begins to be used up in the central regions. The lifetime of a star depends on the rate at which this happens: massive, very luminous stars use up their hydrogen in only about 1,000,000 years, whereas smaller stars, such as the Sun, are more thrifty and can have survived during the age of the Galaxy. Hydrogen is gradually burned to make helium, the core contracts, and the process continues until it becomes so hot and dense that helium itself is ignited to form carbon or oxygen. Contraction of the core is then halted, and the star assumes a complicated structure with two nuclear burning zones. It is probably at this stage that many stars become pulsating variables because of effects occurring in their envelopes while they evolve through a narrow instability strip in the Hertzsprung–Russell diagram, which illustrates the relationship between stellar luminosity and surface temperature. A greater understanding of these problems has been obtained in recent years by detailed numerical calculations using fast computers.

Later stages of stellar evolution have been worked out in rough outline only. The star is expected to become (1) a hot white dwarf (see above), possibly after blowing off its envelope as a planetary nebula; the white dwarf then slowly cools; (2) a neutron star (neither a white dwarf nor a neutron star can have a mass exceeding a certain limit that is not very different from the Sun's mass); (3) a greater mass that will either blow up or simply go on contracting until, in accordance with general relativity theory, it becomes a "black hole" or "collapsar," which can influence its surroundings by gravitational action only. No light or radiation can be received from it at such a stage. Massive stars in any case reach a more advanced state of nuclear evolution, involving successive ignition of heavier nuclei at their centres (see CHEMICAL ELEMENTS, ORIGIN OF). Much of the outer layers may consequently be ejected in an explosion that can be identified with one kind of supernova outburst, leaving a residue at the centre that may become a pulsar. Heavy elements from carbon upward, synthesized in the explosion, may be scattered into the general interstellar medium, gradually enriching the latter in heavy elements that can then be incorporated in later generations of stars; most heavy elements in the universe could have been produced in this way. Supernovae are probably also the main source of cosmic rays.

Outer layers of stars

While the size and total luminosity of a star are governed mainly by its internal constitution, the distribution of radiation over the spectrum depends on the structure of its outer layers, or "atmosphere"—a relatively thin skin from which radiation escapes into space and which contains the only part of the star that is visible. To drive radiation out requires an increase of temperature inward. The region emitting most of the radiation is called the photosphere, and details of the spectrum depend on how the opacity varies with wavelength. The study of stellar atmospheres forms another branch of astrophysics.

The study of stellar atmospheres. Much is learned about stars from their spectra. Measurement of the continuous bright background, or continuum, affords a way of discovering the surface temperatures. It is formed by processes in which radiation is emitted and absorbed in stellar atmospheres: in hotter stars chiefly by the interaction between electrons and protons and in cooler stars by that between electrons and ordinary hydrogen atoms. Superimposed on the continuum are narrow spectral lines (usually dark, caused by absorption by cooler atoms) characteristic of atoms, ions, and molecules that can be identified (and their radial velocities measured) by comparison with lines at similar wavelengths in laboratory sources. Spectral lines also give information on the chemical composition, temperature, and amount of turbulent motion within the star, the star's speed of motion toward the observer or away from him, rotation, progressive changes in novae, and so on. Some stars are spectroscopic binaries showing a periodic shifting or doubling of lines caused by orbital motion; sometimes the components also eclipse each other (*e.g.*, Algol), in which case a great deal of information about mass, radius, and luminosity can be deduced; but very close binaries, such as β Lyrae, are found to be distorted by tidal interaction. Bright lines occur in many stellar spectra, often because an extended gaseous envelope is present.

Finer gradations in the spectrum show differences in pressure caused by differing surface gravity, both being lower in stars of high luminosity; this observation is helpful in estimating absolute magnitudes and, hence, distances. Many of the cooler stars also have two emission lines that are caused by ionized calcium and arise in the outer (chromospheric) layers; their width increases regularly with luminosity, and their intensity in main-sequence stars is inversely related to age, so that deductions regarding age can be made.

Unusual atmospheric compositions in stars

Most nearby stars have about the same atmospheric composition as the Sun, but the spectra of some show composition that may be caused either by the star's own nuclear evolution—a few show the relatively short-lived radioactive element technetium—or by loss of the outer envelope, with consequent exposure of the helium-rich core, or the stars may have been formed long ago before the interstellar medium had been much enriched in heavy elements and hence are very deficient in metals. Some slowly rotating stars, some of which have unusually strong magnetic fields, show extraordinary chemical compositions not easily explained.

The Sun is a quite average star, but it can be studied in great detail and provides an excellent standard to compare with other stars.

An important minority of stars are intrinsically variable, and these are important indicators of distance and stellar population because they are easily identified. The main classes of variable stars are pulsating, irregular, and cataclysmic. They can be readily distinguished by their spectra and light variation.

The study of supernovae. Supernovae are rare and violent, being as bright at maximum as a whole galaxy (say, 10,000,000,000 times the Sun's luminosity) and expending so much energy that the whole star must be disrupted. They are often seen in external galaxies, where they occur perhaps a few times per galaxy per century. Two main types have somewhat different light variations and spectral characteristics. Both expel material at thousands of kilometres per second, and one type may supply heavy elements to the interstellar medium, though there is no direct observational evidence for this. The turbulent Crab Nebula is a radio and X-ray source within the Galaxy, and the nebula and the pulsar discovered in it in 1968 are remnants of a supernova seen in AD 1054. The Crab pulsar is unique in having been detected in light and X-rays as well as in radio waves; a second pulsar, in the constellation Vela, is also located in a recognizable supernova remnant.

THE STUDY OF GALAXIES AND THE UNIVERSE

The study of interstellar material. Rarefied material, such as dust and gas, in interstellar space has important effects because of the vast dimensions of the space in which it is present. Its study depends much on systems of trial and error, because neither the rarefaction nor the extent of space can be reproduced in a laboratory. Studies of complex organic molecules have led to identification of some of these substances in the gas. The exact nature of the dust is uncertain, but it consists of minute grains that may contain ices, graphite, silicates, and more complex organic molecules.

The gas is distributed in the galactic disk, mainly in the form of vast clouds with a few atoms (chiefly hydrogen) per cubic centimetre, which produce sharp interstellar absorption lines in the spectra of remote, hot stars. In the immediate neighbourhood of such stars, the gas glows, especially in light of the red hydrogen line Hα, because stellar ultraviolet radiation ionizes the atoms by stripping off electrons, which are then recaptured with radiation of energy. Such ionized hydrogen clouds are called H II regions, some of which are seen as diffuse nebulae, such as the Orion Nebula. Planetary nebulae emit bright lines by a similar mechanism, but these are smaller, nearly spherical shells expanding from a hot, central contracting old star, whereas diffuse nebulae surround several young stars newly born from them. By studying such useful special regions, a large amount of data about the gas has been gradually built up.

Much has been learned about interstellar gas from radio observations. H II regions emit thermal noise because of the electrical interaction of protons and electrons, but there is general background radio noise from the Milky Way, the continuous spectrum of which indicates a wholly different, nonthermal origin. This noise is synchrotron radiation caused by relativistic (*i.e.*, enormously energetic) electrons spiralling in the magnetic field of the Galaxy at speeds close to that of light. Discrete radio sources within the Galaxy are (1) supernova remnants emitting synchrotron radiation, (2) H II regions and planetary nebulae, or (3) pulsars (discovered in 1967), which emit a short, sharp pulse at very regular but sometimes gradually increasing intervals between 0.03 and two seconds and seem to be rapidly spinning neutron stars with enormous magnetic fields left over after supernova outbursts.

Many spectral lines caused by interstellar gas have been discovered by optical, ultraviolet, and radio observations. Most extensively studied is the 21-centimetre line, produced in ordinary hydrogen atoms. Lines at similar or shorter radio wavelengths are caused by changes in the energy states of various interstellar atoms and molecules, most of which involve some kind of maser action. In a maser the distribution of atoms or molecules among their different possible energy states is maintained at a very different level from that corresponding to thermal equilibrium; this condition makes certain emission sources extraordinarily intense. The unexpected discovery of complex organic molecules in space may shed new light on the question of the origin of life; thus, new discoveries in physics have again been useful in furthering the progress of astrophysics.

Organic molecules in space

Optical spectra of most external galaxies seem to be blends of stellar spectra with a few emission lines from H II regions. Some galaxies, however, have bright star-like nuclei for which the spectrum has a nonstellar continuous background of variable brightness and shows broad emission lines typical of those from a hot, turbulent gas; many of these galaxies have been found to be strong radio, infrared, and X-ray sources, as are certain other extragalactic objects. Many of the radio sources have now been identified optically and fall into two main classes: extended radio galaxies with a peculiar but recognizable optical galaxy at the centre of a double radio source; and quasi-stellar sources or quasars that resemble stars on direct photographs but have emission-line spectra with enormous red shifts (wavelengths up to four times the laboratory wavelength). If these red shifts are of cosmological origin, as for ordinary and radio galaxies, then they indicate enormous distances and hence luminosities. Among other difficulties still unresolved are the following. The total energy stored in relativistic particles and magnetic fields of extended radio galaxies is enormous, a fair fraction of the whole store of nuclear energy in stars of a normal galaxy; quasars are smaller but raise problems of enormous luminosity generated in small volume. Possible sources of gravitational energy on a sufficiently lavish scale might be superstars with perhaps 100,000,000 times the mass of the Sun, disappearance of galactic matter into a central general relativistic singularity or black hole, and a compact cluster of supernovae and neutron stars. A natural scheme of evolution would be for quasars to expand into an extended radio galaxy, perhaps repeatedly.

These problems in high-energy astrophysics have been intensified by infrared, X-ray, and gamma-ray measurements. Many objects in the Galaxy are strong infrared sources at ten to 20 microns' wavelength; these are believed to be newly formed stars surrounded by a "cocoon" of dust, red giants seen through intense obscuration, or compact H II regions containing dust that absorbs ultraviolet and reradiates in the infrared. Some galaxies and certain quasars have a still more dramatic excess of infrared radiation. The Earth's Galaxy and the Andromeda Nebula (Messier 31 [M31]) have compact nuclei that are also quite strong radio and infrared sources, indicating perhaps a quasar-like phenomenon on a much smaller scale.

Enormous energies are also emitted by some objects in the form of X-rays. Galactic X-ray sources are distributed around the Milky Way in the same manner as novae and planetary nebulae, concentrated toward the plane and central bulge. They include objects resembling old novae, the Crab Nebula and its pulsar, and some pulsating sources whose identity is not yet clear. Strong X-rays have also been detected from radio galaxies and others. Finally, there is an X-ray background coming from all directions and caused perhaps by remote quasars and radio galaxies (see X-RAY SOURCES, ASTRONOMICAL).

Another phenomenon of astrophysical importance is the presence of the primary cosmic rays; these are relativistic particles (mostly protons) that are believed to contribute about the same amount to the energy stored in the galactic disk (and exert about the same pressure) as do the interstellar magnetic field, turbulent motions of gas, and the cosmic background radiation. They cannot be directly traced to their sources as electromagnetic radiation can, because, being electrically charged, they must spiral around the lines of force of interstellar, interplanetary, and terrestrial magnetic fields. Two astrophysically im-

Cosmic rays

portant characteristics of cosmic rays are the relative numbers with different energies (energy spectrum) and the chemical composition (or charge and mass distribution). Cosmic rays include some electrons (and a few positrons), and their energy spectrum is of the right kind to explain the spectra of nonthermal radio sources, strengthening the view that the rays originate largely in supernovae. Their chemical composition differs somewhat from that of ordinary material, and the implications of this fact are discussed in the article COSMIC RAYS.

The study of galactic structure and evolution. A galaxy is a large system of stars, dust, and gas; examples are the Earth's Galaxy, of which the Milky Way is part, the two nearby Magellanic Clouds, and the Andromeda Nebula (M31). Most galaxies belong to one of three main structural types: flat spirals with a central bulge, such as the Earth's Galaxy and M31; elliptical galaxies; and irregular systems. In the older literature, external galaxies were often referred to as extragalactic nebulae. The flat disk of spirals, seen projected on the sky as the Milky Way in the case of the Earth's Galaxy, is caused by rotation; the Sun will take about 200,000,000 years to complete one revolution around the galactic centre. Most galaxies are assembled in huge groups or clusters containing tens or hundreds of galaxies.

Knowledge of the Galaxy and of other systems of similar size—the external galaxies—is pieced together from optical data referring to nearby and more distant stars, studies of interstellar dust and gas, radio observations of the gas—at 21 centimetres, the wavelength of a hydrogen atomic transition—and of supernova remnants and other radio sources at other wavelengths. Statistical studies of stellar types and the motions of the stars in painfully limited regions were for many years the only methods of approach to the problem of determining the structure. Now, increasing use of large telescopes has opened up possibilities of direct extension of accurate observations to much greater distances and has also afforded increased possibilities of comparison of the Earth's Galaxy with external systems that can be seen as a whole, yet studied area by area, sometimes star by star, in great detail.

Since 1951, radio studies of the Galaxy have added further evidence; in particular, the radio emission line of atomic hydrogen at 21-centimetre wavelength has been observed in regions of the Milky Way. Because this emission is not absorbed in interstellar space, its measurements provide information of galactic structure out to and beyond the distance of the galactic centre (about 10,000 parsecs). Ideas on spiral structure in these regions can be compared with optical observations of objects in the nearer spiral arms and comparisons traced out from a combination of painstakingly accumulated data.

Evolution of the Galaxy

The evolution of the Galaxy is another aspect of galactic studies, for which the discovery in 1944 of two stellar populations was of great importance, although this idea has since been modified somewhat. Most nearby stars, such as the Sun, have been found to belong to a disk population (Population I) moving in nearly circular orbits close to the galactic plane; the youngest stars are still in spiral arms where they have just been formed by condensation from interstellar clouds. A few nearby stars, however, all about 10,000,000,000 years old, have velocities of more than 150 kilometres per second relative to the Sun because of slower galactic rotation, and these also have a much lower proportion of heavy elements; they move in highly elliptical orbits, often at large angles to the plane of the Milky Way. Because heavy elements presumably result from the cumulative effects of earlier supernova outbursts, these halo, or extreme Population II, stars (resembling stars in globular clusters) are probably among the first stars ever to have been formed. The Galaxy may initially have been a vast gas cloud collapsing in free fall, because the orbits of stars are unmodified by stellar encounters in 10,000,000,000 years and thus are likely to be products of the initial conditions. Gas clouds, on the other hand, undergo collisions, so that their energy of motion would have been changed into heat and radiated away subject to the conservation of angular momentum (the amount of rotational motion); thus, gaseous material would have formed a disk enriched in heavy elements by more massive, short-lived contemporaries of the halo stars. Population I stars were born from the disk and have inherited from it both their orbital motions and their chemical composition. Star formation and presumably heavy-element enrichment are still going on in the disk today, though at a much reduced rate; but the halo population is "fossilized" because there is no interstellar matter present to form new stars. Also, there seems to be no way in which a flattened galaxy can evolve into an elliptical galaxy or vice versa.

Cosmology and cosmogony. Cosmology, the study of the universe as a whole, became a coherent discipline after Einstein's development of the theory of general relativity (1915)—though much can also be understood in Newtonian terms—and the discovery in 1929 that galaxies have spectral red shifts (increased wavelengths of spectral lines indicating velocity of motion away from the observer) proportional to their distances. The distances and times are so great that new problems in their description arise. Three Euclidean dimensions of space and one unidirectional one of uniformly flowing time are no longer necessarily sufficient to frame the phenomena.

Einstein based his theory of gravitation on the principle of equivalence (gravitational accelerations are the same for all test bodies and can therefore be eliminated locally by measuring from an accelerated frame of reference, such as a falling elevator) and expressed the gravitational field mathematically in terms of quantities representing a local curvature of space-time. The theory agrees locally with Newtonian theory (*e.g.*, in applications to the solar system) apart from four effects in which observation favours Einstein's theory (see RELATIVITY; GRAVITATION).

If the universe is assumed to be homogeneous and isotropic (this assumption, called the cosmological postulate, simplifies the situation by supposing the matter to be completely smoothed out), then the theory of general relativity permits the existence of three kinds of space: flat (Euclidean), spherical (closed), and hyperbolic (open). For each curvature there are still several descriptions of a possible universe, called models: expanding, contracting, accelerating, or decelerating (and perhaps oscillating). Different models predict different relationships among red shifts, apparent magnitudes, and angular sizes of very distant galaxies, but the interpretation of these observed quantities involves large and uncertain corrections. Observation does show that the universe is expanding and that if this expansion has been continuous since a point at which it began, the so-called big bang, its age is similar to that of radioactive elements and that of the oldest stars in the Earth's and nearby galaxies, so that the formation of galaxies seems to be connected somehow with the early phases of the universe. The reality of the big bang has been questioned, notably in the steady-state theory, according to which the average density of the universe is kept constant by continuous creation of new matter, and its large-scale appearance always remains the same; but observational results seem rather to favour the big-bang theories (see UNIVERSE, ORIGIN AND EVOLUTION OF).

Models of the universe

III. Trends in modern astronomical investigations

The period since 1950 has produced an enormous widening of the scope of astronomy, thanks to new technologies: electronics, antennas, computers, timekeeping, improvements in the design of radiation and particle detectors, and rocketry. Many observations now involve a combination of sophisticated techniques in different fields, and this situation has led to increasing use of teamwork and interinstitutional collaboration to develop particular experiments or joint national and international facilities. Sophistication has led to escalation in costs, so that most work is done in planned projects funded by specific financial grants; this situation in turn requires increasing effort in costing, justification, and administration.

Another contrast with the past is the increasing role of "invisible" astronomy. Radio astronomy, pioneered in

the 1930s by United States radio engineers Karl Jansky and Grote Reber, was widely ignored until the end of World War II, but since then the use of it has yielded many surprises: nonthermal emission, solar bursts, quasars, pulsars, and complex interstellar molecules. Radio astronomy has also created some of the most impressive astronomical instruments existing today, both dishes and interferometers. To measure angular diameters much less than 1″ of arc—*e.g.*, for hydroxyl (OH) emission sources and quasars—requires interferometry with base lines of thousands of miles stretching across the Earth. Modern standards of timekeeping enable this to be done by recording simultaneously in the Americas and Europe or Australia and combining the recorded signals afterward.

Ultraviolet, X-ray, gamma-ray, and infrared measurements have extended the spectral range of electromagnetic radiation to other invisible wavelengths. Furthermore, astronomers are becoming very interested in different kinds of radiation altogether—*e.g.*, cosmic rays, neutrinos, and gravitational radiation predicted by general relativity theory—and all three are being actively investigated. The result is a much richer universe than man knew in the mid-1940s.

Nevertheless, optical observations have not diminished in importance, partly because a photograph is still the most convenient means of identification, position finding, and study of structure and partly because only a few distances can be found without optical methods. Furthermore, ordinary stars and galaxies, radiating mostly in the near-visible, still present exciting problems. The demand for large optical telescopes is therefore increasing rather than diminishing.

Whether or not ground-based telescopes will be eventually replaced for optical observations by instruments in orbit or on the Moon is not certain. Such a development becomes more probable as space-based instruments and laboratories become more reliable, sophisticated, and flexible. On the other hand, the cost must be considered. Early in the 1970s, a sum of the order of $10,000,000 would buy either a fairly sophisticated instrumented satellite with, say, a 20-inch telescope giving observations for a few years or a fully equipped ground-based 100-inch telescope giving observations for decades; each can do some things that the other cannot. A manned Apollo mission, on the other hand, cost something in the region of $1,000,000,000, and such projects can result only from political decisions by very major powers. For many years to come, therefore, it is plausible that ground-based optical observations will continue to flourish, probably in increasing coordination with radio and space-based observations.

BIBLIOGRAPHY. C. FLAMMARION *et al.*, *Astronomie populaire* (1960; Eng. trans., *The Flammarion Book of Astronomy*, new ed., 1964); F. HOYLE, *Astronomy* (1962); and D.H. MENZEL, *Astronomy* (1970), are excellent introductions for the general reader. L.H. ALLER, *Atoms, Stars and Nebulae* (1971); L. MOTZ and A. DUVEEN, *Essentials of Astronomy* (1966); C. PAYNE-GAPOSCHKIN, *Introduction to Astronomy* (1954); and A. UNSOLD, *The New Cosmos* (1969), are suitable for students with some background in physics. A. PANNEKOEK, *A History of Astronomy* (1961), is comprehensive and scholarly; interesting aspects of astronomical history are described in H.C. KING, *The History of the Telescope* (1955); and SIR BERNARD LOVELL, *The Story of Jodrell Bank* (1968), while readable extracts from classical papers are reproduced in H. SHAPLEY and H.E. HOWARTH (eds.), *A Source Book in Astronomy* (1929); and H. SHAPLEY (ed.), *Source Book in Astronomy, 1900–1950* (1960). Star charts and lists of objects visible in a small telescope are given in A.P. NORTON and J.G. INGLIS, *A Star Atlas and Reference Handbook (Epoch 1950) for Students and Amateurs*, ed. by R.M.G. INGLIS, 15th ed. (1964); and astronomical pictures are collected in B. ERNST and T.E. DE VRIES, *W.P. atlas van het heelal* (1961; Eng. trans., *Atlas of the Universe*, 1961); and in P. MOORE (ed.), *The Atlas of the Universe* (1970). J.B. SIDGWICK, *The Amateur Astronomer's Handbook* (1955); and the *Journal* (6/year) and annual *Handbooks* of the BRITISH ASTRONOMICAL ASSOCIATION provide information for amateur astronomers, as does the journal *Sky and Telescope* (monthly).

An up-to-date account of solar-system astronomy is F.L. WHIPPLE, *Earth, Moon and Planets*, 3rd ed. (1968); aspects of this are described in J.A. WOOD, *Meteorites and the Origin of Planets* (1968). SIR HARRIE S. MASSEY, *Space Physics* (1964), gives a more technical account of research using rockets and artificial satellites. An excellent introduction to theoretical astrophysics is given by R.J. TAYLER, *The Stars: Their Structure and Evolution* (1970). Astronomical instruments are dealt with in D.S. EVANS, *Observation in Modern Astronomy* (1968); and G.R. MICZAIKA and W.M. SINTON, *Tools of the Astronomer* (1961). B.J. and P.F. BOK, *The Milky Way* (1957), explains galactic structure in simple terms; an introduction to this subject at a more technical level is D. MIHALAS and P.M. ROUTLY, *Galactic Astronomy* (1968). A.D. THACKERAY, *Astronomical Spectroscopy* (1961), is a fine introduction to observational astrophysics. The best introductions to radio astronomy are F.G. SMITH, *Radio Astronomy*, 2nd ed. (1962); and J.S. HEY, *The Radio Universe* (1971). D.W. SCIAMA, *Modern Cosmology* (1971), is a good introduction to cosmology. F.D. KAHN and H.P. PALMER, *Quasars* (1967); and T. and L.W. PAGE (eds.), *Beyond the Milky Way* (1969), explain aspects of extragalactic astronomy in simple terms. C.W. ALLEN, *Astrophysical Quantities*, 2nd ed. (1963), summarizes the subject in numbers and formulas.

(B.E.J.P.)

Atacama Desert

The Atacama Desert (Desierto de Atacama) of Chile is a cool, arid region in northern Chile, 600 to 700 miles (1,000 to 1,100 kilometres) long. Its limits are not exact, but it lies mainly within Antofagasta and Atacama provinces between the south bend of the Río Loa and the mountains separating the Salado-Copiapó drainage basins. To the north, the desert region continues into the Province of Tarapacá.

The desert itself consists mainly of salt pans lying at the foot of the Cordillera de la Costa to the west, and of alluvial fans sloping from the foot of the Precordillera ranges to the east; some of the fans are sandy and covered with dunes, but extensive pebble accumulations are more common.

Topography of the Atacama

Several parts of the Atacama Desert may be distinguished. Its western part includes a coastal chain of mountains 5,000 feet or so in height with peaks reaching to 6,560 feet. There is no coastal plain; through much of their extent the mountains end abruptly in cliffs, some of them higher than 1,600 feet, making communication difficult between the coastal ports and the interior. The railroad from the port of Iquique, for example, must ascend about 2,300 feet before it passes through a valley to the interior.

In the interior a raised depression extends north and south and forms a high plain at an altitude of more than 3,000 feet. Farther to the east is the western range of the Andes, preceded by the Cordillera Domeyko. Here there are numerous volcanic cones, some exceeding 16,000 feet in altitude. Along Chile's northeastern frontier with Argentina and Bolivia, between the western and eastern ranges of the Andes, is a high plateau more than 13,000 feet in altitude, called the Puna de Atacama.

The Desert of Atacama forms part of the arid Pacific shoreline of South America. Moist air masses coming from the tropical Amazon Basin tend to be blocked by the Andes, making the desert one of the driest regions in the world. Some artesian waters exist, but their boron content makes them unsuitable for agriculture. In the Pampa del Tamarugal, to the north of the Río Loa, many salt beds have formed over subterranean waters. In the late 1960s about 37,000 acres were planted with tamarugoes, an acacia-like legume that can be used as fodder for sheep. The sprout sends out a root that descends quickly to the subterranean moisture.

On the coast the aridity is the consequence of the Peru (Humboldt) Current that brings cold water from the Antarctic, causing a thermal inversion—cold air at the surface of the ocean and warmer air higher up. This condition produces fog and stratus clouds, but no rain. Heavy rains fall in Iquique or Antofagasta only two to four times a century.

Temperatures in the desert are relatively low compared with those in similar latitudes elsewhere. The average summer temperature at Iquique is only 66° F (19° C) and at Antofagasta 65° F (18° C).

Area of conflicts over mineral resources

For many years in the 19th century, the desert was the source of conflicts between Chile, Bolivia, and Peru because of its valuable resources, particularly sodium nitrate deposits located northeast of Antofagasta and inland from Iquique. Much of the area belonged to Bolivia and Peru, but the mining industry there was controlled by Chilean interests, which were strongly supported by the Chilean government. From the War of the Pacific (1879 to 1883) between the three countries, Chile emerged victorious. The resulting Treaty of Ancon (1884) gave Chile permanent ownership of sectors previously controlled by Peru and Bolivia, the latter losing its whole Pacific coastline.

The area proved to be one of the chief sources of Chile's wealth until World War I. Nitrate deposits in the central depression and in several basins of the coastal range were systematically mined after the middle of the 19th century. Ports were built at Iquique, Caldera, Antofagasta, Taltal, Tocopilla, Mejillones, and, farther north, Pisagua, and railroads penetrated the mountain barriers to the interior. Prior to World War I, Chile had a world monopoly on nitrate; in some years 3,000,000 tons were extracted, and the taxes on its export amounted to half the government's revenues. The development of synthetic methods of fixing nitrogen have since reduced the market to a regional one, and only 780,000 tons per year were produced in the late 1960s. Once, 120,000 nitrate workers were employed; now there are fewer than 10,000. Whole towns were abandoned, and the port of Pisagua had fewer than 25 inhabitants in 1970.

Some sulfur is still mined in the high Cordillera. The region's chief source of revenue, however, is copper mining at Chuquicamata in the Andes and at Paposo on the coast. The industry developed with the entry of U.S. capital and technical skill in the early part of the 20th century. The deposits are abundant, and the government (1964–70) of Pres. Eduardo Frei (Montalva) had an ambitious development program. The succeeding government of Pres. Salvador Allende (Gossens) nationalized all U.S.-owned copper companies.

Some farming is done in the desert, but this supports only a few thousand people. Lemons are grown at Pica, and a variety of products are cultivated on the shores of the salt marshes at San Pedro de Atacama. At Calama, near Chuquicamata, water from the Río Loa irrigates potato and alfalfa fields.

Atacama had become a region of declining population, despite the presence of the copper industry. The development of fishing and the establishment of canning factories, especially at Iquique, provided employment for relatively few. So far had the decline of population gone that government programs for afforestation and sheep breeding were faced with a lack of manpower.

(J.-L.-F.T.)

Atatürk, Kemal

Kemal Atatürk, a distinguished Turkish soldier, reformer, and statesman, was the founder of the Republic of Turkey and its first president. His successful struggle for the liberation of Turkey against the powers of the Entente (an alliance of Britain, France, and Russia) after Turkey's defeat in World War I has inspired many embryonic states in Asia and Africa to fight for their independence.

Early life and career. Kemal Atatürk, born in 1881 in Salonika, Greece, of a Turkish family of humble origin, was named Mustafa. His mother was Zübeyde Hanım, his father Ali Riza, a minor government employee. When Mustafa was still in the primary school he lost his father, and his mother took the boy to live in the country with her brother. Returning to Salonika later, Mustafa finished primary school and entered the military secondary school in that town in order to become an officer in the Ottoman army. It was at this school that one of his teachers, who admired the boy for his skill in mathematics and who was also called Mustafa, suggested he should call himself Mustafa Kemal (maturity and perfection). After finishing secondary school, Mustafa Kemal went on to the military high school in Manastir, where, observing with hatred the continuous attacks of the Christian Macedonian anarchists on the Turkish population, he became, like most of his fellow cadets, an ardent nationalist.

Education

In 1899 Mustafa Kemal entered the Military Academy in Istanbul. There he came to take a close interest in politics; he and his fellow students read secret pamphlets attacking the despotic rule of Sultan Abdülhamid II. He was especially influenced by the patriotic and liberal thinking of the poet Namık Kemal Bey. He also read books on the French Revolution and developed an admiration for Napoleon.

In 1902 Mustafa Kemal graduated from the Military Academy and entered the General Staff College, where his interest in politics continued. After finishing the college with the rank of captain, he was appointed to the cavalry regiment in Damascus. There, together with some of his friends, he founded a secret society called "Fatherland and Freedom," which, however, failed to make much progress. On his return to Salonika, Mustafa Kemal, like many of his fellow officers, joined the secret "Committee of Union and Progress," which spread its revolutionary activities throughout the armed forces and caused the proclamation of the Constitution of 1908.

Mustafa Kemal devoted all his time and energy to his profession in the following years. In 1911, when the Italians attacked Tripoli, an Ottoman province at that time, he hurried there together with some of his officer friends and, forming troops of the natives, launched successful guerrilla raids on the enemy. In the same year Mustafa Kemal was promoted to major. In the Balkan War of 1912 he was charged with the defense of the Gallipoli Peninsula—a task that gave him an excellent opportunity to study the strategic position of this important area. In 1913 he was sent to Sofia as military attaché, and during his stay there he acquired a good knowledge of the Western standards in taste, the arts, and the relations between men and women in polite society. He made good use of this knowledge later when he set about reforming social life in his country. While still in Sofia, Mustafa Kemal was promoted to lieutenant colonel.

Activities during World War I

When World War I broke out, Mustafa Kemal was appointed to the command of the 19th Division at Çanakkale. He defeated the British at Gallipoli twice and gained for himself in the Turkish press the title of "the Saviour of Istanbul." And he was promoted to colonel. In 1916, serving on the eastern front, he stopped the advance of the Russian forces to the south and was promoted to brigadier general.

In 1917 Mustafa Kemal accompanied the crown prince, Vahideddin, on a state visit to Germany. During a tour of the German Western Front he did not hesitate to express openly his view about the vulnerability of the front and Germany's position in the war. On his return to Istanbul Mustafa Kemal fell ill, and for treatment he went to Vienna and Carlsbad (now Karlovy Vary, Czechoslovakia) where he had a further opportunity to observe European civilization.

In 1918 Mustafa Kemal was appointed to command the 7th Army in Palestine; when he took up his duties, however, the fight with the British had all but ended, and the enemy was advancing northward without meeting any resistance. The Arab guerrillas too were launching attacks on the Turkish army. To avoid the capture of the whole 7th Army, Mustafa Kemal withdrew his forces to the north of Aleppo. When, after the Armistice of Mudros (Moúdhros), the German officers and commanders serving in Turkey returned to their country, Mustafa Kemal assumed the command of all the forces of the southeastern front. Disagreeing with the British over the enforcement of the terms of the Armistice, however, he was appointed to a post in the Ministry of War. On his arrival in Istanbul he found the fleet of the Entente anchored in the harbour. The terms of the Armistice were hard enough, but information was now received about a secret agreement reached by the states of the Entente for the partition of the Ottoman territories. Moreover, the minorities in Is-

tanbul and elsewhere had seized the opportunity to organize themselves against the Turks. The Turkish people looked for a means of redress; in some parts of the country they formed organizations called "the Society for the Defence of Rights" to fight against them.

Activities after World War I. In Istanbul there were two main ideas about Turkey's future: the Sultan and his supporters were thinking of placing the country under English protection, while some well-known Turkish journalists and intellectuals were spreading propaganda for placing Turkey under an American mandate. In both cases the aim was to maintain the Ottoman Empire in its cosmopolitan structure. Mustafa Kemal, however, persisted in the idea of an independent Turkish nation living within its national boundaries and believed that this could be achieved if the nation was prepared for a new struggle. Before deciding on a course of action, he had talks with many Turkish and foreign notables, including the Sultan and his ministers. Then he discussed it with his friends, all commanders who were bitterly disillusioned over the abolition of the Ottoman army by the terms of the Armistice, and saw the solution in starting a war of independence in Anatolia.

An excellent opportunity for this presented itself soon: the powers of the Entente were putting pressure on the Turkish government to take measures against riots likely to break out in the eastern provinces. The Sultan appointed Mustafa Kemal as Inspector of the Third Army in Erzurum, endowing him with power over military and civilian authorities. On May 15, 1919, immediately before Mustafa Kemal's departure for Erzurum, the Greeks occupied Izmir.

Mustafa Kemal's secret circular

After a secret interview with the Sultan, Mustafa Kemal left Istanbul with a large suite of staff officers and set foot in Samsun on May 19. In Amasya, with the approval of local corps commanders, he issued a secret circular dated June 22 in which he described the dangers that the country faced: how the government in Istanbul had yielded weakly to the forces of occupation and how the only hope of salvation lay in the nation's own struggle for its liberation. Such a struggle, he added, had already begun, and to make it the decision of the nation itself a national congress would be convoked in Sivas with the participation of three delegates from each province. He ordered all unit commanders to strengthen their forces, disregarding the terms of the Armistice about the demobilization of the Turkish army. Finally, he warned both the military and civilian authorities that henceforth they would take their orders from him alone.

Mustafa Kemal's demands were fervently complied with by the military because his demands meant saving the army from extinction. The army took under its control all postal and telegraphic communications in Anatolia and forced into obedience those civil administrators who tried to resist Mustafa Kemal's orders.

In all the towns and cities he called at on his way to Sivas, Mustafa Kemal met the leading citizens and explained to them his views on a national struggle for independence. He arrived in Sivas amid warm demonstrations of support by the people, and after important talks with the notables of the city, he proceeded to Erzurum, ignoring all orders given by the Sultan's government, under pressure from the states of the Entente for his immediate return to Istanbul. In Erzurum a congress was to be convened on July 23 by the Society for the Defense of Rights in Eastern Anatolia. In the meantime the military and civilian authorities of Erzurum received an order for Mustafa Kemal's arrest and transport to Istanbul. Although this order went unheeded, Mustafa Kemal resigned his commission in the army, deeming it necessary to have more freedom as the leader of the national struggle he had started. Thus he entered the congress as a mere delegate and was elected its president. At his suggestion the congress accepted the National Covenant, which was in the nature of an oath requiring the indivisibility of the fatherland and the successful completion of the national movement. In addition, a Standing Committee of nine members was elected of which Mustafa Kemal was chosen president. On September 4 he opened the National Congress in Sivas, and he was again elected president of the congress. The National Congress adopted all the decisions taken by the Congress of Erzurum and rejected decisively the idea of placing Turkey under American mandate. Resisting a motion for establishing a new state in Anatolia, Mustafa Kemal proposed the joining of all local Societies for the Defense of Rights into one society to be called "the Society for the Defence of Rights in Anatolia and Rumelia." The proposal was accepted, and so the prototype of a political party was formed. The Ottoman cabinet of Ferid Paşa and succeeding Ottoman governments continued to view the national movement as an act of rebellion and Mustafa Kemal's activities as illegitimate.

Participation in nationalist congresses

Role in the founding and reform of modern Turkey. On December 27 Mustafa Kemal transferred the seat of the national struggle to Ankara, thinking it a more convenient location for his purposes. In the meantime, at the general elections of members for the Ottoman Chamber of Deputies in Istanbul, Mustafa Kemal's supporters won an overwhelming majority and succeeded in getting the chamber to proclaim as its own decision the principles of the National Covenant. They also secured the cancellation of the government's former decree about Mustafa Kemal's dismissal from the army. Alarmed at these indications of change in the Ottoman policy, the British occupied Istanbul officially on March 16, 1920, and dissolved the Chamber of Deputies. Mustafa Kemal vehemently protested the British government's action; but, in fact, the occupation of the Ottoman capital and especially the dissolution of the chamber were extremely useful to his aims because they removed the legal obstacle that Istanbul presented to his plan of forming a national government in Anatolia. So, after a new election of deputies, he opened on April 23 in Ankara the first Grand National Assembly of Turkey, and he was elected its president. At Mustafa Kemal's proposal, a constitutional law was passed changing the name of the state to Turkey and stipulating that sovereignty and executive powers would be used on its behalf by the Grand National Assembly. Accordingly, as president of the assembly Mustafa Kemal took upon himself the offices of the prime minister and of the president of the state. Thus ended the Islāmic form of government that had existed in Turkey since the Middle Ages; and, as in the French Revolution, the Turkish people passed suddenly from rule of absolutism and the caliphate to a regime based on national sovereignty. This important change caused serious uprisings in some regions, but they were quickly suppressed by the national forces.

Struggle to consolidate the country

Mustafa Kemal now busied himself with the work of gaining control of such parts of the country as were then under occupation. First, in the east, the Armenians and the Georgians were defeated, and through the mediation of Soviet Russia, a treaty was signed with them that regained for Turkey even the territories she had lost in 1878. After extensive guerrilla warfare, the French in the south evacuated Turkish territories and withdrew to Syria and recognized the legitimacy of the National Government in Ankara. Ignoring Ankara altogether, the British got the Ottoman government to sign the Treaty of Sèvres. The National Government proclaimed that it did not recognize as legitimate a treaty of such terms, whereupon the Greek army extended its area of occupation, advancing within 50 kilometres of Ankara. At this time of great anxiety the National Assembly appointed Mustafa Kemal the commander in chief with extraordinary powers. Indeed, on August 26, 1922, after an all-out offensive planned and directed personally by the Commander in Chief, the Greek army was defeated and forced within two weeks to leave Anatolia completely. Upon this decisive victory and with the mediation of the states of the Entente, an armistice was signed with the Greeks according to which they evacuated all Turkish territories. The British ceded Çanakkale and Istanbul to the National Government. Vahideddin, the last Ottoman sultan, fled abroad, and upon a motion by Mustafa Kemal the National Assembly terminated the 600 years of Ottoman rule in Turkey. The Treaty of Lausanne, signed

on July 24, 1923, established the integrity of Turkey's national frontiers and its complete independence. All privileges granted to the European countries by the Ottomans were cancelled. Thus, Mustafa Kemal realized his dream of founding a completely independent and national Turkish state in place of the Ottoman Empire, that "sick man of Europe" that had been for a long time a subject of strife among the great powers of Europe.

Marriage

In 1922 Mustafa Kemal married Latife Hanım, the well-educated daughter of a wealthy family in Izmir. The marriage was contracted in the modern manner, not in the tradition of Islām. In order to show the Turkish people that the place of women in society was by the side of their men, he took his wife with him on his trips around the country. His marriage did not last long, however. During his long years of single life he had developed an independent habit of living that he found difficult to give up and that his wife could not tolerate.

On one occasion as early as 1917, Mustafa Kemal had remarked that, had he the power and the authority, he would change social life in Turkey at one blow. This opportunity had now presented itself, and he launched on a program of reforms. In place of the Society for the Defense of Rights in Anatolia and Rumelia, he founded the People's Republican Party and became its leader. With the general elections held immediately after the signing of the Treaty of Lausanne, this party, as Turkey's only political party, took complete control of government. On October 29, 1923, Mustafa Kemal proclaimed the Republic and was elected its first president. In 1924 he abolished the caliphate. In the meantime, a group of his friends who were against his drastic methods of reform and who believed in gradual progress over a period of

Reforms

time founded the Progressive Republican Party. Mustafa Kemal went on carrying out his program of reform: he closed down all institutions based on the Muslim canon law, all monasteries, and religious orders. "Science is the most reliable guide in life," he remarked, and abolishing the traditional system of education, which was mainly religious, he established secular schools of the modern type. The whole Ottoman legal system was modernized, and a new civil and penal code was adopted. The Oriental forms of dress that carried a religious significance were discarded in favour of European dress. Dances, balls, and other forms of entertainment involving both men and women were encouraged, and the enlightened classes adopted the European way of life.

Mustafa Kemal's reforms did not go unchallenged. In Eastern Anatolia a man called Şeyh Said stirred up a rebellion to restore the Muslim canon law; in Izmir preparations for a plot to assassinate Mustafa Kemal were reportedly discovered; and there were said also to be some local attempts at rebellion against the use of hats. Mustafa Kemal punished severely all the leaders of these movements, closed down the Progressive Party, and, reverting to the former authoritarian regime, he pursued his program of reform. Setting aside all the old laws and traditions that held women inferior to men, he established complete equality between the sexes, including the right of electing and being elected. In 1928 he substituted Roman characters for the Arabic that had been used in Turkey for centuries. He endeavoured to popularize Western classical music and the theatre in Turkey. In 1930 he made a second attempt at introducing a multiparty regime by allowing the creation of the Free Republican Party; but, as this party soon became a centre for antireformist ideas and activities, it met the same fate as the Progressive Republican Party. Mustafa Kemal also launched a large-scale program of research in the fields of Turkish language and history. By this means he wanted to strengthen in society the ties of national feeling in place of the old ties of religion. In 1933 a law was passed to make the use of family names compulsory, and the National Assembly gave Mustafa Kemal the name Atatürk ("Father of Turks"), which soon became so popular as to supersede his previous name and titles.

Atatürk's foreign policy can be summed up by his motto: "Peace at home, and peace in the world." In economy, he followed a policy of national economy, nationalizing all foreign firms and companies. On the question of Turkey's industrialization, he placed his hope on private domestic capital for a while, but discovering its insufficiency, he decided to encourage etatism (state socialism). In neither case, however, did he achieve any important success. If one or two items of foreign policy are excepted, there was a gradual slowing down in the last five years of Atatürk's life, and his final year passed in serious illness. He died on November 10, 1938, in Istanbul, where he had gone to rest.

Atatürk made major reforms in Turkey in the field of politics, law, and culture that only affected, however, bureaucrats and a minority of well-to-do people in the cities. The poorer part of the population, and especially the peasants who still subsisted in an agricultural order of the medieval type, continued to live much the same as before. Nevertheless, the Western view of life had gained enough power among the educated classes to make a return to the old way of life impossible.

BIBLIOGRAPHY. For a complete bibliography, see J.P.D. KINROSS, *Atatürk: The Rebirth of a Nation* (1964); B. LEWIS, *The Emergence of Modern Turkey* (1961); and UNESCO, *Atatürk* (Eng. trans. 1963). Additional information may be found in the SOCIETE POUR L'ETUDE D'HISTOIRE TURQUE, *Histoire de la république Turque* (1935), written under the patronage of Atatürk.

(M.Ak.)

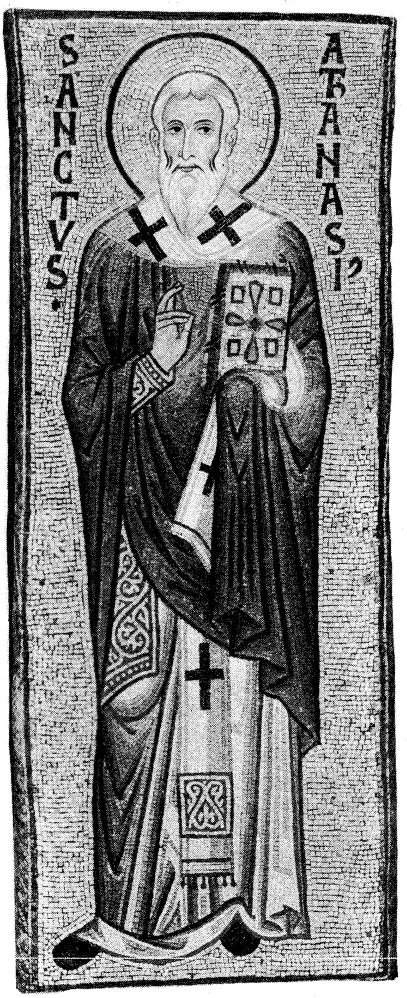
Athanasius, Saint

St. Athanasius, theologian, ecclesiastical statesman, and Egyptian national leader, was the chief defender of Christian orthodoxy in the 4th-century battle against Arianism, the heresy that asserted that the Son of God was a creature of like but not the same substance as God the Father. His two-part work of apologetics, *Against the Heathen* and *The Incarnation of the Word of God*, completed about AD 335, was the first great classic of developed Greek Orthodox theology. In Athanasius' system, the Son of God, the eternal Word through whom God made the world, entered the world in human form to lead men back to the harmony from which they had fallen away. Athanasius reacted vigorously against Arianism, for which the Son was a lesser being, and welcomed the definition of the Son formulated at the Council of Nicaea in AD 325: "consubstantial with the Father."

Life and major works

Athanasius was born at Alexandria, Egypt, about 293 and received his philosophical and theological training there; in 325 he attended Bishop Alexander of Alexandria as deacon at the Council of Nicaea. A recognized theologian and ascetic, Athanasius was the obvious candidate to succeed Alexander when the latter died in 328. The first years of his episcopate were devoted to visitation of his extensive patriarchate, which included all of Egypt and Libya. During this time he established important contacts with the Coptic monks of Upper Egypt and their leader Pachomius. Soon began the struggle with imperialist and Arian churchmen that occupied much of his life. He used political influence against the Meletians, followers of the schismatic bishop Meletius of Lycopolis, who had gone back on the plans made at Nicaea for their reunion with the church; but he refuted specific charges of mistreatment of Arians and Meletians before a hostile gathering of bishops at Tyre (in modern Lebanon) in 335, which he refused to recognize as a general council of the church. When both parties met the emperor Constantine at Constantinople in 336, Athanasius was accused of threatening to interfere with the grain supply from Egypt, and without any formal trial Constantine exiled him to the Rhineland.

The Emperor's death in 337 allowed Athanasius to return to Alexandria, but Constantine's son Constantius, emperor in the East, renewed the order of banishment in 338. Athanasius took refuge at Rome under the protection of Constantius' brother Constans, emperor in the West. An Arian bishop, Gregory, was installed at Alexandria; Athanasius, however, kept in touch with his flock through the annual *Festal Letters* announcing the date of Easter. Pope Julius I wrote in vain on his behalf, and the general council called for AD 343 was no more successful



St. Athanasius, detail of a 12th-century mosaic. In the Palatine Chapel, Palermo, Italy.
Anderson—Alinari

—only Western and Egyptian bishops met at Sardica (modern Sofia, Bulgaria), and their appeal for Athanasius was not accepted in the East. In 346, however, Constans' influence secured his return to Egypt, where he was welcomed as a popular hero. Athanasius' "golden decade" of peace and prosperity followed, during which he assembled documents relating to his exiles and returns in the *Apology Against the Arians*. Nevertheless, after the death of Constans in 350 and the following civil war, Constantius, as sole emperor, resumed his pro-Arian policy. Again political charges were brought against Athanasius, his banishment was repeated, and in 356 an attempt was made to arrest him during a vigil service. This time he withdrew to Upper Egypt, where he was protected in monasteries or friendly houses. In exile he completed his massive theological work *Four Orations Against the Arians* and defended his conduct in the *Apology to Constantius* and *Apology for his Flight*. The Emperor's persistence and reports of persecution at Alexandria under the new Arian bishop George led him, in the more violent *History of the Arians*, to treat Constantius as a precursor of Antichrist.

In 361 the death of Constantius (followed by the murder of the unpopular George) allowed Athanasius to return triumphantly once more to his see. Immediately, he convened a council at Alexandria (362), during which he appealed for unity among those who held the same faith but differed in terminology. The way was thus prepared for the orthodox doctrine of the Trinity—"three Persons in one substance"—which stresses distinctions in the Godhead more than Athanasius usually had done. The new emperor Julian the Apostate rather petulantly ordered Athanasius to leave Alexandria, and he sailed up the Nile again to his exile in Upper Egypt until Julian's death in 363. In 365 the emperor Valens, who favoured Arianism, ordered his exile once more, but this time the popular bishop merely moved to the outskirts of Alexandria for a few months until the local authorities persuaded the Emperor to reconsider. Finally, Athanasius spent a few years in peace until his death on May 2, 373.

Minor works and place in history

Among Athanasius' important works are *The Letters [to Sarapion]* on the divinity of the Holy Spirit and *The Life of St. Antony*, which was soon translated into Latin and did much to spread the ascetic ideal in East and West. Only fragments remain of sermons and biblical commentaries; several briefer theological treatises are preserved, however, and a number of letters, mainly administrative and pastoral. Of special interest are the letter to Epictetus (bishop of Corinth), which anticipates future controversies in defending the humanity of Christ, and the letter to Dracontius, which urges a monk to leave the desert for the active labours of the episcopate. Precision of thought, tireless energy in defense of his convictions and the freedom of the church, and (within certain limits) breadth of understanding have given Athanasius an important place among the teachers and leaders of the church; and as an Egyptian patriot he is also a significant figure in the history of his country.

BIBLIOGRAPHY. The great Benedictine edition of the works of Athanasius (1698), reprinted in J.P. MIGNE, *Patrologia Graeca*, vol. 25–28 (1857), is being replaced by that begun by H.G. OPITZ (*Athanasius Werke*) in 1934; major works are translated in the "Post-Nicene Fathers," Series 2, vol. 4, *Select Writings and Letters of Athanasius* (1892)—the "Prolegomena" to this volume by A. ROBERTSON remains the most complete life. For later discoveries and studies, see H.I. BELL (ed.), *Jews and Christians in Egypt* (1924); and the brilliant summary by F.L. CROSS, *The Study of St. Athanasius* (1945). There are useful sketches in G.L. PRESTIGE, *Fathers and Heretics* (1940, reprinted 1963); and in R.W. THOMSON'S edition of *Contra Gentes and De Incarnatione* (1971). For further references, see E.R. HARDY (ed.), *Christology of the Later Fathers* (1954); and J. QUASTEN, *Patrology*, vol. 3, pp. 20–79 (1960).

(E.R.Ha.)

Atheism

Atheism is the denial of God as the first principle and is thus antitheism, the opposite of theism. The words atheist and atheism can be found as early as in the works of Plato; they recur in the Christian Era in the Letter of Paul to the Ephesians and in the works of the Fathers of the Church, where, for example, St. Ignatius of Antioch attacks the pagans, calling them atheists, and Justin Martyr defends the Christians, claiming that they adore the true God and are thus not atheists. Atheism is opposed to any religion or worship of God; this is indicated by the synonym for atheist: freethinker. To many of its theological critics, atheism is synonymous with impiety, irreligion, disbelief, and even moral corruption, on the ground that every principle and every higher ethical law is rejected.

Divisions and principal forms. Atheism can be divided into theoretical atheism and practical atheism. Theoretical atheism is the denial of God based upon a system of thought that excludes the possibility of the existence of the Absolute. Practical atheism is the denial of God as reflected in the way one conducts his private and public life, leaving the question of God out of consideration and basing conduct solely on finite values.

Theoretical atheism

Theoretical atheism can be either negative or positive. Explicit negative theoretical atheism is attributed to those who unequivocally deny the existence of God and who suppose a concept of the world and of the destiny of man that radically excludes the necessity of the transcendent first principle or of an immortal human soul. Implicit negative theoretical atheism, or cryptoatheism, is attributed to those who, although they affirm the existence of God or of the Absolute, deprive him of some essential attribute. This type of atheism is also known as atheism "by consequence." The encyclopaedia prepared by the 18th-century French philosophers Denis Diderot and Jean d'Alembert equates skepticism (which questions the

ability of the human mind to know), indifferentism (which states that all religions are equally valid), and even agnosticism (which claims that the question of God's existence cannot be answered) with negative atheism because they close off all the paths that lead to God.

Positive theoretical atheism, which predominates in modern thought, replaces the acknowledgment of the transcendent first principle with the autonomy of the subjective thinking element (the *cogito*) within man that leads him to identify knowledge and being (or reality), wisdom and action, freedom and necessity; for the transcendence of God and for personal immortality, it substitutes the emergence of man in the world. Theoretical atheism, insofar as it is a vindication of the total autonomy of man and of his absolute freedom, is also called postulated atheism in the sense that God cannot and must not exist if man is to be guaranteed freedom and responsibility for his duties and his actions. Positive atheism is thus an anthropological atheism in which God is replaced by man.

Practical atheism

Practical atheism consists of ignoring or neglecting any relationship to God in one's actions, or in living as if God did not exist. According to Paul, in the letter to Titus, it is the situation of those who claim to know God but deny him in the things they do. Practical atheism, therefore, involves the orientation of one's life exclusively toward the attainment of earthly goals.

Historical development of atheism. The history of atheism is as complex and obscure as its notion. Atheism seems to have been unknown in the more primitive societies; it appears to be the fruit of civilization and of reflection. Although Plato expressly argues against atheists in the *Laws* and although some lists of men branded as radical atheists have been handed down, explicit atheists were rare in the period of ancient Greece and Rome and in the early Christian Era.

In the Middle Ages, the time of great Christian expansion, explicit atheism was practically unknown. According to some Marxist historians, however, the medieval heresies and the controversies between the schools of thought, especially the arguments concerning the relationships between reason and faith and between freedom and authority, should be interpreted as preparations for modern atheism. The return to a classical pagan conception of life in Humanism and in the Renaissance of the 15th and 16th centuries gave vitality to the atheistic currents that did exist. The Materialism of the ancient Greek philosophers Democritus and Epicurus and the syncretism of the Stoics and the Neoplatonists were fused into a theory that the world is self-contained and self-sufficient (see also STOICISM; PLATONISM AND NEOPLATONISM). The principal inventor and defender of the atheistic conception of the modern state was Niccolò Machiavelli, the Italian statesman and author who died in 1527. In his most famous work, *Il Principe* (written 1513; Eng. trans., *The Prince*, 1954), he defended the principle that the end justifies the means, thus affirming the independence of politics from morals and, in the final analysis, from any form of religion.

Influence of Descartes

Atheism in its modern positive form has been characterized by the thought of the French philosopher and mathematician René Descartes, who, in the first half of the 17th century, attempted to ground his thought on clear and distinct ideas and the proposition *Cogito, ergo sum* ("I think; therefore, I am"). He thus placed the foundation of truth in the evidence and in the freedom of the thinking subject. The Cartesian system is accused of being atheistic because of its mechanistic conception of the world, according to which all natural phenomena can be explained by reference to matter and motion and their laws. The first explicit argumentation defending atheism as moral, as no worse than idolatry, as not leading to corruption or death, was formulated by Pierre Bayle, a late-17th-century Cartesian fideistic skeptic; his defense provided the foundation for modern secularism. At this same time, Deism and British Empiricism were inserted into the intellectual scene by John Locke, who advanced the hypothesis: ". . . whether Omnipotency has not given to some systems of matter fitly disposed a power to perceive and think." This hypothesis became a categorical proposition for the French writer Voltaire and the British freethinker John Toland, who believed in the universal animation of matter and its self-sufficiency. These ideas were the central point of the atheism of the Enlightenment as expressed chiefly by the French Encyclopaedists and the British Deists.

The 18th century witnessed the triumph of the atheism of the French Enlightenment in the intellectual activity of the writers and thinkers referred to as Philosophes, who fused British Deism and Empiricism with Cartesian mechanism. The postulated atheism of these philosophers was expressed systematically and most outspokenly by Baron P.-H.T. d'Holbach, who defined an atheist as "a man who destroys the dreams and chimerical beings that are dangerous to the human race so that men can be brought back to nature, to experience, and to reason."

The German Enlightenment intensified opposition to supernaturalism (belief in an order of existence beyond the observable universe) in order to affirm the necessity of reason and the self-sufficiency of nature against any notion of transcendence. The official theologians became alarmed at this movement and reacted against the philosophers. Even Immanuel Kant, the founder of transcendental, or critical, Idealism, came into conflict with the civil and religious authorities for his *Die Religion innerhalb der Grenzen der blossen Vernunft* (1793; Eng. trans., *Religion Within the Limits of Reason Alone*, 1960), which he later defended in a work published in 1798. In that same year the *Atheismusstreit* ("conflict over atheism") broke out. Friedrich Karl Forberg, an unknown philosopher, published an essay in a periodical of which his friend Johann Gottlieb Fichte, an important German Idealist, was a coeditor. In the essay, Forberg reduced religion to a "practical belief" in the moral order of the world: to have religion is to act as if a just and moral world were possible. Fichte himself conceived of God as the simple, active, moral world order. For this theory Fichte was formally accused of atheism and in 1799 was dismissed from the University of Jena. When Fichte was accused of atheism, Friedrich Jacobi, a contemporary German philosopher who was trying to protect himself, claimed that the danger of atheism was rooted primarily in the influence of Benedict de Spinoza, a 17th-century Rationalist. In so doing, Jacobi was attacking as an atheist not only Fichte but also F.W.J. von Schelling, an Idealist who was influenced by Spinoza, for his notion of God as nature, as "absolute productivity and the sacred original force of the world." The influence of Spinoza was found again in the thought of G.W.F. Hegel, an early-19th-century Idealist, who wrote that "without the world God is not God," and in the work of Friedrich Schleiermacher, a German Protestant theologian of the same period, who in his later years synthesized his thought in the formula "there is no God without a world, just as there is no world without God." The partial atheism of transcendental Idealism was denounced shortly after Hegel's death in 1831 by many of his followers, especially Ludwig Feuerbach, an anti-Christian German philosopher, who in 1839 appraised the development of modern thought and stated that the duty of the modern age was the humanization of God, the transformation of theology into anthropology. Marxist atheism carried the atheism of Feuerbach to its radical extreme by accusing religion of being the "opium of the people" and the principal cause for the alienation and exploitation of man. The final step in this Germanic atheistic revolution was the "will to power" of Friedrich Nietzsche, a Dionysian thinker of the late 19th century, who openly professed the need for the death of God so that the superman (*Übermensch*) could arise.

20th-century development of atheism

The dominant form of atheism in the 20th century has been "radical humanism," which interprets the being of man solely within the confines of the human, or behavioral, sciences. According to the theory of Sigmund Freud, the founder of psychoanalysis, religion—as the affirmation of the transcendental principle and of the spirituality and immortality of man—is just the result of neurotic frustration. Logical Positivism—according to

which scientific knowledge is the only kind of factual knowledge and all traditional metaphysical doctrines are meaningless—also has arrived at the radical negation of God, either because the Absolute is judged to be a concept without meaning or because it is declared to be unknowable. Existentialism has tended to protest against any force in the face of which human beings are regarded as helpless playthings. In the wake of Nietzsche, the German Existentialists Karl Jaspers and Martin Heidegger emphasized the thoroughly ambiguous nature of religious transcendence, but without denying its importance for man; the atheism of the French Existentialists, however, has been even more radical: according to Jean-Paul Sartre, the idea of God is self-contradictory, and, according to Albert Camus, the affirmation of God involves the negation of human reason.

In 1933 a document called the *Humanist Manifesto* was published in the United States; it was substantially a profession of anthropological atheism based on the theory of evolution. The definition of humanism given by the drafter of the *Manifesto* was rather vague: "Humanism is faith in the supreme value and self-perfectibility of human personality." Among the signatories were several of the leading U.S. philosophers, including John Dewey. The movement grew out of the Pragmatism—the philosophical doctrine that the meaning of conceptions is to be sought in their practical consequences—of William James and others at the beginning of the century. An account of the movement in 1949 fully bore out the diagnosis that the new humanism—which was proclaimed as "the real American philosophy"—was a synthesis of scientism (the theory that the methods used in the natural sciences should be used in all fields of investigation), evolutionism, and vitalism (the theory that life processes possess a unique force different from all other forces found outside living things).

The development of atheism in the Western world, therefore, has followed the ascent of man toward the conquest of earthly ideals in which man himself is considered the subject, the source, and the primary object of values.

Critical examination of atheism. The presence of atheism appears to be coextensive with the history of human civilization. The tension between atheism and theism theoretically constitutes the first and fundamental alternative for man; on this choice depend the individual's understanding of the world and of his destiny and his development of a code of life. According to the Bible, the first generations of men did not practice any form of worship or religion; it was Enosh, son of Seth, who was the first to invoke the name of God. For this reason Giambattista Vico, an 18th-century Italian philosopher of history, spoke of the animal state of the first men. But atheism has appeared at very different levels of culture and has assumed very diverse forms. Thus, in India the *Bhagavadgītā* developed the metaphysical Idealism of the Brahman—the impersonal world spirit—of the *Upaniṣads*, the chief documents of ancient Hinduism, into a sort of theism that includes the idea of creation. But Sāṅkhya (the ancient philosophy of numbers), Buddhism, and Jainism profess an animistic atheism in which the universal energy (*karma*) activates both the soul and matter; and also the Buddhist teaching of Nirvāṇa, the ultimate goal of life, implies the negation of a personal God. The atheism of the ancient Chinese wise men was founded on the absolute moralism professed in the 5th and 6th centuries BC by Confucius and on the notion of the original goodness and dignity of human nature that was expressed almost 200 years later by Confucius' follower Mencius. The other great Confucianist of Chinese antiquity, Hsün-tzu, asserted that moral standards are the creations of society that exert a civilizing influence upon the individual, molding him into a disciplined and morally conscious human being.

Atheism's diverse forms in diverse cultures

In the religion of the Bible, the first duty of man is to know and venerate God; atheism is regarded as the highest folly. According to the author of Psalms 14, "The fool says in his heart, 'There is no God,' " and his foolishness consists in the thought that God cannot reach the sinner while he is committing his fault. Also included as atheists are all polytheistic pagans who worship worthless idols. Even the pessimism found in the book of Ecclesiastes has been accused of atheism, although the author expressly declared faith in God and in his Providence. In the New Testament, St. Paul condemned as stupid and unpardonable all those who, after viewing the world, shut their eyes to the manifestations of God in it and refuse to serve him.

The situation of atheism in the classical Greco-Roman world was extremely ambiguous because of the tension between the anthropomorphic notion of the gods and the notion of God as the transcendent first principle, between the religion of the state with its indigenous divinities and the religion of the Creator of the world, the Father of all men, present to all and yet transcendent. Each understanding of God was accused of being atheistic by adherents of the opposing understanding. Further confusion was raised because some ancient Greek and Roman philosophers professed atheism in their doctrines but in practice acknowledged, or at least did not dispute, the official religion and thus avoided any formal accusation of atheism. But other philosophers, such as Plato and Aristotle, the most important thinkers of the ancient world, criticized the polytheism of the popular religion, did not accept the subordination of philosophy to the religion of the state, and tended toward a monotheism with the transcendent principle; they were accused of atheism and in some cases—an example is Socrates, the great Athenian teacher—were condemned for their impiety.

It is not surprising, then, that the Christians, who refused to acknowledge the religion of the state and adored their God as the only true God and Jesus Christ as the God-man, were accused of atheism. The Christians were considered to be the new enemies of the established religion of the state, and this accusation appears to be the primary reason for their persecution by the Roman emperors until the edict of Constantine in 313 granting toleration for all religions. For their part, the Christians accused of atheism not only the pagans who did not believe in their God but also heretics who defended unorthodox doctrines about God. St. Augustine, the greatest thinker of Christian antiquity, formulated a theological judgment of atheism that embraced the whole course of human history extending to the end of time in his doctrine of the two cities, the city of God (*civitas Dei*) and the city of the world (*civitas terrena*), according to which men are divided into the "faithful," who strive for eternal life, and the "impious," who are "lovers of the world."

Atheism as a purgative for religion

Often the psychological origin of atheism, in its philosophical forms, is a critical reaction to superstition, to the multiplicity and variety of religions, and to the aberrations of magical practices and pseudomystical extravagances. Without doubt, atheism in many of its historical forms has been the purgative for such kinds of religion and has thus been historically an indispensable agency for keeping sound and deepening the knowledge of God. This cathartic function of atheism has influenced the thinking of several important 20th-century Protestant theologians. According to Karl Barth, the Swiss theologian, atheism is parallel to mysticism insofar as it replaces the contents of conceptualistic dogmas with a void in which knowledge and object are one and the same. The starting point of faith and of theology, according to Rudolf Bultmann, a German theologian and New Testament scholar, is the destruction of the concept of God by philosophy so that the thinking of the philosophers, who must ignore the God of the Bible, may then, perhaps, come closer to revealing the divine God. Paul Tillich, a German–U.S. philosopher and theologian, asserts that, in the destruction of the metaphysical God accomplished by modern philosophy, man ultimately encounters the theology of the living God in the form of an absolute faith that is without visible authority and without content and that far surpasses the assent of dogmatic theism or atheism. Therefore, the radical atheists (for example, Nietzsche) become the primary witnesses of God, and atheism, in the judgment of these theolo-

gians, becomes intrinsically impossible. The Roman Catholic second Vatican Council (1962–65), in its constitution on the church in the world, admits that modern and contemporary atheism has been in part a reaction to the conduct of those believers who "neglect their own training in the faith, or teach erroneous doctrine, or are deficient in their religious, moral, or social life, and must be said to conceal rather than reveal the authentic face of God and religion." Atheism is thus seen as a historical fault of Christians because of their infidelity to the precepts of love of God and of neighbour. It was the hope of the council that the shock of realizing that atheism has been the result of the scandal of indifferent Christians could and would provoke a return to authentic Christianity.

Theoretical atheism is bound closely to the particular way the philosopher explains the material and formal causes, or bases, of all beings. These explanations govern the approach he takes to the conception of the One and to the relation between reason and faith. Thus, the material pantheism—that God is the prime matter, or potentiality, of all things (*Quod Deus est materia*)—of David of Dinant, a 12th-century Scholastic philosopher, and the formal pantheism—that God is the formal principle of all things (*Quod Deus est omnia*)—of Amalric of Bene, another 12th-century philosopher at Paris, were considered professions of atheism. Al-Ghazālī, the 11th-century Islāmic theologian, in his *Incoherence of the Philosophers*, complained that "skeptical, nihilistic, and sensualistic philosophers" profess atheism. The same accusation was made against all those—including Averroës, the great 12th-century representative of Islāmic philosophy in Spain—who professed the eternity of the world, thereby implying the existence of uncreated matter, and who denied the freedom of human actions, the divine knowledge of singulars, and the immortality of individual souls. In his response to al-Ghazālī, Averroës did not reject the validity of faith but rather affirmed the primacy of reason over faith. The theories of Averroës—especially that there is only one single intellect, or "intellective soul," for the whole of humanity, that Divine Providence does not reach down to individual men, that there is no individual immortality, and that there is a separation-opposition between reason and faith (the "double truth" doctrine, according to which conclusions in natural philosophy were said to be true, while, simultaneously, conclusions affirming the contrary in theological argument were said to be true)—were preserved in Latin Averroism, a term that designates the thought of a number of Western Christian philosophers who, in the later Middle Ages and during the Renaissance, drew their inspiration from Averroës' interpretation of Aristotle. Latin Averroism was undoubtedly the most significant source of atheism during the Renaissance.

Atheism and its refutation as reflections of their times

As the positions of atheism have varied according to the doctrinal background out of which they developed, so the dialectic of the accusations and condemnations of atheism has been a reflection of the cultural situation of the time. Thus, in 1600 Giordano Bruno, an Italian thinker and writer, whose excursions into many areas of investigation constantly brought him into collision with orthodox opinion, was burned at the stake as an atheist for his pantheistic theory that all life is ultimately derived from matter and for his denial of positive religion. In France, François Rabelais, the 16th-century author of *Gargantua and Pantagrue* (1532, 1534), a bitter satire on the effects of popular religion, and other contemporary authors were accused of atheism. In England, Christopher Marlowe, a dramatist, and Sir Walter Raleigh, an author and explorer, were formally accused of atheism in 1593. Marlowe was brought to trial for blaspheming against the truth of the faith and was killed during the course of the proceedings, although it does not seem that he ever openly professed atheism. There is also no clear proof that Raleigh was an atheist, although the Jesuit pamphleteer Robert Parsons complained of his "School of Atheism" in 1592. Raleigh was, however, an enthusiastic reader of skeptical philosophy. There was more substantial argument for the accusation of atheism brought against Thomas Hobbes, a 17th-century British political philosopher. In *Leviathan*, his masterpiece, Hobbes declared that the traditional doctrines of the divine Logos, or wisdom, and of the divine attributes were nonsense and condemned religion as superstition intended to defend the laity from true moral and political institutions.

Critical judgments against modern atheism were directed at the starting points of the speculative systems of the philosophers. Thus, in the 17th century both the Jesuit historian Jean Hardouin and the Calvinist theologian Gisbertus Voetius accused Descartes and his followers of atheism because they conceived of God as *ens infinite perfectum* ("infinitely perfect being") and pure universal form. Spinoza was accused of atheism because he identified God, the one and only substance, with the world. The classical text of rational atheism is considered to be the *Ethics* of Spinoza, who is also considered to be the founder of modern biblical interpretation and demythologization—uncovering the meaning underlying the literary and mythological forms of the Bible. The works of many 17th- and 18th-century English clergymen and theologians, especially Samuel Clarke and Joseph Butler, were aimed at Deism, which denied the unique position of the Judeo-Christian tradition as a divine revelation. In an effort to halt the rising tide of disbelief and atheism, Robert Boyle, a British physicist and chemist who died in 1691, provided in his will for the Boyle Lectures for the defense of Christianity "against those who are notoriously unbelievers, such as the atheists, the deists, the pagans, the Jews, and the Mohammedans." The lecture series, which is still continued, was begun in 1692 by Richard Bentley, a clergyman and classical scholar, who delivered a "refutation of atheism," arguing especially against that form of cryptoatheism that had its starting point in the notion of "thinking matter" and questioning the return to pagan morals and the rejection of the Christian message of salvation.

The anthropological origin of atheism

Although every form of atheism can be traced back to a particular ideology (for example, the cosmic form of the Renaissance, the ethical form of the Deistic Enlightenment, the anthropological form of modern Idealism), in reality atheism always has an anthropological origin, or a tendency to found truth and the value of existence within man. Modern radical atheism is founded on the principle of immanence: human subjectivity—the presence of the self to itself—is the foundation of the truth of being. The atheisms of Marxism, of Existentialism, of Logical Positivism, and of Pragmatism can all be considered as precipitating causes of the modern principle of immanence. But all immanentistic concepts radically eliminate the possibility of the Absolute and thereby also empty man's being of its originality by reducing man to a simple function of his historical situation. Therefore, there is a return to discussion of the need for a "resurrection of metaphysics," or the need for consideration of some transcendent principle, in the writings of Heidegger.

There is also a need to distinguish philosophy from science and even more from morality and religion. According to Positivistic, humanistic, and especially Marxistic atheism, scientific and technological progress (particularly the exploding knowledge about outer space) has destroyed the need for religion. It can be answered, however, that the progress of science has brought man to a deeper and more precise knowledge of the magnitude of the universe and of the complexity of its laws (especially in the fields of microphysics and microbiology); but according to such eminent 20th-century physicists as Max Planck, Albert Einstein, and Werner Heisenberg, man merely discovers these laws, and they presuppose an infinite intellect. The resources capable of catastrophic effects that modern technology has put at man's disposal render more pressing the need to affirm the transcendence of moral and religious values if man wishes to help and not to destroy his own species.

Furthermore, the foundation of the concept of freedom has come to be understood as demanding absolute au-

tonomy. According to Existentialism and secular humanism in their multiple forms, the negation of theological transcendence is necessary in order to guarantee the absolute character of freedom; so state such thinkers as Dewey and Sartre. This radical existential freedom, however, has real meaning only if it is based on each individual's ability to choose his own happiness according to an absolute criterion of good and evil. According to the Christian Existentialist Søren Kierkegaard and the Jewish thinker Martin Buber, such a choice means being and putting oneself "before God."

The problem of evil

Finally, the existence of evil has been at all times the strongest argument on the existential level in favour of the denial of God. Evil refers either to physical evils (such as natural disasters and cataclysms, poverty and war, sickness and death) that deprive man of sensual happiness or especially to moral evils (such as errors and injustices, betrayals and hostilities, hate and violence) that oppress the spirit of man. The horror of evil is that it often strikes the weak and the innocent. How are such phenomena compatible with the existence of an omnipotent and good God? To this question the theist observes that the denial of God and of immortality does not change the sorrow and suffering but only aggravates the situation by depriving the suffering individual of any hope of liberation from evil and by depriving the innocent individual of any right to obtain justice from an absolutely perfect and independent judge, whom men have called God. Christianity further claims that sorrow and suffering have been sanctified by the example of the Passion and death of Jesus Christ.

During the 1960s, in a movement arising under the convergent influence of the German theologians Paul Tillich and Dietrich Bonhoeffer, a "theology without God" was proclaimed. According to one of the theologians espousing it—most of whom are of the English-speaking world—the "death of God" is not a metaphysical category but a cultural one; contemporary man, in this view, is just not fitted to understand any real repentance through an exterior principle like the transcendent God: "Bonhoeffer invites us to accept the world without God as given and unalterable." According to the same author the second idea that marks Bonhoeffer's influence and importance is his plea for a nonreligious or religionless Christianity. Because the world has grown up and transcended its dependency situation, "the God of religion, solving otherwise insoluble problems, meeting otherwise unmeetable needs, is impossible and unnecessary." A last and more profound ground of this "theological atheism" is the kenotic movement, according to which the God who died in Christ is the God who thereby gradually ceases to be present in living form, emptying himself of his original life and power. According to another of the "death of God" theologians, Thomas J.J. Altizer, the death is, thus, "an inevitable consequence of the movement of God into the world, of spirit into flesh."

The possibility of universal atheism and even the atheism of large masses in wealthy modern society have placed all men really for the first time face-to-face with the basic alternative of being either with God or against God. But even for sophisticated modern man the decision is complicated by the danger of nuclear catastrophe, which threatens to destroy all the values of civilization, man himself, and his cosmos.

BIBLIOGRAPHY

Notion of atheism: Encyclopaedia of Religion and Ethics, vol. 2, pp. 173–190 (1928); *Reallexikon für Antike und Christentum*, vol. 1, pp. 866–870 (1950); *Die Religion in Geschichte und Gegenwart*, vol. 1, pp. 670–678 (1957); G. KLAUS and M. BUHR, *Philosophisches Wörterbuch*, vol. 1, pp. 125–129 (1969); and *Enciclopedia filosofica*, 2nd ed., vol. 1, col. 557–562 (1968), are five expositions indispensable for viewing atheism from the point of view of the history of religion, the ancient Christian culture, contemporary theology, and Marxist and contemporary philosophy. J.M. ROBERTSON, *A Short History of Freethought*, 3rd ed. rev., 2 vol. (1915); L. STEPHEN, *A History of English Thought in the Eighteenth Century*, 2nd ed., 2 vol. (1902), are two fundamental works to follow the development of atheism in the various modern philosophical schools. E.S. BRIGHTMAN, *A Philosophy of Religion* (1946); M. SCHELER, *Vom Ewigen in Menschen* (1954; Eng. trans., *On the Eternal in Man*, 1960); and *Philosophische Weltanschauung* (1954; Eng. trans., *Philosophical Perspectives*, 1958), are important studies for a psychological analysis of atheism.

History of atheism: F. MAUTHNER, *Der Atheismus und seine Geschichte im Abendlande*, 4 vol. (1920–23), contains the story of atheism in the Western world, according to the traditional criteria of the positivistic method. See also J. PRESSER, *Das Buch "De tribus impostoribus"* (1926), a primary source for the background and development of modern atheism; H. BUSSON, *Le Rationalisme dans la littérature de la Renaissance (1533–1601)*, new ed. (1957), a fundamental study of atheism from the Middle Ages through modern times; and C. FABRO, *Introduzione all'ateismo moderno* (1964; Eng. trans., *God in Exile: Modern Atheism*, 1968), a critical and speculative analysis of the atheistic essence of modern thought.

Representatives of contemporary atheism: JOHN DEWEY, *A Common Faith* (1934); R. ROBINSON, *An Atheist's Values* (1964); and R. GARAUDY, *Dieu est mort* (1962), represent contemporary secular atheism. DIETRICH BONHOEFFER, *Widerstand und Ergebung* (1951; Eng. trans., *Letters and Papers from Prison*, 1953); T.J.J. ALTIZER and W. HAMILTON, *Radical Theology and the Death of God* (1966); J.A.T. ROBINSON, *Honest to God* (1963); L. DEWART, *The Future of Belief* (1966); THOMAS W. OGLETREE, *The "Death of God" Controversy* (1966); W. HAMILTON, *The New Essence of Christianity* (1966); and G. VAHANIAN, *The Death of God* (1966), are representatives of so-called theological atheism. E. BLOCH, *Atheismus im Christentum* (1968), is an exposition of Marxist atheism; and C. FABRO, *L'uomo e il rischio di Dio* (1967), is a critical evaluation of these modern responses to the problem of God.

(C.F.)

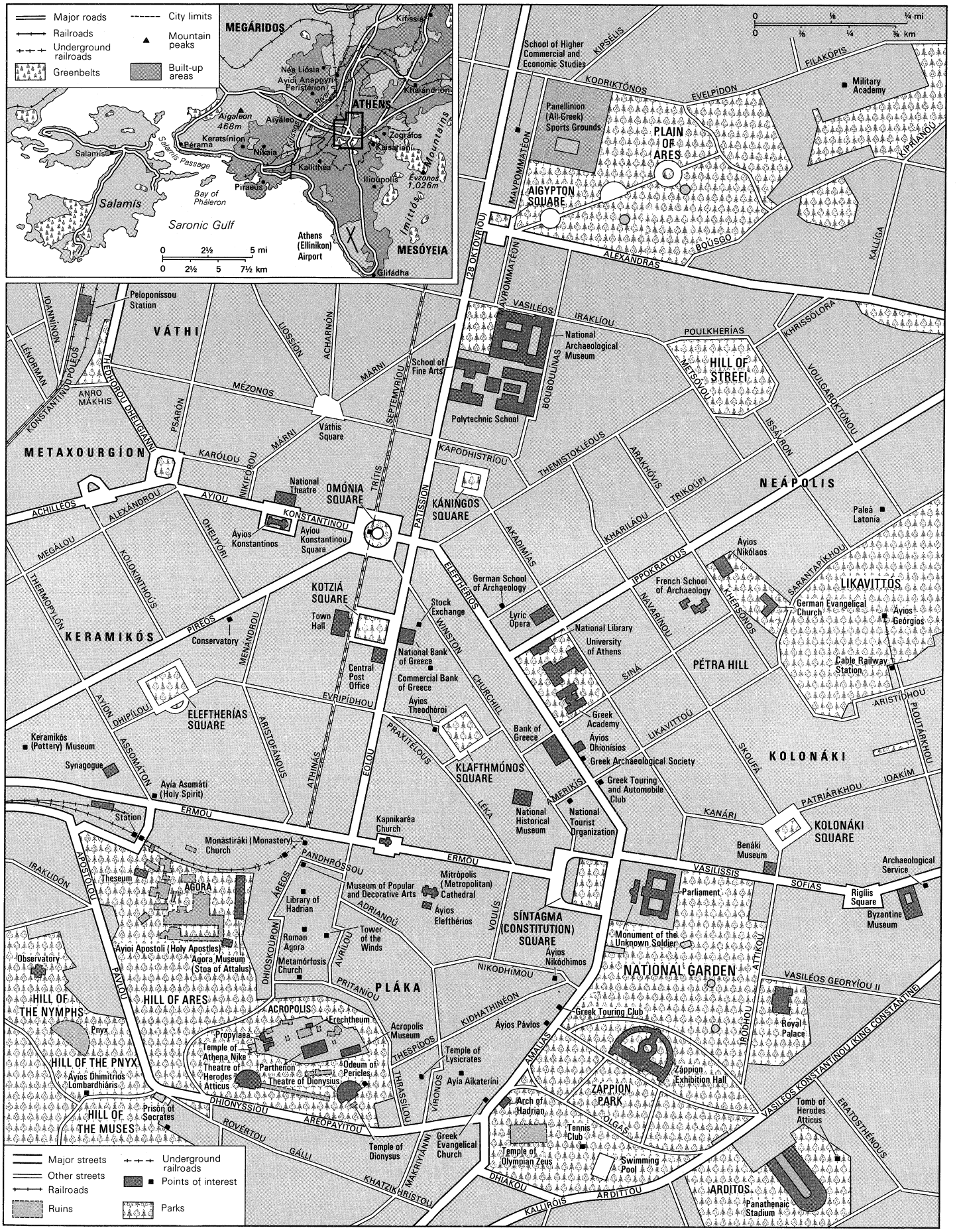
Athens

The capital of the kingdom of Greece and generally considered the nursery of Western civilization, Athens (Greek Athínai) lies five miles from the Bay of Phaleron, an inlet of the Aegean (Aigaion) Sea where Athens' port, Piraeus (Piraiévs) is situated, in a mountain-girt arid basin divided north–south by a line of hills. Greater Athens has an area of 167 square miles (433 square kilometres) and was the home of a little more than 2,500,000 people in the early 1970s. The central city, the old municipality, had over 862,000 inhabitants. The Kifisós River, a trickle in summer, flows through the western half; and the Ilisós River, often actually dry, traverses the eastern half, while the surrounding mountains of Párnis, 4,635 feet (1,413 metres), Pentelicus (Pendéli), 3,631 feet (1,107 metres), Hymettos (Imittós), 3,365 feet (1,026 metres), and Aigáleon, 1,535 feet (468 metres), add to the impression of barrenness. Yet such considerations are superficial when compared with the fecundity of Athens' spiritual bequests to the world, such as its philosophy, its architecture, its literature, and its political ideals.

THE ATHENIAN LEGACY

The Acropolis. Many of these concepts (all, if the theatre of Herodes Atticus may be regarded as an embodiment of the city's literature) are expressed in and around the Acropolis, the natural focus of Athens. Rising some 500 feet above sea level, with springs near the base and a single approach, the Acropolis was an obvious choice of citadel and sanctuary from earliest times. That it could be something more is evidenced in the Parthenon, one of the brightest jewels in mankind's treasury, let alone Athens'. As deceptively simple as Socrates' conversation, this columned, oblong temple is the expression—without a trace of strain or conflict—of a human ideal of clarity and unity. The architectural genius is concentrated in the exterior, for within was a shelter for the goddess Athena—the patroness who lent her name to the city—not a place for mass worship. Its spiritual quality, the sensation of being almost afloat, is enhanced by the lack of a single, straight, vertical line in the peristyle (the colonnade surrounding a building or court); each vertical is almost imperceptibly bowed, theoretically meeting some 11,500 feet up in the sky. The columns, of diminishing thickness toward the centre of the colonnade, with diminishing space between them, lean toward the centre,

The Parthenon



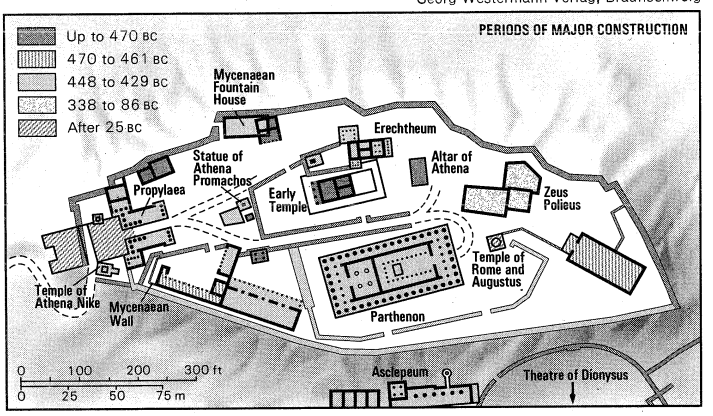
Central Athens and (inset) its metropolitan area.

too; all these differences are virtually invisible to the beholder. Even the 20 flutings (grooves) of each column diminish in width as they rise, and the humblest details of craftmanship are perfect.

On the northeast corner of the interior are faint traces of Christian wall paintings, dating from the temple's service as the Church of St. Mary, and in the southeast corner of the porch is the stair leading to the minaret that poked through the roof when the building was a Turkish mosque. The Parthenon was also used as a powder magazine, when, on September 26, 1687, Venetian artillery, attacking the Turks from the Hill of the Muses, scored a direct hit. A member of the party with the field commander, General Königsmark, wrote, "How it dismayed His Excellency to destroy the beautiful temple which had existed three thousand years!" Conversely, Francesco Morosini, the commander in chief, when reporting to the Venetian government, called it "a fortunate shot." Wishing to bring home more than just good news, he also tried to lower Athena's horses in the centre of the west pediment, but his men's dexterity was not as highly developed as their marksmanship and the masterpieces smashed to bits on the rock below.

The Turks regained possession of the Acropolis the following year, and later began selling "souvenirs" to Europeans. The Duc de Choiseul, lately French ambassador in Constantinople, picked up a piece of the frieze and two metopes (space between two triglyphs, or ornaments in the frieze). Fourteen years later the British ambassador, Lord Elgin, arrived with an imperial decree permitting him to pull down Turkish houses on the Acropolis to seek fragments of sculpture. Among the 50 pieces he took home (the shipping charges were £75,000, a huge sum for those days) was most of the remaining Parthenon sculpture, which he later sold to the British Museum for £35,000. The Greeks have forgiven the clumsiness of the Venetian engineers, the accuracy of her cannoneers, and the vandalism of the Turks, but still nurture rancour against Elgin. His lordship also removed one of the caryatids (draped female figures serving as columns) from the Erechtheum, a temple of Athena called after a shrine dedicated to the legendary King Erechtheus or to Poseidon Erechtheus, but replaced her with a plaster cast. From London he graciously sent a town clock for Athens, duly erected in the Agora (marketplace) and lost in the fire of 1885.

Adapted from *Westermann Grosser Atlas zur Weltgeschichte:* Georg Westermann Verlag, Braunschweig



Plan of the Acropolis, Athens.

The Erectheum (5th century BC), which later became a church under the Byzantines and subsequently the Turkish commander's harem, was originally dedicated to both Athena and Poseidon and was the most venerated of the Acropolis temples.

The Propylaea, the matchless entryway into the Acropolis, was the only opening in the surrounding wall. Just in front of it and to the left is the 27-foot-high pedestal for the thank-offering to Agrippa, the victor of the Battle of Actium, who interceded for Athens, which had supported the loser, Mark Antony. To the right was the temple of Athena Nike (Giver of Victory), 27 feet long and 18½ feet wide, which stood untouched until the Turks demolished it in 1686 to use the stones as defenses against the Venetians. In 1836 it was badly restored, and 100 years later its foundations began to slip into previously undiscovered Turkish cisterns, revealing the 6th-century foundations, of Peisistratus' earlier shrine to Artemis Epipyrgidea (Artemis on the Tower). The temple was then more accurately reconstructed. The northern wing of the Propylaea, the Pinakotheke, was used by the Frankish dukes, who reconstructed the interior to make a two-story building. In the 12th century, Greek Orthodox bishops lived in the Pinakotheke, and in the 14th century, the Acciajuoli dynasty of Athenian dukes from Florence turned the Propylaea into a fortified castle with a Tuscan tower, which Heinrich Schliemann, the German archaeologist who discovered Troy, paid to have dismantled in 1875.

The Pinakotheke

When the Turks, who had occupied Athens since 1456, departed, they left the monuments in a state of ruin, the ground covered with garden plots, and 300 small huts. After Greece won its independence, Otho, the first king of the Hellenes, had everything that postdated the classical period swept away, set scholars to work identifying the remains, and encouraged some reconstruction.

According to Pausanias, the Greek traveller and geographer of the 2nd century AD, the colossal, 30-foot-high bronze seated statue of Athena Promachos (Athena Who Fights in the Foremost Ranks), by the 5th-century Athenian sculptor Phidias, was set up in the open behind the Propylaea, her gleaming helmet and spear visible to mariners off Cape Sunium (30 miles away). The 6th-century Byzantine emperor Justinian carried her off to Constantinople, just as Phidias' ivory and gold statue of Athena had been taken from the Parthenon by his predecessor Theodosius II. Both of these masterpieces were lost to other looters in the Crusaders' sack of Constantinople in 1204. Other statues stood in profusion amid small temples, such as the sculptor Myron's group of Marsyas and Athena, his Perseus, and his heifer; Phidias' Lemnian Athena and his Pericles; and a gigantic bronze effigy of the Trojan horse. There was an altar to Athena Hygeia (the Health Giver), a precinct sacred to the goddess Artemis Brauronia (named after a statue of her, brought from the town of Brauron), the Pandroseum (a building named after Pandrosos, a girl associated with Athena in legend), where the sacred olive tree of Athena grew, and beyond the Parthenon the great altar of Athena.

Other notable buildings. Below the Acropolis sanctuary, on the southwest slope of the hill, Herodes Atticus, a rich Roman, built a 5,000-seat odeum as a memorial to his wife in AD 161. A conventional Roman theatre, save that the semicircular auditorium was hollowed out of the rock, it was roofed in cedar and had a three-story facade of arches. Repaired but roofless, it is now used for the Athens summer festival of music and drama. A 300-yard-long portico stretching toward the theatre of Dionysius had been built some 300 years earlier. The Dionysiac theatre itself, scooped out of the south slope early in the 5th century, replaced the Agora stage as the drama centre. It also replaced the Pnyx as the meeting place for the popular assembly. Rebuilt many times, the ruined theatre now visible is largely Roman, the last construction work on the stage probably dating from the early 3rd century AD. The Dionysia, the spring festival, which drew crowds from many parts of Greece and colonies in Asia Minor and Italy, was held in this theatre, which had 13,000 seats in 67 rows. The jury had larger front seats and the ecclesiastical dignitaries small stone thrones, on which their titles can still be read. Three tragic and four comic plays were presented in competition for the prize. Production costs were met by private sponsors who, when their choruses won the prize tripod, displayed it in an elaborate memorial in the Street of Tripods to the east of the theatre. The only one of these monuments still standing is that of Lysicrates, erected 335/334 BC, a small circular temple 21½ feet high, its six columns an early example of the Corinthian order. Preserved through incorporation into a convent (in which the English poet Lord Byron had a study), the monument influenced British Georgian and Regency architecture through the engravings of the Edin-

burgh artist "Athenian" Stuart. Farther east lay the Odeum of Pericles, and to the west are traces (420 BC) of the precinct of Asclepius, the god of healing, which took the form of a hospital portico for patients and temples decorated with votive reliefs.

On the Hill of Ares, the god of war, to the right of the descent from the Propylaea, a legendary jury of gods spared Ares from execution for the murder of the sea god Poseidon's son. Trials for homicide continued to be heard on this hill through the ages, and the Supreme Court of Greece still bears the name.

Other hills in Athens

Across Apostólou Pávlou (Apostle Paul Avenue) are the Hill of the Nymphs, where an Austro-Greek, Baron Sina, built an observatory in 1842; the Hill of the Muses, crowned with the remains of the marble monument to Philopappus, a Syrian who was Roman consul in the 2nd century AD; and the middle hill, the Pnyx, meaning "tightly crowded together," the meeting place of the Ecclesia, the assembly of 18,000 citizens who heard the great Athenian orators. (In fact, attendance of over 5,000 was rare, but it would still have been crowded.)

The Agora. The avenue leads down to the Agora, which the American School of Classical Studies started restoring in 1931, paying $2,500,000 compensation to the several hundred families living there. Financed by, among others, the Rockefeller Foundation, the Marshall Plan, and the Greek government, the work went on until 1960. It includes what has been called "the pitiless replica of a 180-columned portico of the 2nd century BC," which serves as a museum.

At the approaches to the Agora is the best preserved of all Greek temples, the Theseum (5th century BC). Although virtually intact and absolutely genuine, it has all the deadness of a latter-day reproduction. The beauty, the mystery, and the genius that render the Parthenon incandescent eluded the architects and builders of the Theseum.

The Horologium and the Orthodox cathedrals. Another monument is the octagonal, 42-foot-high marble Horologium of Andronicus of Cyrrhus, usually called the Tower of the Winds because each side bears a weather-beaten figure of the wind from that particular compass point. It used to have a sundial, a water clock for telling the hour on cloudy days, and a weather vane. The Turks left it unchanged, believing it to be the tomb of two local prophets, Sakhratis and Aflatun (Socrates and Plato).

In the shadow of the 19th century, Neo-Byzantine Greek Orthodox Cathedral (Mitrópolis) nestles the old Mitrópolis, Áyios Elefthérios, one of three genuine Byzantine churches still surviving. It is red brick like the others and tiny, its Pentelic marble ruddied with age, its outer walls artfully, if promiscuously, decorated with classical Greek tidbits: panels, votive tablets, and morsels of frieze. Like its sisters, this retired cathedral is charming, unassuming, and comforting. (B.E.)

THE EARLY HISTORY OF ATHENS

The city in antiquity. *Factors inducing settlement.* The climate of Athens is benign: frost is rare (the minimum temperature is 32° F, 0° C) and snow seldom lies, while the summers, though hot (maximum temperature is 99° F, 37° C), are dry, and a fresh northeasterly wind often blows by day. The nights are always cool. All of this permits outdoor activity the year around and has had an important effect on both the style of architecture and the life and political institutions of the city.

Evidence of early inhabitation

The site of Athens has been inhabited since the Neolithic Period (before 3000 BC). Evidence for this has come from pottery finds on and around the Acropolis but particularly from a group of about 20 shallow wells, or pits, on the northwest slope of the Acropolis, just below the Klepsydra spring. These wells contained burnished pots of excellent quality, which show that even at this remote period Athens had a settled population, with high technical and artistic standards. There are similar indications of occupation in the Early and Middle Bronze ages (3000–1500 BC).

The earliest buildings date from the Late Bronze Age, particularly about 1200 BC when the Acropolis was the citadel. Around its top was built a massive wall of Cyclopean masonry (a type of construction using huge blocks without mortar). The construction of this wall probably marks the union of the 12 towns of Attica (the department in which Athens lies) under the leadership of Athens, an event traditionally ascribed to Theseus. The palace of the king was in the area of the later Erechtheum, but almost no traces of it have been identified. The town, insofar as it was outside the Acropolis, lay to the south, where wells and slight remains of houses have been found. The principal cemetery lay to the northwest, and several richly furnished chamber tombs and many smaller ones have been discovered in the area that later became the Agora.

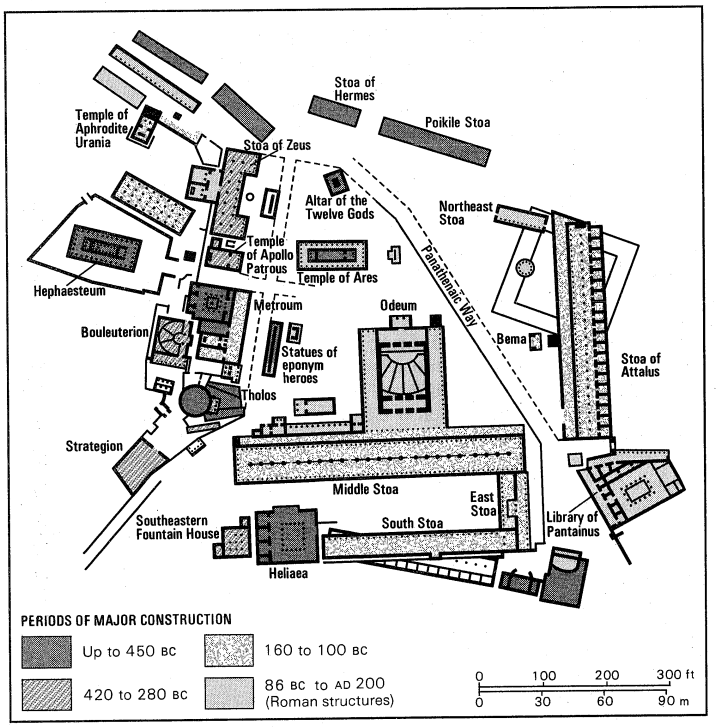
Whether through the strength of its walls, the valour of its citizens, or its geographical position away from the main route to the Peloponnesus, Athens seems to have weathered the Late Bronze and Early Iron ages, troubled times, better than other, more important centres. There is no evidence of complete or widespread destruction, as at Mycenae and Pylos; in fact, the pottery styles show an unbroken development through sub-Mycenaean (a style later than the Mycenaean but not yet Greek) to Protogeometric (the earliest phase of Geometric) and Geometric (a style of pottery in vogue from 1000 BC to 750 BC). Furthermore, there is positive evidence that from about 1000 BC the city began to expand in a northwesterly direction, into the area that had previously been confined to cemeteries. Wells appear, indicating occupation by the living, and any graves in the area are increasingly confined to restricted plots or placed along the roads outside the town limits. The Agora and some of the public buildings seem, to judge from scattered notices in later writers, to have been located west and northwest of the Acropolis. Though there are few remains of buildings, the wealth and prosperity of the city can be appreciated from late Geometric graves found in the area of the later Dipylon and Erian gates. These graves were adorned with large vases, sometimes over five feet high, richly decorated with geometric patterns and having figured scenes of battles, processions, and funeral ceremonies.

Changes in the 6th century BC

Athens' expansion. The 6th century BC was a period of phenomenal growth, particularly during the tyranny of Peisistratus and his sons (*c.* 560–510 BC). On the Acropolis, the old primitive shrines began to be replaced with large stone temples. About 580 BC a temple to Athena, known as the Hekatompedon, or "Hundred-Footer," was erected on the site later to be occupied by the Parthenon. The pediments (triangular spaces forming the gable) of this temple were decorated with large-scale sculpture in gaily coloured, porous limestone, representing groups of lions bringing down bulls, and with snaky-tailed monsters in the angles. These sculptures are now displayed in the Acropolis Museum. In 566 BC Peisistratus reorganized the Panathenaic Games in honour of Athena on a four yearly basis. About 530 BC a large peripteral temple (one having a row of columns on all sides) to Athena Polias (Guardian of the City) was erected near the centre of the Acropolis, on the site of the old Bronze Age palace. It had marble pedimental sculpture representing the battle of the Gods and Giants. Besides these two major temples there were five smaller buildings, treasuries and the like, and a wealth of votive offerings in marble, bronze, and terra-cotta. The Acropolis was now a full-fledged sanctuary.

This change from citadel to sanctuary is also reflected in the arrangement of the entrance at the west. Instead of a winding path suitable for defense, there was, from about the middle of the 6th century BC, a broad ramp, designed as a ceremonial approach, leading up to the gate. This basic change of attitude toward the Acropolis must mean that the whole lower town was now surrounded by a fortification wall, and the Acropolis was no longer needed for defense. The ancient historians Herodotus and Thucydides tell of such a wall, but no trace of it has been found, and its course and date are uncertain.

In the lower town, too, the 6th century was a period of great growth and change. The old Agora, below the western approach to the Acropolis, was now inadequate, and



Ground plan of the Athenian Agora.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte:* Georg Westermann Verlag, Braunschweig

Building of the new Agora

a new one was therefore laid out in the low ground to the northwest. This was accomplished by demolishing houses and filling in wells and gullies, to create a broad, open square, which was used for gatherings of all sorts: political, judicial, religious, and commercial. Dramatic contests were held here, too, before the construction of a separate theatre. Various public buildings and shrines were erected around the borders of the square, including the Basileios (Royal) Stoa, where the archon Basileus, one of the chief magistrates of the city, had his headquarters, the Old Bouleuterion (or Council House), and a large enclosure (100 feet square), which probably housed the Heliaia, the largest of the popular lawcourts. At the southeast corner of the square a small fountain house was built, which received water from outside the city through a conduit of terra-cotta pipes.

In 480 BC this flourishing city was captured and destroyed by the Persians. The Acropolis buildings were burned, and the houses in the lower town mostly destroyed, except for a few that had been spared to house the Persian leaders.

Athens at its zenith. When the Athenians returned, in 479 BC, they immediately rebuilt their fortification wall larger than before. About 20 years later the famous Long Walls were built, connecting the city with its port, Piraeus, four miles away. They were parallel over most of their course, forming a corridor 550 feet wide. These walls played a vital part in Athens' history during the Classical period, for they allowed it to carry the supplies brought in by its powerful fleet in safety to the city, even when enemy forces roamed the Attic countryside.

For 30 years after the Persian destruction, the Athenians built only fortifications and some secular buildings in the Agora, notably the Stoa Poikile, or Painted Colonnade, with its famous paintings by Polygnotus and Micon, one of which represented the Battle of Marathon. The Tholos, the round building that served as the headquarters of the executive committee of the council, was also built at this time. Lack of attention to the Acropolis was partly the result of the oath, sworn before the Battle of Plataea in 479 BC, that sanctuaries destroyed by the Barbarians would not be rebuilt but left as memorials of their impiety. In 449 BC, however, peace with Persia was at last officially established, and the oath was annulled. Athens, moreover, now had ample funds, for the silver mines in the Laurium (Lavrion) Hills of southern Attica were in full production. These mines had always been exploited, but in 483 BC a big strike was made, the proceeds of which were used to build the ships that won the battle of Salamis, in 480 BC. Thereafter, the mines remained productive throughout the 5th and 4th centuries, providing Athens with the sinews of its strength in the great Classical age. Another source of revenue was the tribute that the allies had been paying, as members of the Delian League, to prosecute the war against Persia. Athens had been collecting and administering this money and now, even though the war was officially over, continued to collect it in spite of the protests of the allies, who degenerated into subjects of Athens. Pericles deemed it proper, over the protests of his opponents, to use this money on beautifying the city; in this way he could keep the money in circulation and provide jobs for the whole population. Thus began one of the largest and most enduring works programs in history.

Peace with Persia

In a period of 40 years the Acropolis was entirely rebuilt in gleaming white marble quarried from Mt. Pentelicus, ten miles north of the city. The first great work was the Parthenon, begun in 447 BC and finished, except for some details, in 438 BC. The architects were Ictinus and Callicrates, and Phidias was in charge of the whole artistic program. The building was considerably larger than was usual, having eight columns across the ends and 17 on the long sides, against six by 13 for the average temple. It was richly decorated with sculpture, having a running frieze all around the top of the cella (the walled-in chamber within the colonnade) wall outside, and sculptured metopes and sculptured pediments. Inside the cella stood the cult statue, the great gold and ivory figure of Athena, the work of Phidias. No sooner was the main work on the Parthenon completed than the Propylaea was begun. This was the monumental gateway with five doors at the head of the approach, designed by the architect Mnesicles. Its large outer vestibule was covered by a marble ceiling, supported by marble beams with a free span of 18 feet, about which Pausanias wrote, "the Propylaea has a ceiling of white marble which in the beauty and size of the stones remains supreme even to my time." Work on the Propylaea was nearly finished when stopped by the outbreak of the Peloponnesian War in 432 BC, but, as things began to go well for Athens, the little temple of Athena Nike was erected on the bastion in front of the Propylaea, perhaps in 425 BC. At the time of the Peace of Nicias (421 BC), or perhaps a year or two earlier, the Erechtheum was begun. This was a small Ionic temple of highly irregular plan, which housed various early cults and sacred tokens. When the building was about half-finished, work was suddenly interrupted, probably because of the disastrous Athenian expedition to Sicily (415–413 BC), but it was resumed in 409 BC, and the building was completed in 406 BC. The final defeat of Athens two years later put an end to all building, but the Acropolis had been completed, and in later centuries only secondary buildings and monuments were added.

The building of the Erechtheum

The second half of the 5th century also saw some building activity in the lower town. Even before the Parthenon, work was begun on the temple of Hephaestus (the Greek god of fire), the Theseum, which still stands on a low hill. In the Agora itself, a new Bouleuterion was built, and two colonnades, the Stoa of Zeus and the South Stoa, were constructed. On the south slope of the Acropolis, next to the theatre, Pericles built an odeom, a large enclosed concert hall, its roof supported by a forest of columns. Of the theatre itself there are no identifiable remains, but the arrangements were no doubt quite simple, and it is known that a theatre existed on this spot from the late 6th century BC because of the old temple of Dionysus (the god of wine) nearby, which dates from the same period. A sanctuary of Asclepius was founded on the south slope of the Acropolis in 420 BC.

Athens was slow in recovering from its defeat in the Peloponnesian War, but in 394 BC its admiral, Conon, won a decisive naval victory over Sparta off Cnidus, on the west coast of Asia Minor. As a result he rebuilt the Long Walls, which the Spartans had demolished to the

music of flutes ten years before, believing they were inaugurating the freedom of Greece. The walls of Piraeus were also rebuilt, and those of the city were repeatedly strengthened in the course of the 4th century, notably by the addition of a ditch or moat as protection against siege machinery.

Building in the late 4th century

Apart from military works, there was little building in 4th-century Athens until the years 338–322 BC, when the orator Lycurgus was in control of the state finances and there was great activity. On the Pnyx, the broad-backed hill west of the Acropolis where the Athenian popular assembly had met since the reforms of Cleisthenes (a 6th-century constitution maker), a large auditorium was constructed. At the same time, two large stoas were started on the terrace above. The theatre of Dionysus was rebuilt and greatly enlarged, with stone seats to accommodate the crowds. (Lycurgus did another service to the theatre by having definitive copies made of the old plays.) The Panathenaic stadium was also built about then, partly with state funds and partly by private contributions: the land was donated by a certain Deinias, and one Eudemus of Plataea provided 1,000 yoke of draft animals to level the ground. The period was one of lavish private expenditure in other fields as well. The tripods won in choral contests were displayed on elaborate monuments, sometimes even resembling small temples: the best preserved of these is that of Lysicrates (334 BC), a small round building with six Corinthian columns. Tombs also became increasingly elaborate, often portraying the whole family in high relief. In 315 BC a stop was put to all this extravagance by the sumptuary laws of Demetrius of Phalerum.

Meanwhile, the philosophy schools flourished. Plato (*c.* 428–348/347 BC) established himself in the Academy, a gymnasium that had existed since at least the 6th century BC in the great olive grove about a mile west of the city. Plato himself had a house and garden near by. Aristotle and his Peripatetics occupied the Lyceum, another gymnasium, just outside the city to the east, and his successor Theophrastus lived near by. Antisthenes and the Cynics used the Cynosarges gymnasium to the southeast of the city. Zeno held forth in the heart of the city, in the Stoa Poikile, or Painted Colonnade, in the Agora, and his followers were therefore known as Stoics. Epicurus and his followers had a house and garden in town.

Apart from its temples and public buildings and its great avenues, however, Athens seems to have made a poor impression. A 3rd-century-BC visitor complained that the city was dry and ill-supplied with water, that it was badly laid out because of its great antiquity, and that most of the houses were mean. The streets were in fact narrow and winding, and the houses, it is true, presented a blank wall to the street except for the entrance door, but then they were built around a central courtyard, off which the various rooms opened. There was often an upper story, and the court had a well. Water brought in by the aqueducts was not considered good because it was hard (containing salts of magnesium or calcium) and caused rheumatism. Waste water was carried off in an elaborate system of underground drains beneath the streets. (For related information, see GREEK CIVILIZATION, ANCIENT.)

Hellenistic and Roman times. Athens in Hellenistic and Roman times depended for its embellishment less on its own resources than on the generosity of foreign princes. One of the Ptolemies (rulers of Egypt) gave a gymnasium, erected near the sanctuary of Theseus, and the Ptolemies were probably also instrumental in the founding of the sanctuary of the Egyptian gods Isis and Sarapis. More important were the donations of the Attalids of Pergamum (a dynasty of Asia Minor); Eumenes II (197–159 BC) gave a large, two-story colonnade on the south slope of the Acropolis near the theatre. His brother Attalus II (159–138 BC), who had studied at Athens under the philosopher Carneades, head of the New Academy, likewise gave a colonnade. This was a large, elaborate, two-story building more than 350 feet long with a row of shops at the rear. It was located on the eastern side of the Agora and has been reconstructed in modern times (1953–56) to serve as the Museum of the Agora excavations. The Stoa of Attalus was the first element in a large-scale reconstruction of the Agora. It was followed in quick succession by three buildings, the Middle Stoa, the East Building, and the South Stoa, which together formed a separate South Square in the southern half of the Agora.

Donations of the Attalids

The capture of Athens by the Roman general Sulla in 86 BC was accompanied by great slaughter and much destruction of private houses, but the only public building to be destroyed was the Odeum of Pericles, burned by the defenders of the Acropolis lest its timbers be used by the enemy. The Odeum was rebuilt on the same plan a few years later, through the generosity of King Ariobarzanes of Cappadocia.

Under the Roman Empire, Athens enjoyed imperial favour. A spacious market for the sale of oil and other commodities was laid out east of the old Agora with funds originally provided by Julius Caesar and supplemented by the emperor Augustus. In the old Agora itself, a new odeum, or concert hall, was built in the middle of the square by Marcus Agrippa, the emperor's son-in-law and one of his chief lieutenants. A large building, perhaps a law court, was also erected at the northeast corner. At the southeast corner of the Agora a handsome library was erected about AD 100, the gift of one T. Flavius Pantainus and his family. It was decorated with a group of marble sculpture representing Homer flanked by the Iliad and the Odyssey. On the Acropolis a small round temple was erected to the goddess Roma and the emperor Augustus.

Hadrian's devotion to Athens

The emperor Hadrian (AD 117–138) completed the great temple of Olympian Zeus, started over 600 years earlier by the Peisistratids. This temple formed the chief ornament of the new eastern suburb of Athens, and Hadrian gave the area a monumental entrance through a gateway, the inscriptions on which proclaimed, on one side, "This is the Athens of Theseus, the old city," and, on the other, "This is the city of Hadrian, not of Theseus." Hadrian also built a library, a gymnasium, and a Pantheon (a sanctuary of all the gods). His aqueduct, which brought water from the mountains to the north, has been reconditioned and still serves the modern city.

In the reign of Valerian (AD 253–260), the walls of Athens, which had been neglected since Sulla's capture of the city in 86 BC and had fallen into ruin, were rebuilt, and the circuit was extended to include the new suburb northeast of the Olympieion. This was done because of the threat of a barbarian invasion, but when that invasion came, in AD 267, the walls were of no avail. The Heruli, a Germanic people from northern Europe, easily captured Athens, and though the historian P. Herennius Dexippus rallied 2,000 men on the city outskirts, they could only resort to guerrilla tactics. The lower town was sacked, and all the buildings of the Agora were burned and destroyed. The Acropolis, however, may have held out; at least there is no evidence of extensive damage at this time.

This sack of Athens is comparable only to that by the Persians in 480 BC, but now the reaction was quite different. The Athenians abandoned the outer circuit and established a new and much smaller line north of the Acropolis, leaving even the Agora area outside the walls. This new wall, which, on the evidence of coins, was built in the reign of Probus (AD 276–282), consisted of material taken from ruined buildings in the lower town.

Athens remained confined within this narrow circuit for several generations, but in the 4th and 5th centuries experienced a revival. The old outer circuit of the walls was restored, and many new buildings were erected. Athens at this time was still the cultural capital of the Greek world and a stronghold of Paganism. Its schools of philosophy, which retained their ancient names, however different their outlooks may have been, flourished, attracting students from all parts. These included the emperor Julian the Apostate and two Fathers of the Church, Basil and Gregory Nazianzene. While the schools existed, Athens remained a place of consequence, but when they were closed by the emperor Justinian in AD 529, Athens sank to the level of a small provincial town. Power and

wealth had long since moved to Constantinople, the new centre of the Greek world.

The Byzantine and Turkish periods. Christianity started early in Athens, with the visit of the apostle Paul in AD 51 and the conversion of Dionysius the Areopagite, a former archon and member of the Court of the Areopagus that had heard Paul's defense of his teachings. The little Christian community did not flourish, however, and Athens remained a stronghold of older ways. In the 5th and 6th centuries, however, after the formal establishment of Christianity and the abolition of pagan worship, churches began to be built. These were sometimes ancient temples converted to Christian worship; for example, the Parthenon, the Erechtheum, and the temple of Hephaestus (the "Theseum"). Newly built churches had a basilica plan and a wooden roof, but these now survive only in foundations. In all, some 22 churches of this period are known.

Effects of Christianity on Athens

The 7th to 10th centuries were dark times for Athens. The city is almost never mentioned in the history of the period, and archaeological remains are few. In the 11th and 12th centuries a measure of prosperity returned, and the taste of Athenians then can be gauged by the number of small stone and brick churches surviving, built on the Byzantine cross-in-square plan, such as the Kapnikaréa, and those of the Saint Theodore and the Holy Apostles.

Athens fell to the Crusaders in 1204, remaining in Latin hands for 250 years. The town's outward appearance changed little, except that the Parthenon, now a Catholic not an Orthodox cathedral, received a campanile (bell tower).

When the Turks captured Athens in 1456, the Parthenon became a mosque, and its campanile was turned into a minaret. Other mosques were built in the lower town, but in general the age of gunpowder was to prove disastrous for Athenian architecture, especially on the Acropolis, which was still virtually intact as late as the mid-17th century. (E.V.)

THE MODERN CITY

Athens, when approached from the Middle East, is the first European city, with tall buildings, newspaper kiosks, modern shops, and modishly dressed women. Approached from Europe, it seems, if not exactly the first Oriental city, at any rate not quite European, in its ill-fitting, locally tailored modernity. The European notes a medley of characterless concrete and out-of-style dress, with the smell of spitted meat and spices in narrow streets as clamorous as bazaars and unpaved roads a few streets from the centre.

The Athenian character. Nevertheless, it is wrong to say that Athens is a mixture of East and West: it is Greek and, more particularly, Athenian. The Athenians, after all, nurtured Western civilization. Yet, some three centuries after the death of Pericles (429 BC), they entered upon a period of bondage that lasted almost 2,000 years. Athens was freed in 1833, and in the following 135 years was the scene of 14 revolutions, another brutal foreign occupation, and a civil war of especial savagery. This long history of passion and suffering has had considerable effect on the Athenian character. The core of that character is an implacable will to survive, buttressed by a profound sense of loyalty (especially to the family) and patriotism. The church, which is directed by a Synod sitting in Athens, was a main force in keeping alive the Greek language, tradition, and literature when such things were forbidden, and most people still support it.

The 2,000-year bondage

The millennia of oppression, instead of driving the Athenian into obtuse moroseness, have honed his wit and rendered him tough but supple, while centuries of privation have only preserved his warmth and generosity. The long oral tradition, alive even under the invader, has reflected and stimulated a taste for rich talk. Of course, the poetic impulse to make a good story better leads to considerable exaggeration in daily conversation, suiting a vanity that goes with a sharp-edged sense of personal and family honour and the spoiling of children. The ancient heroes, too, were vain about both themselves and honour, boasting as much about outwitting the enemy as about outfighting him. Cunning, as in the Odyssey, is still a virtue here.

Athens' development as a modern capital. One-quarter of the nation's population, half its urban population, lives in Athens, and, despite sporadic efforts at decentralization, the city remains the capital in every sense. Since 1833, when Athens became the capital of an independent Greece, the population of the country has increased 12 times, that of Athens 150 times.

In 1833, indeed, there was almost no Athens at all. During the fight for independence, it had been entirely evacuated in 1827, and six years later held perhaps 4,000 people in the straggle of little houses on the north slope below the Acropolis. The newly imported King of the Hellenes, Otho, the 18-year-old son of Ludwig I of Bavaria, was installed in the only two-story stone house, while his German architects hurried ahead with plans for a palace and a new Athens far out in the fields.

Below the well-sited but very plain palace, a large garden square, Síntagma (Constitution) Square, was laid out. Today it is garnished in the tourist season with some of Europe's most luxurious cafe chairs, and at all seasons is hemmed in by tall new buildings and elderly luxury hotels. Broad avenues were created and are still the city centre's principal thoroughfares (Winston Churchill [formerly, Stadium] and Elefthérios [formerly, Venizélos and Panepistimíou] streets), between which an orderly grid of narrow side streets was laid out. The housing that developed was generally the sort of architecture familiar in Victorian London: solid, porched, rather imposing, the later imitations graceless and monotonous. In Athens it is called the Othonian style, but there is little of it left as the centre encroaches on old residential areas.

Once the new capital was established, the city grew at a regular rate of about 7 percent a year, soon reaching 50,000 inhabitants, a figure not much exceeded in the days of Athens' greatest power and glory. By 1907 the municipality had a population of 167,479, Omónia Square had been built at the western end of the two main streets, with other broad avenues radiating from it, but it did not develop as the hoped-for balance to Síntagma.

By now the railway to Piraeus (the port of Athens) had been built, its station near the antique Agora and its tracks laid over the Painted Stoa. Indeed, the city plan projected a logical growth southward along this axis, but a real-estate developer beckoned northward—the National Museum is now out this way—and the newly rich followed. The palace garden almost touched the Arch of Hadrian and the 15 mammoth columns (some of them seven feet ten inches in diameter) of the temple of Olympian Zeus, last of the Classical buildings built in Athens, and beyond lay empty fields. The slopes of Mt. Likavittós, outside the town limits, were still pine-clad.

Since then, the garden has become one of the painfully rare public parks in Athens. Likavittós now rears up in the middle of the city (as if Hyde Park or Central Park were a 1,112-foot [339-metre] mountain), its lower slopes built upon, and many of the trees felled for a road leading to a cog railway and restaurant.

Along Elefthérios (Venizélos) Street rose the Academy of Athens in marble from Mt. Pentelicus, its pediments and colonnades gilded. Its new neighbours were the University of Athens (re-founded in 1837), the colonnade adorned with paintings, and the National Library. All were done in Greek Revival style by the Court's German architects. A new Royal Palace was built during 1891–97, a little southeast of the old (which is now a Parliament house) on Herodes Atticus Street. This leads to the 70,000-seat Panathenaic Stadium, reconstructed by an expatriate Greek millionaire in time for the revival of the Olympic Games in 1896.

Population. In 1921 the orderly progress of Athens was overturned and haphazard development began, for ethnic minorities were exchanged between Greece and Turkey, and approximately 1,500,000 Greeks, most of them penniless, came "home" from Asia Minor. Despite government efforts to resettle them elsewhere, many swarmed into shanty towns around the fringes of Athens and Piraeus, and the area's population soared from 473,-

000 to 718,000. The city is still clearing slums that began as refugee colonies. After this, the city began to spread in two directions, south toward Piraeus and north toward the village of Kifissia, which first became a smart suburb when Herodes Atticus built his villa there in the 1st century BC. The 1940 census showed 481,000 inhabitants in Athens proper, and 1,124,000 in the Athens Basin.

In the decade before the next census, hideous things happened in Athens. During the German occupation, a truck toured the streets picking up the bodies of starvation victims, and the city began to fall apart from lack of maintenance. When the Germans left, part of the Allied-equipped Resistance refused to lay down its arms, and the civil war began. For a while the government held only the Parliament building, neighbouring embassies, and a part of Constitution Square, while the palace garden was used as a common grave.

Housing. Shortly after, a construction boom began, which had still not stopped by the early 1970s. New apartment houses pushing up everywhere erased old social boundaries, though the Kolonáki district on the southeast slope of Likavittós remained an enclave of respectable fortunes, and villages that had been attached to the city in the previous expansion lost their physical and political identities. A network of major highways was thrown up. The west side of the historic olive grove by the Kifisós River was shorn, and hillside greenery began to disappear under housing, either unauthorized or made legal through political skullduggery. Open space vanished, without provision for parks, playgrounds, or even schools, and Athens spread down to the sea by Glifada, joining up with Piraeus. Piraeus itself was transformed from one of the world's celebrated honky-tonk ports into a clean, new-built, flower-decorated city.

Enlargement of the city's master plan

The Athens master plan was enlarged several times to keep pace with spread, which in 1964 already attained 75 square miles (195.6 square kilometres), with a built-up area of 17 square miles outside the plan altogether. Land values in the centre quadrupled, then octupled, and rose proportionately elsewhere. Traffic increased almost to the saturation point at rush hours, and the city continued to sprawl over the plain. As international tourism increased, Athens Airport, with 3,008,000 passengers in 1970, was expanded and modernized, and 25,000 hotel beds were still 5,000 too few in 1971's peak season.

The city water supply from an artificial lake at Marathon was insufficient to supply new building construction and the Mornos River 110 miles to the east was dammed and tapped. Installation of a modern sewer system was undertaken, together with controls to check the floods that roar into Athens when heavy rains pour off the denuded mountains.

Economic life. *Commerce and manufacturing.* Since World War I Athens has become the hub of all mercantile business, export and import. With the Piraeus, it is the most important manufacturing town in Greece. There are cloth and cotton mills, distilleries, breweries, potteries, flour mills, soap factories, tanneries, chemical works, and carpet factories. Exports include tobacco, wine, oil, currants, marble, and, recently, bauxite and magnesite. Publishing enterprises are important. The principal imports are coal, grain, and manufactured articles.

Transportation and shipping. By 1970 Athens accounted for more than half the nation's telephones, more than half its cars, trucks, and buses, and half the jobs in industry and handicrafts. Average earnings were three times higher than the national average. Furthermore, between 1967 and late 1970, the number of merchant ships registered in Greece (mostly at Piraeus) increased from 1,771 to 2,315 as Greek shipowners answered the government's call to bring their foreign-registered ships home (though 3,561 Greek ships remained under other flags). Likewise, shipping receipts rose in the same period from $182,400,000 to $269,760,000. In 1970 alone, 130 shipping offices opened in refurbished Piraeus, while on weekends, shipping magnates sailed to the nearby islands of Hydra and Spetse in chrome-fitted luxury yachts, flying Panamanian and Liberian flags. It hardly seemed to matter where the official city limits stood, since Athens was overrunning everything: the population at the 1971 census was 2,530,207.

Environmental and modernization problems. The brilliant Attic light is now dimmed by the pall of pollution hovering over the city. To discourage new factories from adding to the problem and to stimulate modernization of other regions' economies an industrial wage tax was imposed in the Athens area, and a reduction of up to 40 percent tax on turnover was offered to new factories set up in other areas.

Old Athens

The older Athens has not entirely disappeared in all this hubbub. Older men may have given up smoking hookahs (pipes passing through water) in shadowy cafes but not their 33-bead *komboulói* ("worry beads"), a tension-dissipating addiction acquired from the Turks.

Old Athens also lives on in the six streets sidling off Monastiraki, by the excavated Agora. Here are tiny open-fronted shops hung with tinseled folk costumes, boots apparently captured from German soldiers in 1944, and all the monuments of Athens reproduced in copper, plaster, plastic, and paint. There is an alley of antique dealers, a street of smithies, one of hardware merchants, and another of wildly assorted miscellany.

Close to this lively quarter is the Pláka, on the north slope of the Acropolis. Small, one-story houses, dating from about the time of Independence, are clustered together up the hillside in peasant simplicity. There are appropriately tiny squares, with *tavernae*, once celebrated for their folk music, dancing, and simple fare. There are vine-covered pergolas and some unpaved streets too narrow for cars. The baths built by the Turks still function morning and afternoon, but the *bouzouki*, local relative of the lute, is giving way to the electric guitar, the *taverna* signs are multilingual, and the ubiquitous kitchen chair is being replaced by the plastic-ribbed restaurant seat. Progress laps at the Pláka like a vengeful sea, but the Acropolis is just up above, just under the stars. (B.E.)

BIBLIOGRAPHY

General: IRENE FEKETE, *Athens* (1966), on contemporary Athens, including its districts, public gardens, and marketplaces; STEWART HENRY PEROWNE, *The Pilgrim's Companion in Athens* (1964), historical background and details of the city's principal museums, churches, and other places of interest.

History and antiquities: PAUSANIAS, *Description of Greece*, Book I, the description of Athens by the traveller Pausanias (2nd century after Christ) that contains much of interest—available in translation with brief commentary by PETER LEVI, *Guide to Greece*, 2 vol. (1971)—the classic translation with monumental commentary is by J.G. FRAZER, 6 vol. (1898); THOMAS B. WEBSTER, *Everyday Life in Classical Athens* (1969), the Athenian at home and in public; also by Webster, *Art and Literature in Fourth Century Athens* (1956), the cultural life of the city when it was the intellectual capital of the world; ANGELO PROCOPIOU, *Athens, City of the Gods: From Prehistory to 338 B.C.* (1964), richly illustrated with photographs by EDWIN SMITH; JOHN TRAVLOS, *Pictorial Dictionary of Ancient Athens* (1971), an authoritative and up-to-date account of the monuments of Athens, richly illustrated with photographs and drawings; I.C.T. HILL, *Ancient City of Athens: Its Typography and Monuments* (1953); GERHART RODENWALDT, *Akropolis*, 5th ed. (1956; Eng. trans., 2nd ed., 1957), and H.A. THOMPSON and R.E. WYCHERLEY, *The Agora of Athens* (1972), two detailed and learned expositions of classical Athens' important sites, with many illustrations.

Chiefly photographic: MARTIN HURLIMANN, *Athens* (1956), with introductory text by REX WARNER and detailed historical notes accompanying the pictures; JAN LUKAS, *Athens: A Book of Photographs* (1965), emphasizes the city's contemporary life.

For the specialist: A.W. PICKARD-CAMBRIDGE, *Theatre of Dionysus in Athens* (1946), a thorough study of the theatre and the various uses it was put to; HUMFRY PAYNE and GERARD YOUNG, *Archaic Marble Sculpture from the Acropolis* (1950), a scholarly photographic catalog chiefly for the archaeologist and art historian.

(B.E./E.V.)

Atheriniformes

The order Atheriniformes contains 15 families of marine and freshwater spiny-finned fishes, including the flying fishes, needlefishes, silversides, and cyprinodonts. The last group, the Cyprinodontidae, is an abundant tropical

and subtropical family that includes the guppies, mollies, swordtails, and many other aquarium fishes. In addition to the Atheriniformes, this article will treat the three smaller related orders Beryciformes, Zeiformes, and Lampridiformes, the most primitive groups of the superorder Acanthopterygii or spiny-finned fishes.

GENERAL FEATURES

Beryciforms and zeiforms are mostly deep-bodied fishes of small to moderate size, a foot or less in length. The lampridiforms include a few rare, deep-bodied forms, notably the disk-shaped opah, which may reach more than 300 pounds in weight, but the majority are much elongated, ribbonlike fishes, including the giant oarfish, *Regalecus*, which reaches eight metres (25 feet) in length and is the probable source of many sea-serpent legends. The atheriniform silversides, flying fishes, needlefishes, and halfbeaks tend to be slender, elongate fishes, up to two or three feet in length. The cyprinodonts and their relatives are diminutive and include some of the smallest vertebrates. Many cyprinodonts are important as experimental animals in biological research and as useful predators in the control of insect-borne diseases.

From (*Hoplostethus, Stephanoberyx*) *The Fishes of North and Middle America* by David Starr Jordan and Barton Warren Evermann, Bulletin of the U.S. National Museum No. 47, 1900, reprinted by permission of the Smithsonian Institution; (*Gibberichthys*) A.E. Parr, *Bingham Oceanographic Collections*, vol. 14, no. 6; (*Photoblepharon, Monocentris*) P.P. Grasse, *Traite de Zoologie*, vol. 13 (1958), Masson et Cie, Editeurs; (*Capros, Holocentrus, Zenopsis* N.B. Marshall, *The Life of Fishes* (1965), Weidenfeld & Nicolson, Ltd.

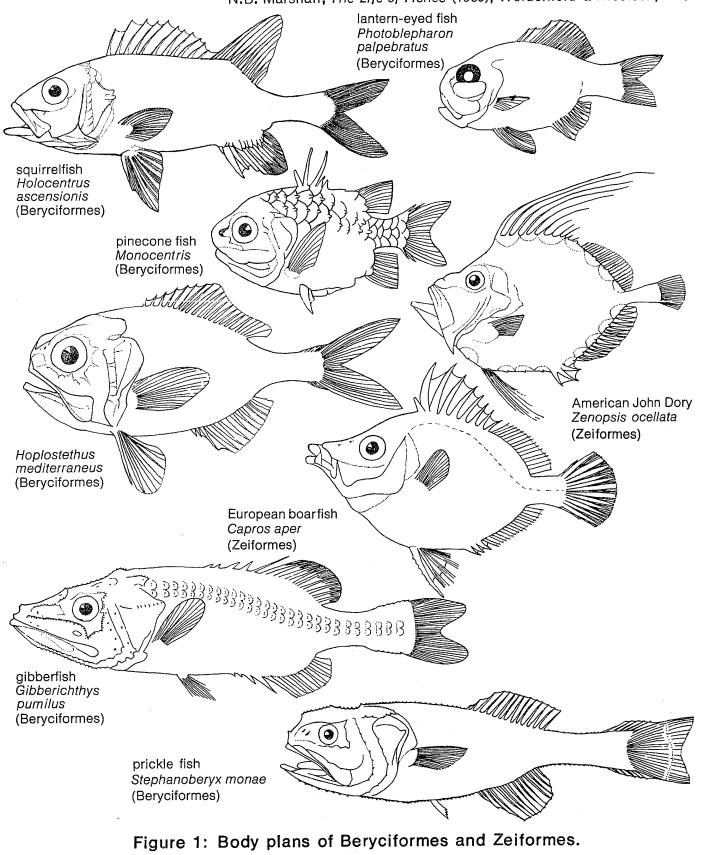


Figure 1: Body plans of Beryciformes and Zeiformes.

NATURAL HISTORY

Most beryciforms, zeiforms, and lampridiforms are inhabitants of the open oceans, usually living at considerable depth, and little is known of their natural history. All appear to produce numerous small eggs. The best known of the beryciform groups are the squirrelfishes and soldierfishes (family Holocentridae), abundant around coral reefs in warm seas. Typical of beryciforms, they are red in colour, with large eyes. Holocentrids are nocturnal, sheltering in crevices during the day and emerging at night to feed. They are notable sound producers, having special drumming muscles attached to the swim bladder, and many have connections between the swim bladder and the ear to improve hearing: presumably these sounds and their reception play some part in courtship. In holocentrids the young (larva) is quite unlike the adult, with a projecting spiny snout and enlarged spines in front of the gill cover. There is a pronounced metamorphosis (a major change in body plan on reaching maturity). It is probable that some of the deep-sea beryciforms undergo similar metamorphoses; the larva of the fanged *Caulolepis* was for many years placed in a different family from the adult, and the genus *Kasidoron*, recently discovered and placed in a distinct family, may be only the larva of *Gibberichthys*. Another family of beryciforms found near the surface is the Anomalopidae, or "strange-eyes," so-called because of a large luminous organ lying directly below the eye, which is switched on by muscular eversion, turning the inside outward (in *Anomalops*) or by the withdrawal of a pigmented cover (in *Photoblepharon*). The Monocentridae, the bizarre pinecone fishes, are another beryciform group with luminous organs, in this case located on the chin. The majority of beryciforms are generalized predators, but a few coral-reef forms are grazers.

Luminescent organs

Among zeiforms, at least one species, *Zeus faber*, produces sounds by drumming muscles and breeds inshore. Little is known of the biology of the oceanic forms, but some certainly undergo metamorphosis, especially the Oreosomatidae, whose larvae are studded with large, spinous tubercles. All zeiforms are highly compressed fishes, with stiff bodies and long dorsal and anal fins: probably they swim by undulating these fins rather than by flexing the body. This slow, stealthy mode of swimming, coupled with their highly protrusile mouths, adapts them for stalking and engulfing prey.

The lampridiforms are all oceanic fishes. Metamorphosis is recorded in dealfishes and oarfishes, the young of which have rather deep bodies and greatly elongated finrays. All are slow swimmers, and the larger forms, the opah and the oarfishes, which are characteristic of surface waters, use their protrusile, toothless mouths as traps for small, planktonic (free-floating) organisms. The deep-sea forms have feebly toothed jaws and are predators. A remarkable modification in one lampridiform, *Lophotus*, is the presence of an ink sac, discharging a viscous, black secretion into the hindgut, thence into the water. These fishes probably use their ink as a defense mechanism, as do squids. *Stylephorus*, a highly modified deep-sea lampridiform, has projecting, telescopic eyes.

Among atheriniforms there is an extraordinary variety of locomotor, reproductive, and ecological adaptations. Locomotor modifications are most marked in the flying fishes, but the origin of the "flying" habit can be traced in flying fish relatives such as the halfbeaks, garfishes, and skippers. All are surface fishes of the open ocean and are capable of leaping or skipping on the surface, sometimes for considerable distances, thus allowing them to escape predators. The tail (caudal) fin is usually asymmetrical, with the lower lobe longer than the upper, and while the body is out of the water the lower lobe vibrates as a scull driving the fish along. True flying fishes have a similar asymmetrical tail, but the pectoral fins are inserted high on the shoulders and are greatly enlarged, with long, stiff fin rays supporting a web of skin. In the most highly evolved flying fishes, the pelvic fins are also enlarged and winglike. The fish accelerates under water by rapid vibration of the tail and fin, with the paired fins furled. On breaking surface, the pectoral fins are expanded, but the lower lobe of the tail remains in the water, sculling rapidly and accelerating the fish. The pelvic fins are then expanded, lifting the tail out of the water and initiating gliding flight. As airspeed is lost, the fish may fall back into the sea or furl its pelvic fins, dropping the lower lobe of the tail into the water and picking up speed for a further glide. Up to five repeated takeoffs have been observed, producing a total flight time of almost half a minute and covering several hundred yards.

Flying fishes

Marine atheriniforms are mostly predators, the predatory habit being most highly developed in the garfishes and needlefishes, with their long, formidably toothed jaws. Freshwater atheriniforms are generally adapted

for feeding at the surface, on insect larvae and small crustaceans.

All atheriniforms are characterized by the production of few, large, adhesive eggs, by mating in pairs, usually accompanied by sexual dimorphism (*i.e.*, the sexes markedly different), and many groups exhibit various reproductive specializations, the most advanced of which is viviparity (the production of functional young, instead of eggs). The young are normally miniatures of the adult and there is no metamorphosis. Sauries, needlefishes, flying fishes and marine halfbeaks are pelagic (*i.e.*, inhabiting open ocean) and breed either in the open sea (sauries, flying fishes) or near the shore (needlefishes, halfbeaks), the eggs often attaching to floating objects by adhesive filaments. The freshwater halfbeaks are mostly viviparous and have an elaborate courtship behaviour.

Atheriniforms of the suborder Atherinoidei fall into two groups, the silversides (Atherinidae and their close relatives) and the more specialized phallostethoids. The silversides are mainly freshwater fishes and show some reproductive specializations in courtship behaviour and sexual dimorphism (coloration and fin shape). They breed near the shore, attaching the eggs to plants. The grunion (*Leuresthes tenuis*) breeds on the California coast, schooling in the surf at extreme spring high water and spawning on the shore, where the female buries the eggs in the sand. The eggs hatch when they are exposed by the next spring tide, two weeks later. In phallostethoids, males have a fleshy, asymmetrical intromittent organ, the priapium, under the throat, formed from the modified pelvic fins. Although fertilization is internal, viviparity is not known to occur.

Breeding specializations of cyprinodonts

Even the most primitive atheriniforms in the suborder Cyprinodontoidei show the usual reproductive specializations of the group: sexual dimorphism and complex behaviour patterns in courtship and spawning. In the Mexican topminnows (Goodeidae) viviparity has developed, the embryos absorbing nourishment within the oviduct of the mother by means of threadlike outgrowths. In the live-bearers (Poeciliidae), an abundant group in the American tropics and subtropics, sexual dimorphism affects many parts of the body. Males have a complex intromittent organ, the gonopodium, formed of modified anal fin rays. One member of the group is oviparous, shedding the eggs while the embryo is only partially developed, but in the guppies, mollies, and swordtails, where the male is much smaller than the female and more brightly coloured, the young are born fully developed, and a series of broods, at about monthly intervals, may result from a single fertilization. In wild cyprinodont populations the sex ratio is frequently unusual, with many females to each male. In *Jenynsia* and *Anableps* (the four-eyed fish) the gonopodium and female reproductive opening are asymmetrical. Both dextral and sinistral forms occur within a species, dextral males mating with sinistral females and vice versa.

Ecological adaptations in atheriniforms are most marked in freshwater species. Cyprinodonts are among the hardiest of fishes and survive in the most rigorous environments. Some cyprinodonts have become adapted to life in hot springs in Africa and America and seem capable of surviving water temperatures approaching the coagulation point of protoplasm. Others survive in stagnant, almost or completely deoxygenated waters, either by taking in water at the surface film, or by breaking surface and gulping air, although no accessory respiratory structures are developed. Some cyprinodonts have overcome the rigours of a seasonal tropical habitat by becoming annuals, growing rapidly and reaching sexual maturity in small temporary bodies of water during the wet season, and on the approach of the dry season, mating and burying the eggs in the mud. The eggs can survive droughts for up to five years, hatching rapidly with the onset of the succeeding wet season. Perhaps another response to rigorous environments is the occurrence in some cyprinodont populations of functional hermaphrodites, capable of self-fertilization and hence of maintaining a population from one surviving parent.

FORM AND FUNCTION

The fishes discussed here share a number of anatomical features typical of the more advanced teleosts. These include a closed swim bladder; separation of the parietal bones by the supraoccipital; jaws that protrude to some extent, with the maxillary bone (toothless except in a few beryciforms) acting as a lever to move the large premaxilla; the pectoral fins inserted high on the flank and the pectoral girdle without a mesocoracoid arch; and a tail skeleton supported by two or less vertebrae. Otherwise, there is considerable structural variation.

Structure of beryciforms

Beryciforms are the most primitive fishes of the four groups under discussion, exhibiting primitive features: the presence of two supramaxillary bones in the upper

From (*Lampris regius*) N.B. Marshall, *The Life of Fishes* (1965), Weidenfeld & Nicolson; (*Mirapinna esau*) P.P. Grasse, *Traite de Zoologie*, Masson et Cie, Paris; (*Kirtlandia vagrans*) D.S. Jordan, *The Study of Fishes*; (others) *The Fishes of North and Middle America* by David Starr Jordan and Barton Warren Evermann, Bulletin of the U.S. National Museum No. 47 (1900), reprinted by permission of the Smithsonian Institution

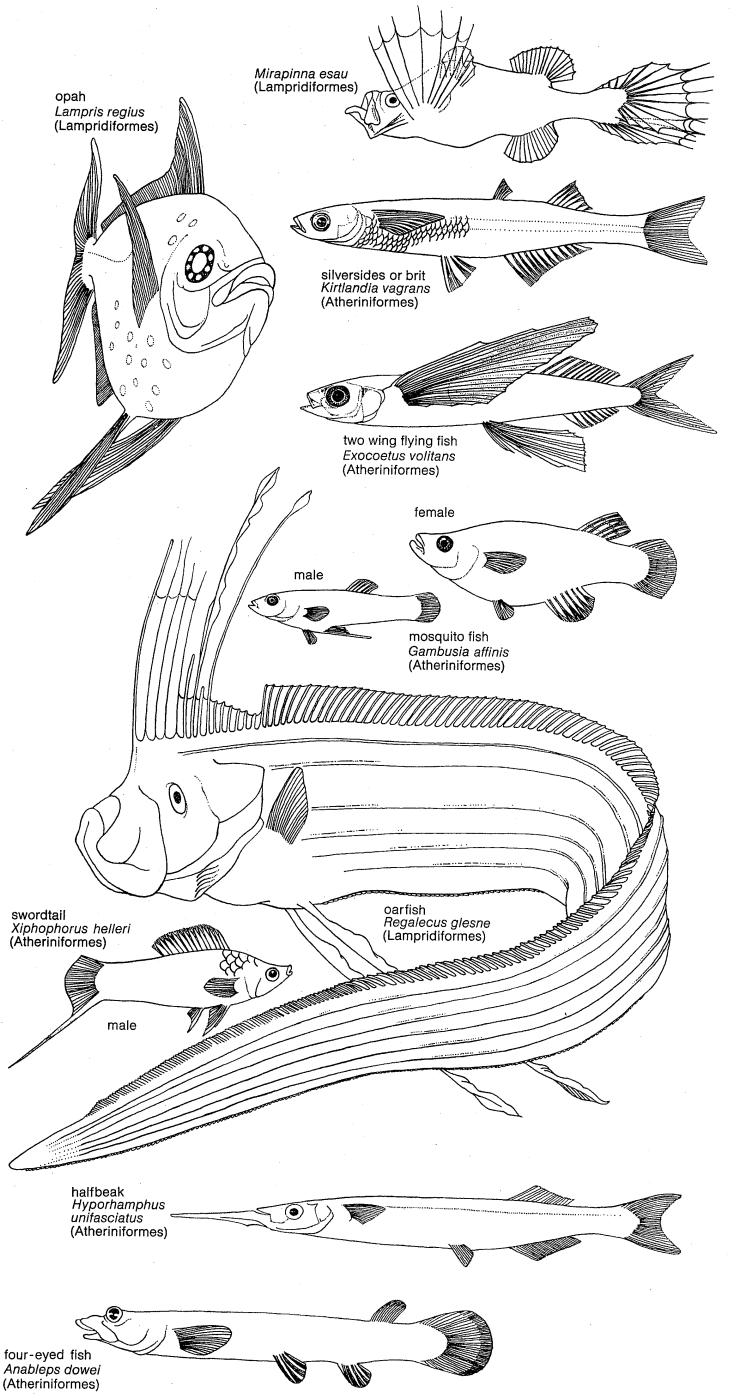


Figure 2: Body plans of Lampridiformes and Atheriniformes.

jaw; an orbitosphenoid bone between the eyes; a tail fin containing 19 principal rays, which insert on six hypural bones supported, in turn, by two vertebrae. Occasionally, they have teeth on the maxillary bone (in the modern holocentrid *Myripristis* and a few Cretaceous fossils). There are many ways in which beryciforms approach the perciforms, the typical "spiny-rayed" fishes. Such resemblances are seen in a number of features: the structure of the mouth, with a normal acanthopterygian pattern of jaw muscles and ligaments; the spiny head bones and ctenoid scales (with a serrated edge); a projection called a subocular shelf on the bones below the eye, stout spines in front of the dorsal, anal, and pelvic fins; bony contact between the pelvic and pectoral girdles; and the short, deep trunk, with about 25 vertebrae. The more generalized beryciforms (holocentrids, trachichthyids, and berycids) exhibit all of these features, but in several lineages degeneration has occurred, associated with life in the deep. In such deep-sea beryciforms as the big-scale fishes (Melamphaeidae), fin spines tend to be absent, the pelvic fins have moved back to the abdomen, and the head bones and scales have become thin and flimsy. Also, in some species primitive structures such as the orbitosphenoid and supramaxillary bones are lacking, and fusions within the tail skeleton have resulted in a condition resembling that of perciforms. The swim bladder is reduced or lost in some.

Structure of zeiforms

Anatomically, the zeiforms resemble perciforms more closely. Almost the only feature that distinguishes zeiforms from perciforms is the presence, in the former, of two or three more rays in the pelvic fins, and in some zeiforms even this distinction fails to hold. Nevertheless, these extra pelvic rays and a few other features, notably the structure of the otoliths ("ear stones," used in maintaining balance), indicate that the zeiforms are beryciform relatives that have independently attained the perciform evolutionary level. Typically, the zeiform has a highly protrusible mouth, a separate spinous dorsal fin, ctenoid scales, and a short, deep trunk; the most primitive members of the order have 24 or less vertebrae.

Structure of lampridiforms

The most primitive lampridiforms are also deep-bodied fishes, with spines in front of the dorsal and anal fins, the pelvic fins directly below the pectorals, an orbitosphenoid bone in the skull, and a tail fin with 19 principal rays, in which they resemble beryciforms. Lampridiforms differ from beryciforms, however, in never having a subocular shelf or pelvic spine, in having more numerous vertebrae, and in having the upper tail fin supports fused with an independent vertebral centrum, a condition resembling that found in the cods and their relatives (Paracanthopterygii). Most lampridiforms have highly protrusile jaws in which depression of the lower jaw dislocates the maxilla of the upper, so that it moves forward bodily, carrying the premaxilla with it. This is a different method from that adopted by other acanthopterygians, hence the name allotriognaths ("strange-jaws") originally applied to the group. A parallel can be drawn between the beryciforms and lampridiforms in certain modifications exhibited by the deep-sea forms, compared with their surface-living relatives. These include the loss of fin spines, reduction in ossification, and reduction of the swim bladder. The most striking features of the more highly evolved lampridiforms, however, are peculiar to the group: great elongation of the trunk, accomplished by increase in vertebral number and elongation of the vertebrae themselves, and reduction of the tail to a small, asymmetrical or filamentous appendage.

Structure of atheriniforms

The atheriniforms are an extremely varied group. There are many structural resemblances to more advanced acanthopterygians, but these are in mosaic distribution, indicating that most have been independently acquired. The jaws of many atheriniforms are protrusile, but the structural modifications by which this is achieved are quite different from those of typical acanthopterygians. The simple, shelflike head of the maxillary bone is attached to the palate only by ligaments, not by a mobile joint. The premaxilla is longer than the maxilla and also has a simple head. Protrusion of the jaws is accomplished by twisting the maxilla and displacing its head forward; the complex system of joints and ligaments characteristic of other acanthopterygians is not developed. The palate is usually toothless, and the series of infraorbital bones incomplete, only the first (lachrymal) and last (dermosphenotic) bones being present. The skull bones are not spiny, but the scales are often ctenoid. The pelvic girdle may have a ligamentous connection with the shoulder girdle but often lies further back, and the girdles never acquire the direct contact that characterizes higher acanthopterygians. The pelvic fin has six or fewer rays, but there is no pelvic spine. The atheriniform tail skeleton is of an advanced type, usually with two large plates emanating from a single supporting centrum, as in some advanced perciforms. The caudal fin contains 17 or less principal rays. There are a few spines in front of the dorsal and anal fins in many atheriniforms, and the members of the Atherinidae and Phallostethidae have a small, separate spinous dorsal fin, but atheriniform spines appear to have evolved independently from those of true acanthopterygians.

An extreme example of adaptation to life near the air–water interface, the habitat of most atheriniforms, is the eye of *Anableps*, the four-eyed fish, so named because each eye is a double structure. The eye is set high on the head and the upper part projects above the water. The cornea is divided by a horizontal band of pigment, separating an upper, strongly convex part from a lower, flatter division. The iris has a pair of projections partially dividing the pupil into two, and the upper is effective for aerial vision, the lower for underwater vision.

Much work has been done on the genetics of atheriniforms, perhaps the most surprising result being the hatching of hybrids between *Fundulus* (Cyprinodontidae) and *Menidia* (Atherinidae), fishes placed in separate suborders. A physiological peculiarity of some marine atheriniforms, garfishes and needlefishes, is a bright green coloration of the bones and even the flesh, due to retention of a bile pigment, biliverdin.

EVOLUTION; PALEONTOLOGY AND CLASSIFICATION

Paleontology. The four orders Beryciformes, Zeiformes, Lampridiformes, and Atheriniformes are primitive groups within the superorder Acanthopterygii. The Beryciformes and Zeiformes apparently form a related group, originating in the Cretaceous, its closest relatives being the Perciformes. The Lampridiformes also originated in the Cretaceous and are of uncertain relationships, being to some extent intermediate between the Acanthopterygii and Paracanthopterygii. The Atheriniformes represent a radiation from near the base of the acanthopterygian stock, but their exact relationships within this group are not known. The present distribution of atheriniforms indicates that the group arose in fresh or brackish waters of the tropical Indo-Pacific region, but little is known of their early fossil history.

Annotated classification.

ORDER BERYCIFORMES

Spiny-rayed fishes with a pelvic spine, an orbitosphenoid, and 19 principal rays in the tail. Of the two main lineages the first contains the Holocentridae, coastal fishes of warm seas. The second is a series of oceanic families centring around the Trachichthyidae. Both groups have fossil records back to the Cretaceous, the 2 lines converging in the Middle Cretaceous.

Family Holocentridae (soldierfishes and squirrelfishes)

Circumtropical, with partly separate spinous dorsal fin. Several extinct genera, Middle Cretaceous onward.

Family Monocentridae (pinecone fishes)

Armoured, very spiny. Teeth on endopterygoid bone. Two genera; Indo-Pacific.

Family Trachichthyidae

Midwater (mesopelagic) or deepwater pelagic fishes, worldwide. Skull bones cavernous, with large mucus cavities. Several extinct genera, Middle Cretaceous onward.

Family Berycidae (*alfonsinos*)

Upper and midwaters in open ocean; worldwide. Pelvic girdle enlarged and tightly joined with the pectoral.

Family Anoplogasteridae

Deep-sea, adults with large fangs; 1 genus.

Family Diretmidae

Very deep bodied, compressed fishes; 1 genus.

Family Anomalopidae (lantern-eyed fishes)

With subocular luminous organ, found near the surface at night; 2 Indo-Pacific genera, 1 Atlantic.

Family Stephanoberycidae (prickle fishes)

Scales and head spiny, fin spines reduced; bathypelagic, worldwide; 3 genera.

Family Melamphaeidae (big-scale fishes)

Abundant deepwater open ocean fishes, worldwide; soft-bodied and black. Fossils in the Miocene.

Family Gibberichthyidae

As Melamphaeidae but with stronger fin spines. Atlantic, 1 or 2 genera.

Family Rondeletiidae (whale fishes)

Head large, no scales, fin spines, or swim bladder; bathypelagic, 1 genus.

Family Cetomimidae (whale fishes)

Mouth enormous, fins without spines, bathypelagic, worldwide.

Family Barbourisiidae (whale fishes)

No fin spines, scales reduced to minute spines, red, bathypelagic, 1 genus.

ORDER ZEIFORMES

As Perciformes but with up to 9 pelvic rays and only 12–13 principal caudal rays.

Family Caproidae (boar fishes)

Most primitive family, 21–23 vertebrae, fossils in the Oligocene; 2 genera, worldwide.

Family Zeidae (John Dories)

Deep bodied and laterally flattened. Mouth large; scales reduced; more than 30 vertebrae. Several genera, worldwide; fossils in the Eocene.

Family Grammicolepididae

Mouth very small, scales drawn out into oblique bands. Two genera, mesopelagic.

Family Oreosomatidae

Larva covered with large tubercles. Four genera, benthic (bottom dwelling); worldwide.

Families Zeniontidae and Macrurocyttidae

Two small families, the first with two genera, the second with one, too poorly known to be characterized.

ORDER LAMPRIDIFORMES

Similar to Beryciformes, but with no pelvic spine; upper hypural bones fused with their supporting centrum.

Suborder Lampridoidei

Deep-bodied forms.

Family Veliferidae

One living genus (*Velifer*) with saillike fins, 33 vertebrae. Fossils from Paleocene and Eocene, several extinct genera.

†*Families Aipichthyidae and Pharmacichthyidae*

Extinct families, each containing a single Upper Cretaceous genus; appear to be primitive lampridiforms, resembling *Velifer* in deep trunk, but with fewer vertebrae and more primitive tail skeletons.

Family Lamprididae (opahs)

One genus (*Lampris*); 15–17 pelvic rays, 46 vertebrae. Fossils from Miocene. Length to 2 m, weight to 300 kg; surface waters (epipelagic) of warm seas; widespread.

Suborder Trachipteroidei

Ribbonlike, about 100 vertebrae.

Family Trachipteridae (dealfishes)

Pelvic fins with 5–9 rays, no anal fin, jaws toothed. Length to 1.2 m; epipelagic. Worldwide in warm seas.

Family Lophotidae (unicorn fishes)

Scales lacking; pelvic fins small or absent, anal fin short. Fossils from Oligocene. Worldwide in warm seas.

Family Regalecidae (oarfishes)

Anal fin lacking; 1 pelvic ray elongated; jaws toothless; length to 9 m; weight to 300 kg. Mesopelagic, tropical.

Suborder Stylephoroidei

Family Stylephoridae

Deep-sea forms with enlarged telescopic eyes, about 50 vertebrae, 2 filamentous caudal rays. Known from only a few specimens.

Suborder Ateleopoidei

Family Ateleopidae

Specialized, deep-sea, bottom-living fishes, Indo-Pacific and Atlantic, usually placed among the primitive teleosts, but probably lampridiform.

Suborder Mirapinnoidei

Families Mirapinnidae and Eutaeniophoridae

Three species of little-known mesopelagic fishes, usually placed as a distinct order of lower teleosts (Mirapinniformes), but probably larval lampridiforms.

Suborder Megalomycteroidei

Family Megalomycteridae

Four rare, little-known, deep-sea genera, probably larval lampridiforms.

ORDER ATHERINIFORMES

Premaxilla greatly expanded between maxilla and mandible, without crossed ligaments controlling the upper jaw, infraorbital bone series incomplete.

Suborder Exocoetoidei

Lateral line complete and low on the flank in marine forms, lower pharyngeal bones fused, no parietals, 9–15 branchiostegals. Worldwide, but especially abundant in the Indo-Pacific.

Family Exocoetidae (halfbeaks and flying fishes)

Lower jaw often extended; snout not modified. Surface marine waters and freshwaters, worldwide; length to 45 cm. Fossil half-beaks in the middle Eocene.

Family Belonidae (garfishes and needlefishes)

Snout bones sutured together, both jaws elongated into a strongly toothed beak. Mostly temperate and tropical marine; a few freshwater; length to 120 cm. Fossils in the Oligocene.

Family Scomberesocidae (sauries, skippers)

Snout and jaws as in Belonidae but feebly toothed; small finlets behind dorsal and anal fins. Inshore temperate and tropical marine waters; length to 35 cm. Fossils in the Miocene.

Suborder Cyprinodontoidei

Lateral line represented by pits on the flank, 4–7 branchiostegal bones. Families mostly distinguished by reproductive specializations.

Family Oryziatidae (medakas)

Most primitive cyprinodonts; a single genus in freshwaters and brackish waters in Indonesia.

Family Adrianichthyidae

Mouth and snout enlarged and shovellike. Two genera in lakes in Celebes; length 7–20 cm. Fossils in Late Tertiary in Celebes.

Family Horaichthyidae

Small fishes with anal fin modified, in males, for clasping female in mating. One genus, freshwater, India.

Family Cyprinodontidae (killfishes or egg-laying topminnows)

Circumtropical and temperate marine and freshwater, many genera. Many popular aquarium fishes; length to 15 cm. Fossils in the Oligocene.

Family Goodeidae (Mexican topminnows)

Live-bearing, but male lacks elaborate intromittent organ found in poeciliids. About 10 genera, in rivers draining the Mexican Plateau; length to about 10 cm.

Family Jenynsiidae

Small fishes with asymmetrical genital organs; 1 genus; rivers of South America.

Family Anablepidae (four-eyed fishes)

Characterized by specialized eye structure (see above *Form and function*); 1 genus, 2 species; surface waters in rivers and estuaries of South America.

Family Poeciliidae (live bearers or viviparous topminnows)

Native to tropical and subtropical America but introduced elsewhere for mosquito control. Freshwaters and coastal marine waters. Length 1.5 to about 15 cm. Family includes mollies (*Mollienesia*), guppies (*Lebistes*), swordtails (*Xiphophorus*), any many other popular aquarium fishes, as well as the mosquito fishes (*Gambusia*).

Suborder Atherinoidei

Lateral line variable; 5–7 branchiostegal bones; separate spinous dorsal fin.

Family Melanotaeniidae

Many species; freshwater bodies of New Guinea and Australia. Compressed, deep-bodied; pointed snout; 5–20 cm.

Family Atherinidae (silversides)

Lateral line absent; pelvic fins midway along belly; length 7–70 cm. Coastal and freshwater, worldwide in warmer regions. Many genera. Fossils from middle Eocene.

Family Isonidae

Pectoral fins unusually high on body. Small marine fishes; Indian and Pacific Oceans. Two genera.

Families Phallostethidae and Neostethidae

Males with priapium, an organ derived from pectoral and pelvic girdles, functioning to clasp female. Tiny fishes (3–5 cm long); confined to freshwaters and brackish waters in Thailand, Indonesia, and Philippines.

Critical appraisal. The whale fishes (cetomimids, rondeletiids, barbourisiids) are often placed in a separate order Cetomimiformes, thought to be more primitive than Beryciformes, but their "primitive" features appear to be due only to degeneration. The stephanoberycids, melamphaeids, and gibberichthyids are usually placed in a suborder Stephanoberycoidei, all other beryciforms being placed in the Berycoidei, but the major phyletic cleft in Beryciformes seems to be between the holocentrids and the remainder, which form a related group.

BIBLIOGRAPHY. R.M. ALEXANDER, "Mechanisms of the Jaws of Some Atheriniform Fish," *J. Zool.*, 151:233–255 (1967), an account of methods of jaw protrusion; C.M. BREDER and D.E. ROSEN, *Modes of Reproduction in Fishes* (1966), especially good on reproductive modifications in atheriniforms, with full bibliography; D.S. JORDAN and C.L. HUBBS, "A Monographic Review of the Family Atherinidae or Silversides," *Stanford Univ. Publs., Univ. Ser., Studies in Ichthyology*, 1:1–87 (1919), a classic review of atherinoids; C.T. REGAN, "On the Anatomy, Classification, and Systematic Position of the Teleostean Fishes of the Suborder Allotriognathi," *Proc. Zool. Soc. Lond.*, pp. 634–643 (1907), a classic paper in which the lampridiforms were first grouped together; R.R. ROFEN, "The Whale-Fishes: Families Cetomimidae, Barbourisiidae and Rondeletiidae (order Cetunculi)," *Galathea Rep.*, 1:255–260 (1959), an illustrated account of these deep-sea forms; D.E. ROSEN, "The Relationships and Taxonomic Position of the Halfbeaks, Killifishes, Silversides and their Relatives," *Bull. Am. Mus. Nat. Hist.*, 127:219–267 (1964), a monograph, with bibliography, in which the Atheriniformes were first grouped together, and with R.M. BAILEY, "The Poeciliid Fishes (Cyprinodontiformes), their Structure, Zoogeography, and Systematics," *ibid.*, 126:1–176 (1963), a monographic account of peociliids, and with C. PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *ibid.*, 141:359–474 (1969), discussions of the relationships of fossil and living beryciforms, lampridiforms and atheriniforms, with bibliography.

(C.P.)

Athletic Games and Contests

This article covers the historical background of athletic games and contests from their beginnings in ancient Greece through the renewal of the quadrennial Olympic Games at Athens in 1896 and their development into the pre-eminent athletic events of the world. Tables at the end of the article list the Olympic champions from 1896.

Just how far back in history organized athletic contests were first held remains a matter of doubt, but it is reasonably certain that they occurred in Greece, at least, some 1,500 years before Christ.

Classical games

However ancient in origin, by the end of the 6th century BC at least four of the Greek sporting festivals, sometimes known as classical games, had achieved major importance. They were the Olympic Games, held at Olympia; the Pythian Games at Delphi; the Nemean Games at Nemea; and the Isthmian Games at Corinth. Later, similar festivals were held in nearly 150 cities as far afield as Rome, Naples, Odessus, Antioch, and Alexandria.

The Olympic Games in particular were to become famous throughout the Greek world. There are records of the champions at Olympia from 776 BC to AD 217. The Games, held every four years, apparently continued till AD 394 when they were abolished by the Roman emperor Theodosius I. For the first 100 or 200 years, Olympic champions came from a dozen or more Greek cities—the majority from Sparta and Athens; but in the next three centuries, athletes were drawn from 100 cities in the Greek empire. And in the final 100 years or so before the games were discontinued champions came from as far afield as Antioch, Alexandria, and Sidon.

Fifteen centuries after the abolition of the Olympic Games, thought was given to organizing events of a comparable character. In 1887 the 24-year-old Baron Pierre de Coubertin conceived the idea of reviving the Olympic Games and spent seven years preparing public opinion in France, England, and the United States to support his plan. At an international congress in 1894, his plan was accepted and the International Olympic Committee was founded. The first modern Olympic Games were held in Athens in April 1896, with 13 nations sending nearly 300 representatives to take part in 42 events and 10 different sports. Seventy-six years later, more than 100 nations sent 6,000 athletes to Munich to take part in almost 200 events in more than 20 different sports. In 1976 at Montreal, the number of participants was about the same as at Munich despite the last minute withdrawal of 22 countries involving nearly 450 athletes.

The revival of the Olympic Games led to the formation of many international bodies controlling their own amateur sports and to the creation of National Olympic Committees throughout the world. By the 1960s and '70s, through the media—the press, radio, and especially television—the quadrennial Games were brought into millions of homes.

History

CLASSICAL GAMES

Greece. *The Olympic Games.* Of all the games held throughout Greece, those staged at Olympia in honour of Zeus are the most famous. Held every four years between August 6 and September 19, they occupied such an important place in Greek life that time was measured by the interval between them—an "Olympiad." Although the first Olympic champion listed in the records was one Coroebus of Elis, a cook, who won the sprint race in 776 BC, it is generally accepted that the Games were probably at least 500 years old at that time. According to one legend they were founded by Heracles, son of Alcmene. The Games, like all Greek games, were an intrinsic part of a religious festival. They were held at Olympia in the city-state of Elis, on a track about 30 yards (27 metres) wide and one stade (about 600 feet [183 metres]) long. Later, the word *stadion* was sometimes employed to describe the *dromos* (a race one length of the track) and also the arena itself.

The first Olympic champion

At the meeting in 776 BC, there was apparently only one event, the *dromos*; but other events were added over the ensuing decades. In 724 BC a two-length race, the *diaulos*, roughly similar to the modern 400-metre race, was included, and four years later the *dolichos*, a long-distance race possibly to be compared to the modern 1,500- or even 5,000-metre event, was added. Wrestling and the pentathlon were introduced in 708 BC. The latter was an all-round competition consisting of five events—the long jump, javelin throw, discus throw, foot race, and wrestling.

Boxing was introduced in 688 BC, and in 680 a chariot race. In 648 the pancratium (Greek, *pankration*), a kind of all-strength, or no-holds-barred, wrestling, was included. Kicking and hitting were allowed; only biting and gouging (thrusting a finger or thumb into an opponent's eye) were forbidden. Between 632 and 616 BC events for boys were introduced. From time to time further events were added, including contests for fully armed soldiers, for heralds, and for trumpeters. The program must have been as varied as that of the modern Olympics, although the athletics (track and field) events were limited; there was no high jumping in any form and no individual field event, except in the pentathlon.

Until the 77th Olympiad (472 BC) all of the contests took place on one day; later they were spread over four days, with a fifth devoted to the closing-ceremony presentation of prizes and a banquet for the champions. Women were not allowed as competitors or, except for the priestess of Demeter, as spectators.

The Olympic Games were originally restricted to "free-born" Greeks. The competitors, including those who came from the Greek colonies, were "amateur" in the sense that the only prize was a wreath or garland. The athletes underwent a most rigorous period of supervised training, however, and eventually the contestants were true professionals. Not only were there substantial prizes in kind or in money but the Olympic champion was handsomely rewarded for his prowess and received adu-

Early professionalism

lation and unlimited benefits from his city. Athletes became full-time specialist performers. An ominous parallel in the course of the modern Olympic Games has been suggested.

Pythian Games. From earliest times there had been a religious festival at Delphi in connection with the oracle of Apollo. Originally the games associated with the festival were held every eight years, but in 582 BC a four-year cycle was introduced, in the third year of an Olympiad. In addition to sporting events and contests, there were competitions in singing, the playing of musical instruments, and recitation. The prize was a crown of laurel leaves.

Nemean Games. According to one legend, these games were founded by Adrastus of Argos when he led the Seven against Thebes; another founder is said to have been Heracles, after the slaying of the Nemean lion. The games, conducted on similar lines to the Olympics, were in honour of Zeus. They took place every two years, in the second and fourth years of an Olympiad. The prizes were crowns of wild celery.

Isthmian Games. These athletic and musical competitions were held in Corinth in honour of Poseidon. They took place every two years, in the same years as the Olympic and Nemean Games.

Other games. Similar games were held in cities not only in Greece proper, but in Macedonia, Egypt, Syria, Arabia, Italy, and many other areas. While there was variation in the types of events included, the pattern was much the same, and the semi-religious character of the games was retained. The games for the most part took place on fixed dates at a sanctuary and were integrated with religious ceremonies. Every Greek city of any size held what might be described as local games, while larger ventures were the combined effort of several cities or states. Participation soon became a career, and outstanding young athletes were sponsored by the city in which they lived in the hope that one day they would bring it glory by their victories.

Local games

Rome. Greece lost its independence to Rome in the middle of the 2nd century BC, and the support for the competitions at Olympia and other places fell off considerably in the next century. The Romans looked on athletics with contempt—to strip naked and to contend in public was degrading in the eyes of the Roman citizen. The Romans realized the value of the Greek festivals, however, and Augustus, who had a genuine love for athletics, staged athletic games in a temporary wooden stadium erected near the Circus Maximus. Nero was also a keen patron of the festivals in Greece. By the 4th century AD, Rome, with its population of more than 1,000,000, had well over 150 holidays for games. There was chariot racing in the hippodrome and horse racing in the Circus Maximus, with room for more than 250,000 spectators. In an amphitheatre with accommodation for 50,000, animals and human beings were maimed and slaughtered.

Indeed, public games were held in abundance—games in honour of the gods, some of which were said to go back to the foundation of the city. But through all this, athletic events occupied a secondary position. The only ones that really interested the Romans were the fighting events—wrestling, boxing, and the pancratium. The main difference between the Greek and Roman attitude was that the Roman festivals were described as *ludi* (games), the Greek as *agōnes* (contests). The Greeks originally organized their games for the competitors, the Romans for the public. One was primarily competition, the other entertainment; and it is not unreasonable to suggest that the Greeks took an "amateur" view of sport, the Romans a professional.

GAMES FROM THE 5TH TO THE 19TH CENTURY

From the fall of Rome in the 5th century to the 19th century, the history of sport is but thinly documented. Organized games and contests as religious observances and events sponsored by the state seem to have disappeared, and athletic contests and sports generally to have declined. All that the records show is that in the Middle Ages, religious festivals were often the scene of primitive games between towns and guilds. On holy days, there were national holidays, and assuredly sporting matches and contests took place on these occasions. Games of all kinds were practiced in England and practiced extensively, as the various proclamations and laws forbidding sports and the playing of games testify. In the reign of Elizabeth I, for example, a statute of Oxford University forbade the playing of football by any scholar over the age of 16 on penalty of a fine and imprisonment, and expulsion for a third offense.

Class distinction was much in evidence. The aristocracy engaged in tournaments and jousts and in hunting. Royalty played court, or royal, tennis, as the English diarist Samuel Pepys testifies:

> *Thence to the Tennis Court and there saw the King* play at tennis and others. To see how the King's play was extolled without any cause at all was a loathesome sight, though sometimes he did play right well and deserved to be commended; but such open flattery is beastly.

Pepys mentions in an entry for May 1663:

> . . . great thronging upon Banstead Downs upon a great horse race and foot race. The foot race was between Lee, the Duke of Richmond's footman and a tyler, a famous runner. And Lee hath beat him, though the King and Duke of York and almost all men did bet 3 to 4 to 1 upon this tyler's head.

The common people played football, while prizefights, bearbaiting, and cockfighting drew large crowds. Competitions between pedestrians (called "peds," that is to say, professional runners) were frequent, as were contests against the clock. At the beginning of the 19th century Captain Barclay (Robert Barclay Allardice), a famous English pedestrian, walked 1,000 miles in 1,000 consecutive hours before thousands of spectators on Newmarket Heath. There must have been hundreds of local sports meetings with competitions for men, women, and youths of both sexes. But real championships were a product of the late 19th century, and then mainly in England and the United States.

Pedestrianism

The real growth of organizations concerned with the national control of sport came in Great Britain in the latter half of the 19th century with, for example, the formation of governing bodies for association football in 1863, rugby football in 1871, athletics in 1880, boxing in 1884, and lawn tennis in 1888. The Amateur Athletic Union of the United States, the biggest national governing body in the world, was founded in 1888.

For the most part, these governing bodies were formed to control amateur sport at a time when amateur performers competed primarily for the love of competition. But the improvement in the standard of amateur performers inevitably led to public popularity, the building of stadiums to accommodate paying crowds, and the subsequent great inducements to become professional.

MODERN GAMES AND CONTESTS

It is to amateur sport that credit for the initiation of international competition is due, for while from time to time individual professional performers from one country pitted their skill against individuals from another, such encounters were rare. One can instance the boxing match in 1860 between Tom Sayers from England and John Heenan of the United States, which went to 42 rounds before the police intervened; and the sensational performances of an American Indian, who called himself Deerfoot, on the track in England in the same year. But it was the formation of the various amateur clubs in different sports from 1860 onward that led to the creation of organized international competitions. A cricket team from England visited Canada and the United States in 1859 and Australia two years later. The first Test match (*i.e.*, between representative national teams) against Australia took place in March 1877. Yale University sent a track team to meet Oxford in London in 1894, followed in 1895 by a match in New York City between the oldest American athletic club, the New York AC, and the oldest British club, the London AC. The Americans won every event.

Beginnings of international competitions

The revival of the Olympic Games in 1896 gave the real impetus to international competition and led in time to

the formation of the many national and international governing bodies.

Revival of the Olympics. The architect of the modern Olympics was Pierre de Coubertin, born in Paris on New Year's Day, 1863. As a young man he was intensely interested in literature and in education and sociology. Family tradition pointed to an army career or possibly politics, but at the age of 24 de Coubertin decided that his future lay in education. At the same time he had the idea of reviving the Olympic Games, and in 1892 he propounded his desire for a new era in international sport when on November 25th, at a meeting of the Union des Sports Athlétiques in Paris, he said:

> Let us export our oarsmen, our runners, our fencers into other lands. This is the true Free Trade of the future; and the day it is introduced into Europe the cause of Peace will have received a new and strong ally. It inspires me to touch upon another step I now propose and in it I shall ask that the help you have given me hitherto you will extend again, so that together we may attempt to realise, upon a basis suitable to the conditions of our modern life, the splendid and beneficent task of reviving the Olympic Games.

The speech did not produce any appreciable activity, but de Coubertin was not fainthearted. At a conference on international sport in Paris in June 1894 at which de Coubertin raised the possibility of the revival of the Olympic Games, there were 79 delegates representing 49 organizations from nine countries. De Coubertin himself wrote that except for his co-workers Dimítrios Vikélas of Greece, who was to be the first president of the International Olympic Committee, and Professor William M. Sloane of the U.S. from the College of New Jersey (later Princeton University), nobody had real interest in the revival of the Games. Nevertheless, and to quote de Coubertin again, "a unanimous vote in favour of revival was rendered at the end of the Congress chiefly to please me."

It was at first agreed that the Games should be held in Paris in 1900. Six years seemed a long time to wait, however, and it was decided to change the venue—what better site than Athens, the capital of Greece—and the date, to April 1896.

First games of the modern cycle

A great deal of indifference, if not opposition, had to be overcome, including a refusal by Athens to stage the Games at all. But de Coubertin and his newly elected International Olympic Committee of 14 members won through, and the Games were opened by the king of Greece in the first week of April 1896.

Regional games related to the Olympics. The International Olympic Committee (IOC) grants its patronage to certain regional games that contribute to the development of amateur sports in the areas in which they are organized; that are conducted in a dignified manner and without commercial exploitation; and that follow the policies, rules, and regulations of the IOC and the international sports federations concerned.

Three of the regional games that have received Olympic patronage were first held in 1951: the Pan-American Games, the Mediterranean Games, and the Asian Games. The Central American and Caribbean Games were first held in 1926 and the Bolivarian Games in 1938. The Far East Games, encompassing China, Japan, and the Philippines, were first held in 1913 and last held in 1930.

From time to time, often at somewhat irregular intervals, a number of area games have been staged, of which the following may be mentioned: Pan-African, Pan-Arabian, Pan-Asian, Southeast Asian Peninsular, South Pacific, Pacific Conference, Balkan, International Youth, and World Student games.

Other regional games. *Commonwealth Games.* In 1928 plans were set up for what were to be known as British Empire Games. The first meeting was held in Hamilton, Canada, in 1930, and 11 countries took part with a total of 400 competitors. The sports included lawn bowling, or bowls, in addition to the more typical events in boxing, rowing, swimming and diving, track and field, and wrestling. The 1954 games, in Vancouver, were renamed the British Empire and Commonwealth Games; in 1970 the title was changed to British Commonwealth Games, and later the word British was omitted.

Highland and Hibernian games. The Braemar Gathering, the most famous of all Highland Games, dates from the 8th century AD. About 1100, according to tradition, King Malcolm of Scotland built a hunting lodge in the highlands, at Braemar in Aberdeenshire. One day he required a courier to take some dispatches to Edinburgh. In order to obtain a suitable runner for the royal task, a gathering of local inhabitants organized a race to the top of a hill a few minutes away through scrub and heather, for a prize of a sword, a belt, and a purse of gold. The race is said to have been the first of the traditional highland hill races. By the end of the 14th century the area had been depopulated by the Black Death, and the modern games date from the 19th century, when they were revived to raise money for the relatives of men killed at the Battle of Waterloo (1815). Regular gatherings have been held since.

Events staged at most Highland Games—and there are as many as 40 such gatherings in Scotland annually—are putting the stone, throwing the hammer, tossing the caber, and the pole vault, the high jump and broad jump, wrestling, and hill races. The caber is a pole, usually from 16 to 19 feet (5 to 6 metres) long and weighing 90 to 120 pounds (40 to 54 kilograms). Distance is of no significance; the object is to carry and toss the caber so that the end of the pole in the hands of the thrower passes through an angle of 180 degrees and lands away from the thrower. "Highland" Games have not been confined to Scotland. They have been held in many parts of the world including North Carolina and Nova Scotia.

Caber tossing and wheel throwing

Hibernian Games, held originally in Ireland, are of great antiquity. The Tailtean Games at Tara, in memory of Queen Tailte, were first held in the 19th century BC and continued for many centuries until sometime in the Middle Ages. The Hibernian weight-throwing event was the *roth cleas*, or wheel feat. Legend tells of the Celtic hero Cú Chulainn, who gripped a chariot wheel by its axle, whirled it around his head, and threw it farther than could any other mortal. Wheels were later replaced by boulders attached to wooden handles; modern versions are weight throwing and hammer throwing.

Maccabiah Games. This international Jewish festival held under the auspices of the Maccabi World Union was first staged in Czechoslovakia in 1929. Meetings were held in Belgium (1930) and thereafter in Israel.

International and national contests. The revival of the Olympic Games gave an impetus to sport that gradually spread throughout the world, so that the number of encounters between countries in different sports must be many hundreds annually. It is not only the world championships in many sports, both amateur and professional, but far more the interchange of club and national teams. The tremendous growth in air travel has made such contacts possible. In track and field most countries have a number of international matches every year; and from time to time contests are set up between the United States and Europe, and the United States against the British Commonwealth. In 1977 the first World Cup of the International Amateur Athletic Federation was held in Düsseldorf, West Germany, with competitors representing African, North American, South American, Oceanian, Asian, and European countries. The tendency toward many competitions is making it increasingly difficult for a top-class performer to remain an amateur, even with the gradual loosening of the strict and now absolutely unenforceable rules. The greatest difficulty is the ever-increasing number of competitions coupled with the public demand for the best performers. This tends to make sport more of an entertainment than a recreation, and it is questionable that the demand for the superlative by the public, including the world television audience, can be combined with the traditional idea that sport is for the competitor.

The modern Olympics

ORGANIZATION

The International Olympic Committee. At the Congress of Paris in 1894, the control and development of the modern Olympic Games was entrusted to the Inter-

national Olympic Committee (IOC; Comité International Olympique), with headquarters at Mon-Repos, Lausanne, Switzerland. It is responsible for maintaining the regular celebration of the Olympic Games; seeing that the Games are carried out in the spirit that inspired their revival; and promoting the development of amateur sport throughout the world. The original committee in 1894 consisted of 14 members and de Coubertin.

Convinced that the downfall of the ancient Olympic Games had been caused by outside influences that undermined the spirit of the Games, de Coubertin felt that the revived Games would go the same way unless they were in the hands of people whose concern was to keep the spirit of amateur sport alive, and who were responsible in no way to any outside influences. Thus IOC members are regarded as ambassadors from the IOC to their national sports organizations. They are in no sense delegates *to* the committee and may not accept from the government of their country, or from any organization or individual, any instructions that in any way affect their independence.

Independence of International Committee members

The IOC is a permanent organization that elects its own members. Each member—the total is about 70—must speak French or English and be a citizen or reside in a country that has a National Olympic Committee. With a very few exceptions, there is only one member from any one country. Members were originally elected for life, but anyone elected after 1965 must retire at 72.

The IOC elects its president for a period of eight years, at the end of which he is eligible for re-election for further periods of up to four years each. Michael Morris, Lord Killanin (Ireland) was elected in 1972. Previous presidents were Dimítrios Vikélas (1894–96, Greece), Baron Pierre de Coubertin (1896–1925, France), Count Henri de Baillet-Latour (1925–42, Belgium), J. Sigfrid Edström (1946–52, Sweden), Avery Brundage (1952–72, U.S.).

The executive board of nine members holds periodic meetings with the international federations and National Olympic Committees. The IOC as a whole meets when summoned by the president, and he must convene a meeting if one-third of the members so request.

National Olympic Committees. Each country that desires to participate in the Olympic Games must have an Olympic committee accepted by the IOC. In the 1970s there were more than 120 such committees.

A National Olympic Committee is composed of at least five national sporting federations, each affiliated to an appropriate international federation. The ostensible purpose of these National Olympic Committees is the development and promotion of the Olympic movement and of amateur sport. National committees arrange to equip, transport, and house their country's representatives at the Olympic Games. According to the rules of the committee, they must be not-for-profit organizations; must not associate themselves with affairs of a political or commercial nature; and must be completely independent and autonomous and in a position to resist all political, religious, or commercial pressure.

A person who has ever competed in sports as a professional, who has ever coached sports competitors for payment, or who is engaged in or connected with sport for personal profit is not eligible to serve on a national committee. The rules provide that

> exceptions to these categories may be made by the Executive Board of the I.O.C. on the recommendation of the National Olympic Committee concerned.

National Olympic Committees that do not conform to IOC rules and regulations forfeit their recognition and their right to send participants to the Olympic Games.

IOC awards. In individual Olympic Games events, the award for first place is a gold (silver-gilt, with six grams of fine gold) medal, for second place a silver medal, and for third place a bronze medal. Solid gold medals were last given in 1912. Diplomas are awarded for fourth, fifth, and sixth places. All competitors and officials receive a commemorative medal.

Since 1974 there have been two noncompetitive awards, the Olympic Cup and the Olympic Order. The former was instituted by Baron de Coubertin in 1906. It is awarded to an association or institution that has a general reputation for merit and integrity, that has been active in the service of amateur sport, and that has made a substantial contribution to the development of the Olympic movement. The first recipient of the Olympic Cup was the Touring Club de France. It was awarded to the town of Innsbruck, Austria, for its organization of the XII Winter Games of 1976.

The Olympic Order, created in 1974, is intended for living persons, excluding active members of the IOC. There are three degrees of award—gold, silver, and bronze—to be made "to any person who has illustrated the Olympic ideal through his action, has achieved remarkable merit in the sporting world, or has rendered outstanding services to the Olympic cause, either through his own personal achievement or his contribution to the development of sport." The gold medal was awarded posthumously to Avery Brundage, for 20 years president of the IOC. Ten silver and 13 bronze awards have been made, including silver medals to Jesse Owens, winner of four gold medals at the 1936 Olympics, and to Dan Ferris, for more than 50 years a prominent administrator in United States and world track and field.

Olympic Order

Five previous awards were suspended in 1974: the Olympic Diploma of Merit, first awarded in 1905; the Sir Thomas Fearnley Cup and the Mohammed Taher Trophy, first awarded in 1951; the Count Alberto Bonacossa Trophy, established in 1955; and the Tokyo Trophy, first awarded in 1967.

The Games and participants. The Olympic Games (Summer) are held every four years. There is no age limit for competitors, and in theory no discrimination is allowed against any country or person on grounds of race, religion, or political affiliation. The Games are contests between individuals and not between countries. Separate Winter Games have been held since 1924.

The Olympic Games celebrate an Olympiad, or period of four successive years. The first Olympiad of modern times was celebrated in 1896, and subsequent Olympiads are numbered consecutively, even though no Games take place (as was the case in 1916, 1940, and 1944).

The period of the Summer Games must not exceed 15 days; the Winter Games are limited to 10 days.

The maximum number of entries permitted for individual events is three per nation. The number is fixed (but can be varied) by the IOC in consultation with the international federation concerned. In team events, only one team per country is allowed. In general, a National Olympic Committee may only enter a citizen of the country concerned.

Problems of amateurism. To be eligible to compete, a competitor must be an amateur as defined by the international body of the particular sport and also by the rules of the IOC.

The IOC rules were substantially changed in 1976 and now provide that a competitor must not have received any financial reward or material benefit in connection with his or her sports participation except as permitted by the IOC bylaws. These bylaws permit the athlete to receive personal sports equipment and clothing, travelling and hotel expenses, and compensation (authorized by a National Olympic Committee or a National Federation). Compensation may be given in case of necessity to cover financial loss resulting from absence from work or basic occupation because of preparation for, or participation in, the Olympic Games and international sports competitions.

IOC rule changes

This wide concession for payment, known as "broken time," in effect authorizes a competitor (subject, of course, to the rules of each individual International Federation, or of any national controlling body) to devote as much time as he or she wishes to training and competitions and to be recompensed to the full for any loss of earnings. The only limitation is that "in no circumstances shall payment made exceed the sum which the competitor would have earned in the same period." In the extreme, therefore, a competitor may abandon his or her normal occupation and be in fact, if not in name, a professional.

For decades, it had been generally agreed that many of the rules of amateurism were more honoured in the breach than in the observance. Originally these rules were laid down in England more than a century ago by the leisured and, in many cases, financially independent class of society. Numerous followers of amateur sport have felt that since the Olympic Games take place only once in four years, the rules governing general participation in amateur sport should be left entirely to the individual international bodies, and that the IOC should abandon any attempt to produce a series of rules applying to more than a score of different sports. The extensive changes made by the IOC in 1976 materially altered the whole position, and indicated that yet one more step had been taken toward making the Olympic Games open to all comers.

It now appears to many of those who have been closely associated with the Games that they have become too vast, too nationalistic, too expensive, and too commercial. It is felt that no longer is the important thing "to take part," but that winning a gold medal is the only thing that matters. The primary consideration appears no longer to be the competitors, more than 95 percent of whom cannot win gold medals, but the public. Entertainment now seems to supersede the enjoyment of competition, and the demand for entertainment may well be the death of amateur sport.

The Olympic sports

Programs and events. An official Olympic program must include at least 15 of the following sports: archery, athletics, basketball, boxing, canoeing, cycling, diving, equestrian sports, fencing, football, gymnastics, handball (team handball), hockey, judo, modern pentathlon, rowing, shooting, swimming, volleyball, water polo, weight lifting, wrestling, and yachting. Women can participate in the following sports: archery, athletics, basketball, canoeing, diving, equestrian sports, fencing, gymnastics, handball (team handball), hockey, luge, rowing, shooting, figure and speed skating, skiing, swimming, volleyball, and yachting. An Olympic program may also include two demonstration sports and in addition must include exhibitions and demonstrations of fine arts (architecture, literature, music, painting, sculpture, photography, and sports philately).

The particular events included in the different sports are a matter for agreement between the IOC and the international federation. For many years there were 24 events for men in track and field. In 1976, however, the men's 50,000-metre walk was excluded (on the claim that the total number of competitors in the Games must be reduced), while women's rowing events were included for the first time. It appears that although the IOC feels that in some way there must be a halt to the total number of competitors, it has not faced the fact that the only way to produce substantial reductions must be to exclude some sports. The suggestion that all team events be omitted is the kind of action that should be seriously contemplated.

The Olympic Village. The Olympic Village was first introduced at Los Angeles in 1932. The organizing committee provides the village so that competitors and team officials can be housed together and fed at a reasonable price. The villages are located as close as possible to the main stadium and other facilities, and have separate accommodations for men and women. Only competitors and officials may live in the village, and the number of team officials is strictly limited; *e.g.*, a team of 250 competitors is allowed 45 officials.

Ceremonies. *The Olympic flag.* In the stadium and its immediate surroundings, the Olympic flag is flown freely together with the flags of the nations taking part. The Olympic flag presented by Baron de Coubertin in 1914 is the prototype: it has a white background and in the centre there are five interlaced rings—blue, yellow, black, green, and red. The blue ring is on the left next to the pole. These rings represent the five continents joined together in the Olympic Movement. The Olympic motto is *Citius—Altius—Fortius* ("Faster—Higher—Stronger").

The opening ceremony. The form of the opening ceremony is laid down by the IOC in great detail, from the moment when the chief of state of the host country is received by the president of the IOC and the organizing committee at the entrance to the stadium, to the end of the proceedings when the last team files out. The rules provide that participants are not permitted to carry cameras into the arena, but this provision is always ignored.

When the head of state has reached his place in the tribune, he is greeted with the national anthem of his country, and the parade of competitors begins. The Greek team is always the first to enter the stadium, and the nations follow in alphabetical order as determined by the language of the organizing country. Each contingent, dressed in its official uniform, is preceded by a shield with the name of its country, while an athlete carries its national flag. The competitors march around the stadium and then form up in the centre of the ground facing the tribune.

The president of the organizing committee then delivers a brief speech of welcome (not more than two minutes), followed by a speech of not more than three minutes from the president of the IOC, who asks the chief of state to proclaim the Games open.

A fanfare of trumpets is sounded as the Olympic flag is raised slowly; pigeons are released, symbolically to fly to the countries of the world with the news that the Games are open.

The Olympic flame

The Olympic flame is then carried into the stadium by the last of the runners who have brought it from Olympia, Greece. The runner circles the track, mounts the steps, and lights the Olympic fire that burns night and day during the Games. In 1968 a woman carried the flame into the stadium, and in 1976 the flame was jointly borne by a male and a female athlete.

Victory ceremony. Medals are presented during the Games at the various venues and as soon as possible after the conclusion of the event. The competitors who have occupied the first three places proceed to the rostrum, with the winner (gold medalist) in the centre, the silver medalist on his or her right, and the bronze medalist on the left. The medals attached to a chain or ribbon are hung round the necks of the winners by a member of the IOC, and the flags of the nations concerned are raised to the top of the flagpoles while an abbreviated form of the national anthem of the winner is played. The spectators are expected to stand and face the flags, as do the three successful athletes.

Closing ceremony. The closing ceremony takes place after the final event, which is usually the equestrian Prix des Nations. Since the Melbourne Olympics in 1956, the closing ceremony has been a less formal affair, and after certain formalities have been observed, the athletes taking part stage their own demonstrations.

The ceremonies include a parade of six athletes from each nation, marching eight or ten abreast without distinction of nationality, and signifying the friendly bonds of Olympic sport. The president of the IOC calls the youth of the world to assemble in four years to celebrate the Games of the next Olympiad. A fanfare is sounded, the Olympic fire is extinguished, and to the strains of the Olympic anthem the Olympic flag is lowered and the Games are over.

The awarding of the Olympic Games. The honour of holding the Olympic Games is entrusted to a city and not to a country. The choice of the city lies solely with the IOC. Application to hold the Games is made by the chief authority of the city, with the support of the national government.

Applications must state that no political meetings or demonstrations will be held in the stadium or other sports grounds or in the Olympic Village, and it must be promised that every competitor shall be given free entry without any discrimination on grounds of religion, colour, or political affiliation. This involves the assurance that the national government will not refuse visas to any of the competitors. At the Montreal Olympics in 1976, however, the Canadian government refused visas to the representatives of Taiwan because they were unwilling to forgo the title of the Republic of China, under which their National Olympic Committee was admitted to the IOC.

This Canadian decision, in the opinion of the IOC, did great damage to the Olympic Games, and it was later resolved that any country in which the Games are organized must undertake strictly to observe the rules. Enforcement will be difficult unless, for example, the IOC should decide that the Games would not be held, even at the last minute, unless the rules were obeyed.

DEVELOPMENT OF THE GAMES

Since the revival of the Games, their growth in the number of competitions, of competitors, and of participating countries has been almost continuous. The history of the modern Games may be conveniently divided into three periods: (1) from the revival in 1896 to the outbreak of World War I, (2) the period between the two World Wars, and (3) the post-World War II period.

First period: 1896–1912. Of the five celebrations in the first period, the first three, Athens in 1896, Paris in 1900, and St. Louis, U.S., in 1904 were somewhat haphazard affairs. The entries were unlimited and hardly ever "national" in the sense of representing each nation's best performers. The actual events included varied considerably, and it is not possible to determine with any degree of confidence which of the events were really of Olympic significance. Women competed at golf and lawn tennis in the Paris Games and at archery in the St. Louis Games.

London, 1908

The Games of 1908 in London were the first to be organized by the various sporting bodies concerned, and were not regarded as an appendage to a world fair and controlled by the promoters. Though the holding of a Franco-British Exhibition at Shepherd's Bush, London, made possible the construction of a new stadium—the famous White City, with a crowd capacity of 66,000—the governing bodies of sport were in control. Twenty-two nations sent more than 2,000 athletes (but only 36 women) to compete in over 100 events in 21 different sports. Archery and lawn tennis were included in the 1908 program.

At Stockholm in 1912, the number of nations increased to 28, the number of competitors passed the 2,500 mark, and for the first time three swimming events for women were included. Because of World War I, no Games were held in 1916.

Second period: 1920–36. In this period, too, there were five celebrations. In less than two years after the armistice in November 1918, Belgium organized the first post-World War I Olympics, opened at Antwerp by King Albert I. Twenty-nine nations, including many new ones, sent more than 2,000 competitors to take part in over 150 events in 20 sports. Women competitors were still a mere handful (about 60). Four years later in Paris, the number of countries increased to 44, with more than 3,000 competitors for 137 events, a decrease because of a reduction in the excessive number of shooting and yachting events. For the first time more than 100 women took part, appearing in fencing as well as swimming and lawn tennis. After 1924, intrusion of professionalism in lawn tennis caused the IOC to omit the sport from future Games.

Increase of women's events

In 1928 in Amsterdam, women's track and field competitions (five events) and one event for women in gymnastics were introduced. Again there were 3,000, from 46 countries, but the number of women (290) was double that at Paris (136). Four years later the Games for the second time crossed the Atlantic, to Los Angeles, and there was a great falling off in the number of competitors —under 1,500 from 37 countries. The long journey from Europe, in those days 3,000 miles by sea and then the same distance by land, and the heavy cost were responsible for the large decreases. In 1932 the Olympic Village was first introduced. The return to Europe and to Germany produced 4,000 athletes at Berlin in 1936.

Recent games: 1948–1976. Because of World War II, no Games were held in 1940 or 1944. London was given the task of organizing the first postwar Games in 1948 with a bare three years available. This was accomplished despite extremely difficult postwar conditions, with rationing of building and other materials and supplies still in effect and a strict limitation on expenditure. The number of competing countries was 59, with 4,700 competitors of whom 385 were women. Track and field events for women were increased to nine by the addition of the 200-metre run, the broad (now long) jump, and shot put.

Finland, which had hoped to organize the 1940 Games, was host to the 1952 Games, at Helsinki. The number of countries participating rose to 69, competitors totalled within 100 of 5,000, and women topped the 500 mark.

In 1956 for the first time the venue was in the Southern Hemisphere. The Games in Melbourne were celebrated in November and December, and for the first time one of the sports had to take place in another country. Owing to quarantine regulations that prohibited the importation of horses to Australia, the equestrian events were decided in Stockholm. Again the distance from Europe reduced the competitors in Melbourne by 1,500.

In Rome four years later, however, the number of competitors passed the 5,000 mark, the number of nations rose to more than 80, and of events to 150. Originally Rome was to have staged the 1908 Games, so the city had waited over 50 years for the honour.

The Games were in Asia for the first time in 1964, at Tokyo, again in a city where the previous award (in 1940) had not been fulfilled. Once more there were record figures: 94 countries; 162 events; 5,500 competitors, 700 of them women.

There was considerable and mounting criticism when Mexico City was awarded the 1968 Games. Controversy surrounded charges that the altitude, more than 7,500 feet (2,300 metres), would adversely affect the majority of distance runners, who would not have been able to spend many months living in a comparably rare atmosphere. The number of competitors rose to more than 6,000 from 112 countries. Women numbered 800, the total events 172.

In Munich in 1972 there were nearly 200 different events, with 6,000 competitors from 124 countries. Tragedy struck the Games when Arab (Palestinian) terrorists invaded the Olympic Village and killed two and seized nine Israeli athletes as hostages for the release of 200 Arab prisoners in Israel; all nine, five of their captors, and a West German policeman were slain when police rescue attempts failed.

In Montreal in 1976 the number of entries exceeded those of Munich, but the last minute withdrawals reduced the actual competitors by more than 400. The countries endeavoured to persuade the IOC not to permit competitors from New Zealand to take part because the New Zealand rugby team had played matches against South Africa. The IOC rejected the plea, and later decreed that "such occurrences cannot be tolerated in the future." The disqualification of weight-lifting competitors for use of anabolic steroids was another sad incident of the games. Two of the competitors were gold medallists, who were ordered to return their medals.

Montreal, 1976

The Winter Olympic Games. While some skating events were included in the 1908 and 1920 Games, the Winter Games were accepted as a celebration comparable to the Summer Games and given the official blessing of the IOC in 1924. The first Winter Games were held at Chamonix, France, and consisted of 16 events. There were 16 participating countries, and the participants numbered less than 300. Subsequent Winter Games were held at St. Moritz, Switzerland (1928); Lake Placid, U.S. (1932); Garmisch-Partenkirchen, Germany (1936); St. Moritz (1948); Oslo (1952); Cortina d'Ampezzo, Italy (1956); Squaw Valley, U.S. (1960); Innsbruck, Austria (1964); Grenoble, France (1968); Sapporo, Japan (1972); and Innsbruck (1976). The 1980 Winter Games were to be held at Lake Placid.

So much dissension has arisen over the commercialism of the Winter Games that, from time to time, there has been talk of revising, or even discontinuing, them.

THE OLYMPIC RECORD

The accompanying tables of the Olympic record list the gold medal winners of each event in each of the games held since the revival of the Olympic idea in 1896. Unofficial, or demonstration, events are not included.

Olympic Champions, 1896–1976

Athletics (track-and-field) (men)

100 metres

			sec
1896	T. Burke	U.S.	12.0
1900	F. Jarvis	U.S.	11.0
1904	A. Hahn	U.S.	11.0
1908	R. Walker	S.Africa	10.8
1912	R. Craig	U.S.	10.8
1920	C. Paddock	U.S.	10.8
1924	H. Abrahams	Gt.Brit.	10.6
1928	P. Williams	Canada	10.8
1932	E. Tolan	U.S.	10.3
1936	J. Owens	U.S.	10.3
1948	H. Dillard	U.S.	10.3
1952	L. Remigino	U.S.	10.4
1956	R. Morrow	U.S.	10.5
1960	A. Hary	Ger.*	10.2
1964	R. Hayes	U.S.	10.0
1968	J. Hines	U.S.	9.9
1972	V. Borzov	U.S.S.R.	10.14†
1976	H. Crawford	Trinidad and Tobago	10.06

200 metres

			sec
1900	J. Tewksbury	U.S.	22.2
1904	A. Hahn	U.S.	21.6
1908	R. Kerr	Canada	22.6
1912	R. Craig	U.S.	21.7
1920	A. Woodring	U.S.	22.0
1924	J. Scholz	U.S.	21.6
1928	P. Williams	Canada	21.8
1932	E. Tolan	U.S.	21.2
1936	J. Owens	U.S.	20.7
1948	M. Patton	U.S.	21.1
1952	A. Stanfield	U.S.	20.7
1956	R. Morrow	U.S.	20.6
1960	L. Berruti	Italy	20.5
1964	H. Carr	U.S.	20.3
1968	T. Smith	U.S.	19.8
1972	V. Borzov	U.S.S.R.	20.00†
1976	D. Quarrie	Jamaica	20.23

400 metres

			sec
1896	T. Burke	U.S.	54.2
1900	M. Long	U.S.	49.4
1904	H. Hillman	U.S.	49.2
1908	W. Halswelle	Gt.Brit.	50.0
1912	C. Reidpath	U.S.	48.2
1920	B. Rudd	S.Africa	49.6
1924	E. Liddell	Gt.Brit.	47.6
1928	R. Barbuti	U.S.	47.8
1932	W. Carr	U.S.	46.2
1936	A. Williams	U.S.	46.5
1948	A. Wint	Jam.	46.2
1952	G. Rhoden	Jam.	45.9
1956	C. Jenkins	U.S.	46.7
1960	O. Davis	U.S.	44.9
1964	M. Larrabee	U.S.	45.1
1968	L. Evans	U.S.	43.8
1972	V. Matthews	U.S.	44.66†
1976	A. Juantorena	Cuba	44.26

800 metres

			min	sec
1896	E. Flack	Australia	2	11.0
1900	A. Tysoe	Gt.Brit.	2	01.2
1904	J. Lightbody	U.S.	1	56.0
1908	M. Sheppard	U.S.	1	52.8
1912	J. Meredith	U.S.	1	51.9
1920	A. Hill	Gt.Brit.	1	53.4
1924	D. Lowe	Gt.Brit.	1	52.4
1928	D. Lowe	Gt.Brit.	1	51.8
1932	T. Hampson	Gt.Brit.	1	49.7
1936	J. Woodruff	U.S.	1	52.9
1948	M. Whitfield	U.S.	1	49.2
1952	M. Whitfield	U.S.	1	49.2
1956	T. Courtney	U.S.	1	47.7
1960	P. Snell	N.Z.	1	46.3
1964	P. Snell	N.Z.	1	45.1
1968	R. Doubell	Australia	1	44.3
1972	D. Wottle	U.S.	1	45.9
1976	A. Juantorena	Cuba	1	43.5

1,500 metres

			min	sec
1896	E. Flack	Australia	4	33.2
1900	C. Bennett	Gt.Brit.	4	06.2
1904	J. Lightbody	U.S.	4	05.4
1908	M. Sheppard	U.S.	4	03.4
1912	A. Jackson	Gt.Brit.	3	56.8
1920	A. Hill	Gt.Brit.	4	01.8
1924	P. Nurmi	Finland	3	53.6
1928	H. Larva	Finland	3	53.2
1932	L. Beccali	Italy	3	51.2
1936	J. Lovelock	N.Z.	3	47.8
1948	H. Eriksson	Sweden	3	49.8
1952	J. Barthel	Luxembourg	3	45.1
1956	R. Delany	Ireland	3	41.2
1960	H. Elliott	Australia	3	35.6
1964	P. Snell	N.Z.	3	38.1
1968	K. Keino	Kenya	3	34.9
1972	P. Vasala	Finland	3	36.3
1976	J. Walker	N.Z.	3	39.2

5,000 metres

			min	sec
1912	H. Kolehmainen	Finland	14	36.6
1920	J. Guillemot	France	14	55.6
1924	P. Nurmi	Finland	14	31.2
1928	V. Ritola	Finland	14	38.0
1932	L. Lehtinen	Finland	14	30.0
1936	G. Höckert	Finland	14	22.2
1948	G. Reiff	Belgium	14	17.6
1952	E. Zátopek	Czech.	14	06.6
1956	V. Kuts	U.S.S.R.	13	39.6
1960	M. Halberg	N.Z.	13	43.4
1964	R. Schul	U.S.	13	48.8
1968	M. Gammoudi	Tunisia	14	05.0
1972	L. Viren	Finland	13	26.4
1976	L. Viren	Finland	13	24.8

10,000 metres

			min	sec
1912	H. Kolehmainen	Finland	31	20.8
1920	P. Nurmi	Finland	31	45.8
1924	V. Ritola	Finland	30	23.2
1928	P. Nurmi	Finland	30	18.8
1932	J. Kusocinski	Poland	30	11.4
1936	I. Salminen	Finland	30	15.4
1948	E. Zátopek	Czech.	29	59.6
1952	E. Zátopek	Czech.	29	17.0
1956	V. Kuts	U.S.S.R.	28	45.6
1960	P. Bolotnikov	U.S.S.R.	28	32.2
1964	W. Mills	U.S.	28	24.4
1968	N. Temu	Kenya	29	27.4
1972	L. Viren	Finland	27	38.4
1976	L. Viren	Finland	27	40.4

marathon

			hr	min	sec
1896	S. Louis	Greece	2	58	50.0
1900	M. Theato	France	2	59	45.0
1904	T. Hicks	U.S.	3	28	53.0
1908	J. Hayes	U.S.	2	55	18.4
1912	K. McArthur	S.Africa	2	36	54.8
1920	H. Kolehmainen	Finland	2	32	35.8
1924	A. Stenroos	Finland	2	41	22.6
1928	A. El Ouafi	France	2	32	57.0
1932	J. Zabala	Arg.	2	31	36.0
1936	K. Son	Japan	2	29	19.2
1948	D. Cabrera	Arg.	2	34	51.6
1952	E. Zátopek	Czech.	2	23	03.2
1956	A. Mimoun	France	2	25	00.0
1960	A. Bikila	Eth.	2	15	16.2
1964	A. Bikila	Eth.	2	12	11.2
1968	M. Wolde	Eth.	2	20	26.4
1972	F. Shorter	U.S.	2	12	19.8
1976	W. Cierpinski	E.Ger.	2	09	55.0

110-metre hurdles

			sec
1896‡	T. Curtis	U.S.	17.6
1900	A. Kraenzlein	U.S.	15.4
1904	F. Schule	U.S.	16.0
1908	F. Smithson	U.S.	15.0
1912	F. Kelly	U.S.	15.1
1920	E. Thomson	Canada	14.8
1924	D. Kinsey	U.S.	15.0
1928	S. Atkinson	S.Africa	14.8
1932	G. Saling	U.S.	14.6
1936	F. Towns	U.S.	14.2
1948	W. Porter	U.S.	13.9
1952	H. Dillard	U.S.	13.7
1956	L. Calhoun	U.S.	13.5
1960	L. Calhoun	U.S.	13.8
1964	H. Jones	U.S.	13.6
1968	W. Davenport	U.S.	13.3
1972	R. Milburn	U.S.	13.24†
1976	G. Drut	France	13.30

400-metre hurdles

			sec
1900	J. Tewksbury	U.S.	57.6
1904§	H. Hillman	U.S.	53.0
1908	C. Bacon	U.S.	55.0
1920	F. Loomis	U.S.	54.0
1924	F. Taylor	U.S.	52.6
1928	Lord Burghley	Gt.Brit.	53.4
1932	R. Tisdall	Ireland	51.7
1936	G. Hardin	U.S.	52.4
1948	R. Cochran	U.S.	51.1
1952	C. Moore	U.S.	50.8
1956	G. Davis	U.S.	50.1
1960	G. Davis	U.S.	49.3
1964	W. Cawley	U.S.	49.6
1968	D. Hemery	Gt.Brit.	48.1
1972	J. Akii-Bua	Uganda	47.82†
1976	E. Moses	U.S.	47.64

3,000-metre steeplechase

			min	sec
1920	P. Hodge	Gt.Brit.	10	00.4
1924	V. Ritola	Finland	9	33.6
1928	T. Loukola	Finland	9	21.8
1932	V. Iso-Hollo	Finland	10	33.4‖
1936	V. Iso-Hollo	Finland	9	03.8
1948	T. Sjöstrand	Sweden	9	04.6
1952	H. Ashenfelter	U.S.	8	45.4
1956	C. Brasher	Gt.Brit.	8	41.2
1960	Z. Krzyszkowiak	Poland	8	34.2
1964	G. Roelants	Belgium	8	30.8
1968	A. Biwott	Kenya	8	51.0
1972	K. Keino	Kenya	8	23.6
1976	A. Garderud	Sweden	8	08.0

4 × 100 metre relay

			sec
1912	Gt.Brit.		42.4
1920	U.S.		42.2
1924	U.S.		41.0
1928	U.S.		41.0
1932	U.S.		40.0
1936	U.S.		39.8
1948	U.S.		40.6
1952	U.S.		40.1
1956	U.S.		39.5
1960	Ger.*		39.5
1964	U.S.		39.0
1968	U.S.		38.2
1972	U.S.		38.19†
1976	U.S.		38.33

4 × 400 metre relay

			min	sec
1912	U.S.		3	16.6
1920	Gt.Brit.		3	22.2
1924	U.S.		3	16.0
1928	U.S.		3	14.2
1932	U.S.		3	08.2
1936	Gt.Brit.		3	09.0
1948	U.S.		3	10.4
1952	Jam.		3	03.9
1956	U.S.		3	04.8
1960	U.S.		3	02.2
1964	U.S.		3	00.7
1968	U.S.		2	56.1
1972	Kenya		2	59.8
1976	U.S.		2	58.7

20,000-metre walk

			hr	min	sec
1956	L. Spirin	U.S.S.R.	1	31	27.4
1960	V. Golubnichy	U.S.S.R.	1	34	07.2
1964	K. Matthews	Gt.Brit.	1	29	34.0
1968	V. Golubnichy	U.S.S.R.	1	33	58.4
1972	P. Frenkel	W.Ger.	1	26	42.6
1976	D. Bautista	Mexico	1	24	40.6

*Joint East–West German team. †Race first timed in hundredths of a second. ‡Distance, 100 metres. §Hurdles were two feet six inches high, not three feet.
‖An extra lap of 460 metres was run in error.

Olympic Champions, 1896–1976 (continued)

Athletics (track-and-field) (men) (continued)

*50,000-metre walk**

			hr	min	sec
1932	T. Green	Gt.Brit.	4	50	10.0
1936	H. Whitlock	Gt.Brit.	4	30	41.4
1948	J. Ljunggren	Sweden	4	41	52.0
1952	G. Dordoni	Italy	4	28	07.8
1956	N. Read	N.Z.	4	30	42.8
1960	D. Thompson	Gt.Brit.	4	25	30.0
1964	A. Pamich	Italy	4	11	12.4
1968	C. Hohne	E.Ger.	4	20	13.6
1972	B. Kannenberg	W.Ger.	3	56	11.6

high jump

			metres
1896	E. Clark	U.S.	1.81
1900	I. Baxter	U.S.	1.90
1904	S. Jones	U.S.	1.80
1908	H. Porter	U.S.	1.90
1912	A. Richards	U.S.	1.93
1920	R. Landon	U.S.	1.93
1924	H. Osborn	U.S.	1.98
1928	R. King	U.S.	1.94
1932	D. McNaughton	Canada	1.97
1936	C. Johnson	U.S.	2.03
1948	J. Winter	Australia	1.98
1952	W. Davis	U.S.	2.04
1956	C. Dumas	U.S.	2.12
1960	R. Shavlakadze	U.S.S.R.	2.16
1964	V. Brumel	U.S.S.R.	2.18
1968	R. Fosbury	U.S.	2.24
1972	Y. Tarmak	U.S.S.R.	2.23
1976	J. Wszola	Poland	2.25

pole vault

			metres
1896	W. Hoyt	U.S.	3.30
1900	I. Baxter	U.S.	3.30
1904	C. Dvorak	U.S.	3.50
1908	E. Cooke		
	A. Gilbert	U.S.	3.71
1912	H. Babcock	U.S.	3.95
1920	F. Foss	U.S.	4.09
1924	L. Barnes	U.S.	3.95
1928	S. Carr	U.S.	4.20
1932	W. Miller	U.S.	4.31
1936	E. Meadows	U.S.	4.35
1948	O. Smith	U.S.	4.30
1952	R. Richards	U.S.	4.55
1956	R. Richards	U.S.	4.56
1960	D. Bragg	U.S.	4.70
1964	F. Hansen	U.S.	5.10
1968	B. Seagren	U.S.	5.40
1972	W. Nordwig	E.Ger.	5.50
1976	T. Slusarski	Poland	5.50

long jump

			metres
1896	E. Clark	U.S.	6.35
1900	A. Kraenzlein	U.S.	7.18
1904	M. Prinstein	U.S.	7.34
1908	F. Irons	U.S.	7.48
1912	A. Gutterson	U.S.	7.60
1920	W. Pettersson	Swed.	7.15
1924	H. de Hubbard	U.S.	7.44
1928	E. Hamm	U.S.	7.73
1932	E. Gordon	U.S.	7.64
1936	J. Owens	U.S.	8.06
1948	W. Steele	U.S.	7.82
1952	J. Biffle	U.S.	7.57
1956	G. Bell	U.S.	7.83
1960	R. Boston	U.S.	8.12
1964	L. Davies	Gt.Brit.	8.07
1968	R. Beamon	U.S.	8.90
1972	R. Williams	U.S.	8.24
1976	A. Robinson	U.S.	8.35

triple jump

			metres
1896	J. Connolly	U.S.	13.71
1900	M. Prinstein	U.S.	14.47
1904	M. Prinstein	U.S.	14.35
1908	T. Ahearne	Gt.Brit.	14.91
1912	G. Lindblom	Sweden	14.76
1920	V. Tuulos	Finland	14.50
1924	A. Winter	Australia	15.53
1928	M. Oda	Japan	15.21
1932	C. Nambu	Japan	15.72
1936	N. Tajima	Japan	16.00
1948	A. Ahman	Sweden	15.40
1952	A. da Silva	Brazil	16.22
1956	A. da Silva	Brazil	16.35
1960	J. Schmidt	Poland	16.81
1964	J. Schmidt	Poland	16.85
1968	V. Saneyev	U.S.S.R.	17.39
1972	V. Saneyev	U.S.S.R.	17.35
1976	V. Saneyev	U.S.S.R.	17.29

shot put

			metres
1896	R. Garrett	U.S.	11.22
1900	R. Sheldon	U.S.	14.10
1904	R. Rose	U.S.	14.81
1908	R. Rose	U.S.	14.21
1912	P. McDonald	U.S.	15.34
1920	V. Pörhölä	Fin.	14.81
1924	C. Houser	U.S.	14.99
1928	J. Kuck	U.S.	15.87
1932	L. Sexton	U.S.	16.00
1936	H. Woellke	Ger.	16.20
1948	W. Thompson	U.S.	17.12
1952	P. O'Brien	U.S.	17.41
1956	P. O'Brien	U.S.	18.57
1960	W. Nieder	U.S.	19.68
1964	D. Long	U.S.	20.33
1968	R. Matson	U.S.	20.54
1972	W. Komar	Poland	21.18
1976	U. Beyer	E.Ger.	21.05

discus throw

			metres
1896	R. Garrett	U.S.	29.15
1900	R. Bauer	Hung.	36.04
1904	M. Sheridan	U.S.	39.28
1908	M. Sheridan	U.S.	40.89
1912	A. Taipale	Fin.	45.21
1920	E. Niklander	Fin.	44.68
1924	C. Houser	U.S.	46.15
1928	C. Houser	U.S.	47.32
1932	J. Anderson	U.S.	49.49
1936	K. Carpenter	U.S.	50.48
1948	A. Consolini	Italy	52.78
1952	S. Iness	U.S.	55.03
1956	A. Oerter	U.S.	56.36
1960	A. Oerter	U.S.	59.18
1964	A. Oerter	U.S.	61.00
1968	A. Oerter	U.S.	64.78
1972	L. Danek	Czech.	64.40
1976	M. Wilkins	U.S.	67.50

hammer throw

			metres
1900	J. Flanagan	U.S.	49.73
1904	J. Flanagan	U.S.	51.23
1908	J. Flanagan	U.S.	51.92
1912	M. McGrath	U.S.	54.74
1920	P. Ryan	U.S.	52.87
1924	F. Tootell	U.S.	53.30
1928	P. O'Callaghan	Ireland	51.39
1932	P. O'Callaghan	Ireland	53.92
1936	K. Hein	Germany	56.49
1948	I. Németh	Hung.	56.07
1952	J. Csermák	Hung.	60.34
1956	H. Connolly	U.S.	63.19
1960	V. Rudenkov	U.S.S.R.	67.10
1964	R. Klim	U.S.S.R.	69.74
1968	G. Zsivótzky	Hung.	73.36
1972	A. Bondarchuk	U.S.S.R.	75.50
1976	Y. Sedykh	U.S.S.R.	77.52

javelin throw

			metres
1908	E. Lemming	Sweden	54.83
1912	E. Lemming	Sweden	60.64
1920	J. Myyrä	Finland	65.78
1924	J. Myyrä	Finland	62.96
1928	E. Lundkvist	Sweden	66.60
1932	M. Järvinen	Finland	72.71
1936	G. Stöck	Germany	71.84
1948	T. Rautavaara	Finland	69.77
1952	C. Young	U.S.	73.78
1956	E. Danielson	Norway	85.71
1960	V. Tsybulenko	U.S.S.R.	84.64
1964	P. Nevala	Finland	82.66
1968	Y. Lusis	U.S.S.R.	90.10
1972	K. Wolfermann	W.Ger.	90.48
1976	M. Nemeth	Hung.	94.58

decathlon

1912	H. Wieslander	Swed.
1920	H. Lövland	Nor.
1924	H. Osborn	U.S.
1928	P. Yrjölä	Fin.
1932	J. Bausch	U.S.
1936	G. Morris	U.S.
1948	R. Mathias	U.S.
1952	R. Mathias	U.S.
1956	M. Campbell	U.S.
1960	R. Johnson	U.S.
1964	W. Holdorf	Ger.†
1968	W. Toomey	U.S.
1972	N. Avilov	U.S.S.R.
1976	B. Jenner	U.S.

Athletics (track-and-field) events no longer included (men)

60 metres

			sec
1900	A. Kraenzlein	U.S.	7.0
1904	A. Hahn	U.S.	7.0

3,000-metre team race

		individual winner		min	sec
1912	U.S.	T. Berna	U.S.	8	44.6
1920	U.S.	R. Brown	U.S.	8	45.4
1924	Finland	P. Nurmi	Finland	8	32.0

3-mile team race

				min	sec
1908	Gt.Brit.	J. Deakin	Gt.Brit.	14	39.6

5,000 metre team race

				min	sec
1900	Gt.Brit.	C. Bennett	Gt.Brit.	15	20.0

4-mile team race

				min	sec
1904	U.S.	A. Newton	U.S.	21	17.8

5 miles (individual)

			min	sec
1908	E. Voight	Gt.Brit.	25	11.2

1,600-metre relay (200 × 200 × 400 × 800 metres)

			min	sec
1908	U.S.		3	29.4

cross-country (team)

		individual winner		min	sec
1912	(about 5 mi)	Swed.	H. Kolehmainen	Fin.	45	11.6
1920	(8,000 m)	Fin.	P. Nurmi	Fin.	27	15.0
1924	(10,000 m)	Fin.	P. Nurmi	Fin.	32	54.8

steeplechase

				min	sec
1900	(2,500 m)	G. Orton	U.S.	7	34.4
1900	(4,000 m)	J. Rimmer	Gt.Brit.	12	58.4
1908	(3,200 m)	A. Russell	Gt.Brit.	10	47.8

walks

				hr	min	sec
1908	(3,500 m)	G. Larner	Gt.Brit.		14	55.0
1908	(10 mi)	G. Larner	Gt.Brit.	1	15	57.4
1912	(10,000 m)	G. Goulding	Canada		46	28.4
1920	(3,000 m)	U. Frigerio	Italy		13	14.2
1920	(10,000 m)	U. Frigerio	Italy		48	06.2
1924	(10,000 m)	U. Frigerio	Italy		47	49.0
1948	(10,000 m)	J. Mikaelsson	Swed.		45	13.2
1952	(10,000 m)	J. Mikaelsson	Swed.		45	02.8

200-metre hurdles

			sec
1900	A. Kraenzlein	U.S.	25.4
1904	H. Hillman	U.S.	24.6

standing high jump

			metres
1900	R. Ewry	U.S.	1.65
1904	R. Ewry	U.S.	1.50
1908	R. Ewry	U.S.	1.57
1912	P. Adams	U.S.	1.63

*Not held in 1976. †Joint East–West German team.

Olympic Champions, 1896–1976 (continued)

Athletics (track-and-field) events no longer included (continued)

standing long jump

			metres
1900	R. Ewry	U.S.	3.21
1904	R. Ewry	U.S.	3.47
1908	R. Ewry	U.S.	3.33
1912	C. Tsiklitiras	Greece	3.37

standing triple jump

			metres
1900	R. Ewry	U.S.	10.58
1904	R. Ewry	U.S.	10.54

shot-putting (both hands)

			metres
1912	R. Rose	U.S.	27.70

throwing the discus (Greek style)

			metres
1908	M. Sheridan	U.S.	37.99

throwing the discus (both hands)

			metres
1912	A. Taipale	Finland	82.86

throwing the javelin (free style)

			metres
1908	E. Lemming	Sweden	54.44

throwing the javelin (both hands)

			metres
1912	J. Saaristo	Finland	109.42

throwing the 56 lb weight

			metres
1904	E. Desmarteau	Canada	10.46
1920	P. McDonald	U.S.	11.26

tug-of-war

1900	Sweden-Denmark	
1904	U.S.	
1908	Great Britain	
1912	Sweden	
1920	Great Britain	

pentathlon

1912	F. Bie	Norway	
1920	E. Lehtonen	Finland	
1924	E. Lehtonen	Finland	

Athletics (track-and-field) (women)

100 metres

			sec
1928	E. Robinson	U.S.	12.2
1932	S. Walasiewicz	Poland	11.9
1936	H. Stephens	U.S.	11.5
1948	F. Blankers-Koen	Neth.	11.9
1952	M. Jackson	Australia	11.5
1956	B. Cuthbert	Australia	11.5
1960	W. Rudolph	U.S.	11.0
1964	W. Tyus	U.S.	11.4
1968	W. Tyus	U.S.	11.0
1972	R. Stecher	E.Ger.	11.07*
1976	A. Richter	W.Ger.	11.08

200 metres

			sec
1948	F. Blankers-Koen	Neth.	24.4
1952	M. Jackson	Australia	23.7
1956	B. Cuthbert	Australia	23.4
1960	W. Rudolph	U.S.	24.0
1964	E. McGuire	U.S.	23.0
1968	I. Szewinska	Poland	22.5
1972	R. Stecher	E.Ger.	22.40*
1976	B. Eckert	E.Ger.	22.37

400 metres

			sec
1964	B. Cuthbert	Australia	52.0
1968	C. Besson	France	52.0
1972	M. Zehrt	E.Ger.	51.08*
1976	I. Szewinska	Poland	49.29

800 metres

			min	sec
1928	L. Radke-Batschauer	Germany	2	16.8
1960	L. Lysenko-Shevtsova	U.S.S.R.	2	04.3
1964	A. Packer	Gt.Brit.	2	01.1
1968	M. Manning	U.S.	2	00.9
1972	H. Falcke	W.Ger.	1	58.6
1976	T. Kazankina	U.S.S.R.	1	54.9

1,500 metres

			min	sec
1972	L. Bragina	U.S.S.R.	4	01.4
1976	T. Kazankina	U.S.S.R.	4	05.5

4 × 100-metre relay

		sec
1928	Canada	48.4
1932	U.S.	47.0
1936	U.S.	46.9
1948	Neth.	47.5
1952	U.S.	45.9
1956	Australia	44.5
1960	U.S.	44.5
1964	Poland	43.6
1968	U.S.	42.8
1972	W.Ger.	42.8
1976	E.Ger.	42.55*

4 × 400-metre relay

		min	sec
1972	E.Ger.	3	23.0
1976	E.Ger.	3	19.2

80-metre hurdles (100 metres from 1972)

			sec
1932	M. Didrikson	U.S.	11.7
1936	T. Valla	Italy	11.7
1948	F. Blankers-Koen	Neth.	11.2
1952	S. Strickland de La Hunty	Australia	10.9
1956	S. Strickland de La Hunty	Australia	10.7
1960	I. Press	U.S.S.R.	10.8
1964	K. Balzer	Germany†	10.5
1968	M. Caird	Australia	10.3
1972	A. Ehrhardt	E.Ger.	12.6
1976	J. Schaller	E.Ger.	12.77*

high jump

			metres
1928	E. Catherwood	Can.	1.59
1932	J. Shiley	U.S.	1.66
1936	I. Csák	Hung.	1.60
1948	A. Coachman	U.S.	1.68
1952	E. Brand	S.Af.	1.67
1956	M. McDaniel	U.S.	1.76
1960	I. Balas	Rom.	1.85
1964	I. Balas	Rom.	1.90
1968	M. Rezkova	Czech.	1.82
1972	U. Meyfarth	W.Ger.	1.92
1976	R. Ackermann	E.Ger.	1.93

long jump

			metres
1948	V. Gyarmati	Hung.	5.69
1952	Y. Williams	N.Z.	6.24
1956	E. Krzesinska	Poland	6.35
1960	V. Krepkina	U.S.S.R.	6.37
1964	M. Rand	Gt.Brit.	6.76
1968	V. Viscopoleanu	Romania	6.82
1972	H. Rosendahl	W.Ger.	6.78
1976	A. Voigt	E.Ger.	6.72

discus throw

			metres
1928	H. Konopacka	Poland	39.62
1932	L. Copeland	U.S.	40.58
1936	G. Mauermayer	Germany	47.63
1948	M. Ostermeyer	France	41.92
1952	N. Romashkova	U.S.S.R.	51.42
1956	O. Fikotova	Czech.	53.69
1960	N. Ponomareva	U.S.S.R.	55.10
1964	T. Press	U.S.S.R.	57.27
1968	L. Manoliu	Rom.	58.28
1972	F. Melnik	U.S.S.R.	66.62
1976	E. Schlaak	E.Ger.	69.00

shot put

			metres
1948	M. Ostermeyer	France	13.75
1952	G. Zybina	U.S.S.R.	15.28
1956	T. Tyshkevich	U.S.S.R.	16.59
1960	T. Press	U.S.S.R.	17.32
1964	T. Press	U.S.S.R.	18.14
1968	M. Gummel	E.Ger.	19.61
1972	N. Chizhova	U.S.S.R.	21.03
1976	I. Christova	Bulgaria	21.16

javelin throw

			metres
1932	M. Didrikson	U.S.	43.68
1936	T. Fleischer	Germany	45.18
1948	H. Bauma	Austria	45.57
1952	D. Zatopkova	Czech.	50.47
1956	I. Yaunzeme	U.S.S.R.	53.86
1960	E. Ozolina	U.S.S.R.	55.98
1964	M. Penes	Romania	60.54
1968	A. Nemeth	Hung.	60.36
1972	R. Fuchs	E.Ger.	63.88
1976	R. Fuchs	E.Ger.	65.94

pentathlon

1964	I. Press	U.S.S.R.
1968	I. Becker	W.Ger.
1972	M. Peters	Gt.Brit.
1976	S. Siegl	E.Ger.

Archery

men's individual

			points
1972	J. Williams	U.S.	2,528
1976	D. Pace	U.S.	2,571

women's individual

1972	D. Wilbert	U.S.	2,424
1976	L. Ryon	U.S.	2,499

For archery events no longer included, *see* page 293.

Boxing

light flyweight

1968	F. Rodríguez	Venezuela
1972	G. Gedo	Hung.
1976	J. Hernández	Cuba

flyweight

1904	G. Finnegan	U.S.
1920	F. De Genaro	U.S.
1924	F. La Barba	U.S.
1928	A. Kocsis	Hung.
1932	I. Enekes	Hung.
1936	W. Kaiser	Ger.
1948	P. Perez	Arg.
1952	N. Brooks	U.S.
1956	T. Spinks	Gt.Brit.
1960	G. Török	Hung.
1964	F. Atzori	Italy
1968	R. Delgado	Mexico
1972	G. Kostadinov	Bulg.
1976	L. Randolph	U.S.

bantamweight

1904	O. Kirk	U.S.
1908	H. Thomas	Gt.Brit.
1920	C. Walker	S.Af.
1924	W. Smith	S.Af.
1928	V. Tamagnini	Italy
1932	H. Gwynne	Canada
1936	U. Sergo	Italy
1948	T. Csik	Hung.
1952	P. Hämäläinen	Finland
1956	W. Behrendt	Germany†
1960	O. Grigoryev	U.S.S.R.
1964	T. Sakurai	Japan
1968	V. Sokolov	U.S.S.R.
1972	O. Martínez	Cuba
1976	Y.J. Gu	N.Kor.

featherweight

1904	O. Kirk	U.S.
1908	R. Gunn	Gt.Brit.
1920	P. Fritsch	France
1924	J. Fields	U.S.
1928	L. van Kleveren	Neth.
1932	C. Robledo	Arg.
1936	G. Casonovas	Arg.
1948	E. Formenti	Italy
1952	J. Zachara	Czech.
1956	V. Safronov	U.S.S.R.
1960	F. Musso	Italy
1964	S. Stepashkin	U.S.S.R.
1968	A. Roldan	Mexico
1972	B. Kausnetsov	U.S.S.R.
1976	A. Herrera	Cuba

*Race first timed in hundredths of a second. †Joint East–West German team.

Olympic Champions, 1896–1976 (continued)

Boxing (continued)

lightweight

1904	H. Spanger	U.S.
1908	F. Grace	Gt.Brit.
1920	S. Mosberg	U.S.
1924	H. Nielsen	Den.
1928	C. Orlandi	Italy
1932	L. Stevens	S.Africa
1936	I. Harangi	Hung.
1948	G. Dreyer	S.Africa
1952	A. Bolognesi	Italy
1956	R. McTaggart	Gt.Brit.
1960	K. Pazdzior	Poland
1964	G. Grudzien	Poland
1968	R. Harris	U.S.
1972	J. Szczepanski	Poland
1976	H. Davis	U.S.

light welterweight

1952	C. Adkins	U.S.
1956	V. Engibaryan	U.S.S.R.
1960	B. Nemecek	Czech.
1964	J. Kulej	Poland
1968	J. Kulej	Poland
1972	R. Seales	U.S.
1976	R. Leonard	U.S.

welterweight

1904	A. Young	U.S.
1920	T. Schneider	Canada
1924	J. Delarge	Belgium
1928	E. Morgan	N.Z.
1932	E. Flynn	U.S.
1936	S. Suvio	Finland
1948	J. Torma	Czech.
1952	Z. Chychla	Poland
1956	N. Linca	Romania
1960	G. Benvenuti	Italy
1964	M. Kasprzyk	Poland
1968	M. Wolke	E.Ger.
1972	E. Correa	Cuba
1976	J. Bachfeld	E.Ger.

light middleweight

1952	L. Papp	Hung.
1956	L. Papp	Hung.
1960	W. McClure	U.S.
1964	B. Lagutin	U.S.S.R.
1968	B. Lagutin	U.S.S.R.
1972	D. Kottysch	W.Ger.
1976	J. Rybicki	Poland

middleweight

1904	C. Mayer	U.S.
1908	J. Douglas	Gt.Brit.
1920	H. Mallin	Gt.Brit.
1924	H. Mallin	Gt.Brit.
1928	P. Toscani	Italy
1932	C. Barth	U.S.
1936	J. Despeaux	France
1948	L. Papp	Hung.
1952	F. Patterson	U.S.
1956	G. Chatkov	U.S.S.R.
1960	E. Crook	U.S.
1964	V. Popenchenko	U.S.S.R.
1968	C. Finnegan	Gt.Brit.
1972	V. Lemechev	U.S.S.R.
1976	M. Spinks	U.S.

light heavyweight

1920	E. Eagan	U.S.
1924	H. Mitchell	Gt.Brit.
1928	V. Avendano	Arg.
1932	D. Carstens	S.Af.
1936	R. Michelot	France
1948	G. Hunter	S.Af.
1952	N. Lee	U.S.
1956	J. Boyd	U.S.
1960	C. Clay	U.S.
1964	C. Pinto	Italy
1968	D. Pozdniak	U.S.S.R.
1972	M. Pavlov	Yugos.
1976	L. Spinks	U.S.

heavyweight

1904	S. Berger	U.S.
1908	A. Oldman	Gt.Brit.
1920	R. Rawson	Gt.Brit.
1924	O. Von Porat	Norway
1928	R. Jurado	Arg.
1932	A. Lovell	Arg.
1936	H. Runge	Germany
1948	R. Iglesias	Arg.
1952	E. Sanders	U.S.
1956	P. Rademacher	U.S.
1960	F. de Piccoli	Italy
1964	J. Frazier	U.S.
1968	G. Foreman	U.S.
1972	T. Stevenson	Cuba
1976	T. Stevenson	Cuba

Canoeing (men)

kayak singles (500 metres)

			min	sec
1976	V. Diba	Romania	1	46.41

kayak pairs (500 metres)

1976	E.Ger.		1	35.87

kayak singles (1,000 metres)

1936	G. Hradetzky	Austria	4	22.90
1948	G. Fredriksson	Sweden	4	33.20
1952	G. Fredriksson	Sweden	4	07.90
1956	G. Fredriksson	Sweden	4	12.80
1960	E. Hansen	Denmark	3	53.00
1964	R. Peterson	Sweden	3	57.13
1968	M. Hesz	Hung.	4	03.58
1972	A. Shaparenko	U.S.S.R.	3	48.06
1976	R. Helm	E.Ger.	3	48.20

kayak pairs (1,000 metres)

			min	sec
1936	Austria	4		03.80
1948	Sweden	4		07.30
1952	Finland	3		51.10
1956	Germany*	3		49.60
1960	Sweden	3		34.70
1964	Sweden	3		38.54
1968	U.S.S.R.	3		37.54
1972	U.S.S.R.	3		31.23
1976	U.S.S.R.	3		29.01

kayak fours (1,000 metres)

			min	sec
1964	U.S.S.R.	3		14.67
1968	Norway	3		14.38
1972	U.S.S.R.	3		15.07
1976	U.S.S.R.	3		08.69

slalom kayak singles†

1972	S. Horn	E.Ger.

Canadian singles (500 metres)

			min	sec
1976	A. Rogov	U.S.S.R.	1	59.23

Canadian singles (1,000 metres)

1936	F. Amyot	Can.	5	32.10
1948	J. Holeček	Czech.	5	42.00
1952	J. Holeček	Czech.	4	56.30
1956	L. Rottman	Rom.	5	05.30
1960	J. Parti	Hung.	4	33.03
1964	J. Eschert	Ger.*	4	35.14
1968	T. Tatai	Hung.	4	36.14
1972	I. Patzaichin	Rom.	4	08.94
1976	M. Ljubek	Yugos.	4	09.51

Canadian pairs (500 metres)

			min	sec
1976	U.S.S.R.		1	45.81

Canadian pairs (1,000 metres)

1936	Czech.	4		50.10
1948	Czech.	5		07.10
1952	Denmark	4		38.30
1956	Romania	4		47.40
1960	U.S.S.R.	4		17.04
1964	U.S.S.R.	4		04.65
1968	Romania	4		07.18
1972	U.S.S.R.	3		52.60
1976	U.S.S.R.	3		52.76

slalom Canadian singles†

1972	R. Eiben	E.Ger.

slalom Canadian pairs†

1972	E.Ger.	

Canoeing (women)

kayak singles (500 metres)

			min	sec
1948	K. Hoff	Denmark	2	31.90
1952	S. Saimo	Finland	2	18.40
1956	E. Dementieva	U.S.S.R.	2	18.90
1960	A. Seredina	U.S.S.R.	2	08.08
1964	L. Khvedosyuk	U.S.S.R.	2	12.87
1968	L. Pinayeva	U.S.S.R.	2	11.09
1972	Y. Ryabchinskaya	U.S.S.R.	2	03.17
1976	C. Zirzow	E.Ger.	2	01.05

kayak pairs (500 metres)

			min	sec
1960	U.S.S.R.		1	54.76
1964	Germany*		1	56.95
1968	W.Ger.		1	56.44
1972	U.S.S.R.		1	53.50
1976	U.S.S.R.		1	51.15

slalom kayak singles†

1972	A. Bahmann	E.Ger.

Canoeing events (men) no longer included

Canadian singles (10,000 metres)

			min	sec
1936	G. Hradetzky	Austria	50	01.2
1948	F. Čapek	Czech.	62	05.2
1952	F. Havens	U.S.	57	41.1
1956	L. Rottman	Romania	56	41.0

kayak singles (10,000 metres)

			min	sec
1936	E. Krebs	Ger.	46	01.6
1948	G. Fredriksson	Swed.	50	47.7
1952	T. Strömberg	Fin.	47	22.8
1956	G. Fredriksson	Swed.	47	43.4

Canadian pairs (10,000 metres)

			min	sec
1936	Czech.	50		33.5
1948	U.S.	55		55.4
1952	France	54		08.3
1956	U.S.S.R.	54		02.4

kayak fours (10,000 metres)

			min	sec
1936	Ger.	41		45.0
1948	Swed.	46		09.4
1952	Fin.	44		21.3
1956	Hung.	43		37.0

double collapsible (10,000 metres)

			min	sec
1936	Swed.		45	48.9

single kayak relay (4 × 500 metres)

			min	sec
1960	Ger.*	7		39.43

single collapsible (10,000 metres)

			min	sec
1936	E. Hradetzky	Austria	50	01.2

Cycling

1,000-metre sprint†

1900	G. Taillandier	France
1920	M. Peeters	Neth.
1924	L. Michard	France
1928	R. Beaufrand	France
1932	J. Van Egmond	Neth.
1936	T. Merkens	Ger.
1948	M. Ghella	Italy
1952	E. Sacchi	Italy
1956	M. Rousseau	France
1960	S. Gaiardoni	Italy
1964	G. Pettenella	Italy
1968	D. Morelon	France
1972	D. Morelon	France

1,000-metre time trial

			min	sec
1928	W. Falck-Hansen	Denmark	1	14.40
1932	E. Gray	Australia	1	13.00
1936	A. van Vliet	Neth.	1	12.00
1948	J. Dupont	France	1	13.50
1952	R. Mockridge	Australia	1	11.10
1956	L. Faggin	Italy	1	09.80
1960	S. Gaiardoni	Italy	1	07.27
1964	P. Sercu	Belgium	1	09.59
1968	P. Trentin	France	1	03.91
1972	N. Fredbord	Denmark	1	06.40
1976	K.-J. Grünke	E.Ger.	1	05.927

2,000-metre tandem†

1908	M. Schilles, A. Auffray	France
1920	H. Ryan, T. Lance	Gt.Brit.
1924	J. Cugnot, L. Choury	France
1928	B. Leene, D. van Dijk	Neth.
1932	M. Perrin, L. Choury	France
1936	E. Ihbe, C. Lorenz	Germany
1948	P. Perona, F. Teruzzi	Italy
1952	R. Mockridge, L. Cox	Australia
1956	J. Browne, A. Marchant	Australia
1960	S. Bianchetto, G. Beghetto	Italy
1964	A. Damiano, S. Bianchetto	Italy
1968	D. Morelon, P. Trentin	France
1972	V. Sements, I. Tseldvalnkov	U.S.S.R.

4,000-metre individual pursuit

1964	J. Daler	Czech.
1968	D. Rebillard	France
1972	K. Knudsen	Nor.
1976	G. Braun	W.Ger.

*Joint East–West German team. †Not held in 1976.

Olympic Champions, 1896–1976 (continued)

Cycling (continued)

4,000-metre team pursuit

		min	sec
1920	Italy	5	20.0
1924	Italy	5	12.0
1928	Italy	5	06.25
1932	Italy	4	52.9
1936	Fr.	4	45.0
1948	Fr.	4	57.8
1952	Italy	4	46.1
1956	Italy	4	37.4
1960	Italy	4	37.4
1964	Ger.*	4	35.7
1968	Den.	4	22.4
1972	W.Ger.	4	22.1
1976	W.Ger.	4	21.1

road race (individual)†

			hr	min	sec
1896	A. Konstantinidis	Greece	3	22	31.00
1912	R. Lewis	S.Africa	10	42	39.00
1920	H. Stenqvist	Sweden	4	41	01.80
1924	A. Blanchonnet	France	6	20	48.00
1928	H. Hansen	Den.	4	47	18.00
1932	A. Pavesi	Italy	2	28	05.60
1936	R. Charpentier	France	2	33	05.00
1948	J. Beyaert	France	5	18	12.60
1952	A. Noyelle	Belg.	5	06	03.90
1956	E. Baldini	Italy	5	21	17.00
1960	V. Kapitanov	U.S.S.R.	4	20	37.00
1964	M. Zanin	Italy	4	39	51.63
1968	P. Vianelli	Italy	4	41	25.24
1972	H. Kuiper	Neth.	4	14	38.00
1976	B. Johansson	Sweden	4	46	52.00

100-kilometre (team time trial)

		hr	min	sec
1960	Italy	2	14	33.53
1964	Neth.	2	26	31.19
1968	Neth.	2	07	49.06
1972	U.S.S.R.	2	11	17.8
1976	U.S.S.R.	2	08	53.00

Cycling events no longer included

440 yards	1904	M. Hurley	U.S.
586⅔ yards	1904	M. Hurley	U.S.
880 yards	1904	M. Hurley	U.S.
Mile	1904	M. Hurley	U.S.
2,000 metres	1904	M. Hurley	U.S.
2 miles	1904	B. Downing	U.S.
5,000 metres	1908	B. Jones	Gt.Brit.
5 miles	1904	C. Schlee	U.S.

10,000 metres	1896	P. Masson	France
20,000 metres	1908	C. Kingsbury	Gt.Brit.
25 miles	1904	B. Downing	U.S.
50,000 metres	1920	H. George	Belgium
	1924	J. Williams	Neth.
100,000 metres	1896	L. Flameng	France
	1908	C. Bartlett	Gt.Brit.
12 hours	1896	F. Schmal	Austria

time trials

333.3 metres	1896	P. Masson	France
660 yards	1908	W. Johnson	Gt.Brit.

team pursuit

1900	1,500 metres	U.S.
1908	1 mile 1 furlong	Gt.Brit.

road race (team)

		hr	min	sec
1912	Swed.	44	35	33.6
1920	France	19	16	43.2
1924	France	19	30	14.0
1928	Den.	15	09	14.0
1932	Italy	7	27	15.2
1936	France	7	39	16.2
1948	Belg.	15	58	17.4
1952	Belg.	15	20	46.6
1956	France	5	21	17.0

Equestrian sports

grand prix (dressage)

1900	C. Haegeman	Belg.
1912	C. Bonde	Sweden
1920	J. Lundblad	Sweden
1924	E. Linder	Sweden
1928	C. von Langen	Germany
1932	F. Lesage	France
1936	H. Pollay	Germany
1948	H. Moser	Switz.
1952	H. St. Cyr	Sweden
1956	H. St. Cyr	Sweden
1960	S. Filatov	U.S.S.R.
1964	H. Chammartin	Switz.
1968	I. Kizimov	U.S.S.R.
1972	L. Linsenhoff	W.Ger.
1976	C. Stueckelberger	Switz.

grand prix (dressage) team

1928	Germany
1932	France
1936	Germany
1948	France
1952	Sweden
1956	Sweden
1964	Germany*
1968	W.Germany
1972	U.S.S.R.
1976	W.Germany

grand prix (jumping)

1912	J. Cariou	France
1920	T. Lequio	Italy
1924	A. Gemuseus	Switz.
1928	F. Ventura	Czech.
1932	T. Nishi	Japan
1936	K. Hasse	Ger.
1948	H. Mariles Cortes	Mex.
1952	P. Jonquères d'Oriola	France
1956	H. Winkler	Ger.*
1960	R. d'Inzeo	Italy
1964	P. Jonquères d'Oriola	France
1968	W. Steinkraus	U.S.
1972	G. Mancinelli	Italy
1976	A. Schockemoehle	W.Ger.

grand prix (jumping) team‡

1912	Sweden
1920	Sweden
1924	Sweden
1928	Spain
1936	Germany
1948	Mexico
1952	Gt.Brit.
1956	Germany*
1960	Germany*
1964	Germany*
1968	Canada
1972	W.Ger.
1976	France

three-day event (individual)

1912	A. Nordlander	Sweden
1920	H. Mörner	Sweden
1924	A. van der Voort van Zijp	Neth.
1928	F. Pahud de Mortanges	Neth.
1932	F. Pahud de Mortanges	Neth.
1936	L. Stubbendorff	Germany
1948	B. Chevallier	France
1952	H. von Blixen-Finecke	Sweden
1956	P. Kastenman	Sweden
1960	L. Morgan	Australia
1964	M. Checcoli	Italy
1968	J. Goyon	France
1972	R. Meade	Gt.Brit.
1976	E. Coffin	U.S.

three-day event (team)

1912	Sweden
1920	Sweden
1924	Neth.
1928	Neth.
1932	U.S.
1936	Germany
1948	U.S.
1952	Sweden
1956	Gt.Brit.
1960	Australia
1964	Italy
1968	Gt.Brit.
1972	Gt.Brit.
1976	U.S.

Equestrian events no longer included

high jump	1900	G. Gardère	France
		G. Trissino	Italy
long jump	1900	Van Langendonck	Belg.
figure riding	1920	T. Bouckaert	Belg.

Fencing

foil individual (men)

1896	E. Gravelotte	France
1900	C. Coste	France
1904	R. Fonst	Cuba
1912	N. Nadi	Italy
1920	N. Nadi	Italy
1924	R. Ducret	France
1928	L. Gaudin	France
1932	G. Marzi	Italy
1936	G. Gaudini	Italy
1948	J. Buhan	France
1952	C. d'Oriola	France
1956	C. d'Oriola	France
1960	V. Zhdanovich	U.S.S.R.
1964	E. Franke	Poland
1968	I. Drimba	Romania
1972	W. Woyda	Poland
1976	F. Dal Zotto	Italy

foil team (men)

1904	Cuba
1920	Italy
1924	France
1928	Italy
1932	France
1936	Italy
1948	France
1952	France
1956	Italy
1960	U.S.S.R.
1964	U.S.S.R.
1968	France
1972	Poland
1976	W.Ger.

épée (individual) men

1900	R. Fonst	Cuba
1904	R. Fonst	Cuba
1908	G. Alibert	France
1912	P. Anspach	Belg.
1920	A. Massard	France
1924	C. Delporte	Belg.
1928	L. Gaudin	France
1932	C. Cornaggia-Medici	Italy
1936	F. Riccardi	Italy
1948	L. Cantone	Italy
1952	E. Mangiarotti	Italy
1956	C. Pavesi	Italy
1960	G. Delfino	Italy
1964	G. Kriss	U.S.S.R.
1968	G. Kulcsár	Hung.
1972	C. Fenyvési	Hung.
1976	A. Pusch	W.Ger.

épée team (men)

1908	France
1912	Belg.
1920	Italy
1924	France
1928	Italy
1932	France
1936	Italy
1948	France
1952	Italy
1956	Italy
1960	Italy
1964	Hung.
1968	Hung.
1972	Hung.
1976	Sweden

sabre individual (men)

1896	J. Georgiadis	Greece
1900	G. de la Falaise	France
1904	M. Diaz	Cuba
1908	J. Fuchs	Hung.
1912	J. Fuchs	Hung.
1920	N. Nadi	Italy
1924	S. Posta	Hung.
1928	O. Tersztyánszky	Hung.
1932	G. Piller	Hung.
1936	E. Kabos	Hung.
1948	A. Gerevich	Hung.
1952	P. Kovács	Hung.
1956	R. Kárpáti	Hung.
1960	R. Kárpáti	Hung.
1964	T. Pézsa	Hung.
1968	J. Pawlowski	Pol.
1972	V. Sidiak	U.S.S.R.
1976	V. Krovopouskov	U.S.S.R.

sabre team (men)

1908	Hung.
1912	Hung.
1920	Italy
1924	Italy
1928	Hung.
1932	Hung.
1936	Hung.
1948	Hung.
1952	Hung.
1956	Hung.
1960	Hung.
1964	U.S.S.R.
1968	U.S.S.R.
1972	Italy
1976	U.S.S.R.

foil individual (women)

1924	E. Osiier	Den.
1928	H. Mayer	Ger.
1932	E. Preis	Austria
1936	I. Schacherer-Elek	Hung.
1948	I. Elek	Hung.
1952	I. Camber	Italy
1956	G. Sheen	Gt.Brit.
1960	A. Schmid	Ger.*
1964	I. Ujlaki-Rejtö	Hung.
1968	E. Novikova	U.S.S.R.
1972	A. Ragno Lonzi	Italy
1976	I. Schwarczenberger	Hung.

foil team (women)

1960	U.S.S.R.
1964	Hung.
1968	U.S.S.R.
1972	U.S.S.R.
1976	U.S.S.R.

*Joint East–West German team. †Distances varied from 87–320 km. ‡In 1932 no team completed the course.

Olympic Champions, 1896–1976 (continued)

Fencing events no longer included

1896		
individual foil professional	L. Pyrgos	Greece
1900		
individual foil professional	L. Merignac	France
individual épée professional	A. Ayat	France
individual sabre professional	A. Conte	Italy
individual épée open	A. Ayat	France
1904		
singlestick	A. Van Zo Post	Cuba
individual foil, junior	A.G. Fox	U.S.

Gymnastics (men)

combined, or all-around (individual)

1900	S. Sandras	France
1904	J. Lenhardt	U.S.
1908	A. Braglia	Italy
1912	A. Braglia	Italy
1920	G. Zampori	Italy
1924	L. Stukelj	Yugos.
1928	G. Miez	Switz.
1932	R. Neri	Italy
1936	A. Schwarzmann	Germany
1948	V. Huhtanen	Finland
1952	V. Chukarin	U.S.S.R.
1956	V. Chukarin	U.S.S.R.
1960	B. Shakhlin	U.S.S.R.
1964	Y. Endo	Japan
1968	S. Kato	Japan
1972	S. Kato	Japan
1976	N. Andrianov	U.S.S.R.

all-around (team)

1920	Italy
1924	Italy
1928	Switz.
1932	Italy
1936	Germany
1948	Finland
1952	U.S.S.R.
1956	U.S.S.R.
1960	Japan
1964	Japan
1968	Japan
1972	Japan
1976	Japan

floor exercises

1932	I. Pelle	Hung.
1936	G. Miez	Switz.
1948	F. Pataki	Hung.
1952	K. Thoresson	Swed.
1956	V. Muratov	U.S.S.R.
1960	N. Aihara	Japan
1964	F. Menichelli	Italy
1968	S. Kato	Japan
1972	N. Andrianov	U.S.S.R.
1976	N. Andrianov	U.S.S.R.

horizontal bar

1896	H. Weingärtner	Ger.
1904	A. Heida	U.S.
	E. Henning (tied)	U.S.
1924	L. Stukelj	Yugos.
1928	G. Miez	Switz.
1932	D. Bixler	U.S.
1936	A. Saarvala	Finland
1948	J. Stalder	Switz.
1952	J. Günthard	Switz.
1956	T. Ono	Japan
1960	T. Ono	Japan
1964	B. Shakhlin	U.S.S.R.
1968	M. Voronin	U.S.S.R.
	A. Nakayama (tied)	Japan
1972	M. Tsukahara	Japan
1976	M. Tsukahara	Japan

parallel bars

1896	A. Flatow	Germany
1904	G. Eyser	U.S.
1924	A. Güttinger	Switz.
1928	L. Vácha	Czech.
1932	R. Neri	Italy
1936	K. Frey	Germany
1948	M. Reusch	Switz.
1952	H. Eugster	Switz.
1956	V. Chukarin	U.S.S.R.
1960	B. Shakhlin	U.S.S.R.
1964	Y. Endo	Japan
1968	A. Nakayama	Japan
1972	S. Kato	Japan
1976	S. Kato	Japan

side, or pommeled, horse

1896	L. Zutter	Switz.
1904	A. Heida	U.S.
1924	J. Wilhelm	Switz.
1928	A. Hanggi	Switz.
1932	I. Pelle	Hungary
1936	K. Frey	Germany
1948	P. Aaltonen	Finland
	V. Huhtanen	
	H. Savolainen	
1952	V. Chukarin	U.S.S.R.
1956	B. Shakhlin	U.S.S.R.
1960	B. Shakhlin	U.S.S.R.
	E. Ekman	Finland
1964	M. Cerar	Yugoslavia
1968	M. Cerar	Yugoslavia
1972	V. Kilmenko	U.S.S.R.
1976	Z. Magyar	Hungary

long, or vaulting, horse

1896	K. Schuhmann	Germany
1904	A. Heida	U.S.
	G. Eyser (tied)	U.S.
1924	F. Kriz	U.S.
1928	E. Mack	Switz.
1932	S. Guglielmetti	Italy
1936	K. Schnorzmann	Germany
1948	P. Aaltonen	Finland
1952	V. Chukarin	U.S.S.R.
1956	V. Muratov	U.S.S.R.
	H. Bantz	Germany*
1960	T. Ono	Japan
	B. Shakhlin	U.S.S.R.
1964	H. Yamashita	Japan
1968	M. Voronin	U.S.S.R.
1972	K. Koeste	E.Ger.
1976	N. Andrianov	U.S.S.R.

rings

1896	J. Mitropoulos	Greece
1904	H. Glass	U.S.
1924	F. Martino	Italy
1928	L. Stukecj	Yugos.
1932	G. Gulack	U.S.
1936	A. Hudec	Czech.
1948	K. Frei	Switz.
1952	G. Chaguinian	U.S.S.R.
1956	A. Azaryan	U.S.S.R.
1960	A. Azaryan	U.S.S.R.
1964	T. Hayata	Japan
1968	A. Nakayama	Japan
1972	A. Nakayama	Japan
1976	N. Andrianov	U.S.S.R.

Gymnastics (women)

combined, or all-around (individual)

1952	M. Gorokhovskaya	U.S.S.R.
1956	L. Latynina	U.S.S.R.
1960	L. Latynina	U.S.S.R.
1964	V. Čáslavská	Czech.
1968	V. Čáslavská	Czech.
1972	L. Tourischeva	U.S.S.R.
1976	N. Comaneci	Romania

combined, or all-around (team)

1928	Neth.
1936	Ger.
1948	Czech.
1952	U.S.S.R.
1956	U.S.S.R.
1960	U.S.S.R.
1964	U.S.S.R.
1968	U.S.S.R.
1972	U.S.S.R.
1976	U.S.S.R.

balance beam

1952	N. Bocharova	U.S.S.R.
1956	Á. Keleti	Hung.
1960	E. Bosáková	Czech.
1964	V. Čáslavská	Czech.
1968	N. Kuchinskaya	U.S.S.R.
1972	O. Korbut	U.S.S.R.
1976	N. Comaneci	Romania

uneven parallel bars

1952	M. Korondi	Hung.
1956	Á. Keleti	Hung.
1960	P. Astakhova	U.S.S.R.
1964	P. Astakhova	U.S.S.R.
1968	V. Čáslavská	Czech.
1972	K. Janz	E.Ger.
1976	N. Comaneci	Romania

vaulting horse

1952	Y. Kalinchuk	U.S.S.R.
1956	L. Latynina	U.S.S.R.
1960	M. Nikolayeva	U.S.S.R.
1964	V. Čáslavská	Czech.
1968	V. Čáslavská	Czech.
1972	K. Janz	E.Ger.
1976	N. Kim	U.S.S.R.

floor exercises

1952	Á. Keleti	Hung.
1956	L. Latynina	U.S.S.R.
	Á. Keleti (tied)	Hung.
1960	L. Latynina	U.S.S.R.
1964	L. Latynina	U.S.S.R.
1968	V. Čáslavská	Czech.
	L. Petrik (tied)	U.S.S.R.
1972	O. Korbut	U.S.S.R.
1976	N. Kim	U.S.S.R.

For gymnastics events no longer included, *see* page 293.

Judo†

lightweight

1964	T. Nakatani	Japan
1972	T. Kawaguchi	Japan
1976	H. Rodríguez	Cuba

light middleweight‡

1972	T. Nomura	Japan
1976	V. Nevzorov	U.S.S.R.

middleweight

1964	I. Okano	Japan
1972	S. Sekine	Japan
1976	I. Sonoda	Japan

light heavyweight

1972	S. Chochoshvili	U.S.S.R.
1976	K. Ninomiya	Japan

heavyweight

1964	I. Inokuma	Japan
1972	W. Ruska	Neth.
1976	S. Novikov	U.S.S.R.

open

1964	A. Geesink	Neth.
1972	W. Ruska	Neth.
1976	H. Uemura	Japan

*Joint East–West German team. †Not contested in 1968. ‡Called welterweight in 1972.

Olympic Champions, 1896–1976 (continued)

Modern pentathlon

individual

1912	G. Lilliehöök	Swed.
1920	G. Dyrssen	Swed.
1924	B. Lindman	Swed.
1928	S. Thofelt	Swed.
1932	J. Oxenstierna	Swed.
1936	G. Handrick	Ger.
1948	W. Grut	Swed.
1952	L. Hall	Swed.
1956	L. Hall	Swed.
1960	F. Nemeth	Hung.
1964	F. Torok	Hung.
1968	B. Ferm	Swed.
1972	A. Balczo	Hung.
1976	J. Pyciak-Peciak	Poland

team

1952	Hung.
1956	U.S.S.R.
1960	Hung.
1964	U.S.S.R.
1968	Hung.
1972	U.S.S.R.
1976	Gt.Brit.

Rowing (men)*

single sculls

			min	sec
1900	M. Barrelet	France	7	35.6
1904	F. Greer	U.S.	10	08.5
1908	H. Blackstaffe	Gt.Brit.	9	26.0
1912	W. Kinnear	Gt.Brit.	7	47.6
1920	J. Kelly	U.S.	7	35.0
1924	J. Beresford	Gt.Brit.	7	49.2
1928	H. Pearce	Australia	7	11.0
1932	H. Pearce	Australia	7	44.4
1936	G. Schäfer	Germany	8	21.5
1948	M. Wood	Australia	7	24.4
1952	Y. Chukalov	U.S.S.R.	8	12.8
1956	V. Ivanov	U.S.S.R.	8	02.5
1960	V. Ivanov	U.S.S.R.	7	13.96
1964	V. Ivanov	U.S.S.R.	8	22.51
1968	A. Wienese	Neth.	7	47.8
1972	Y. Malishev	U.S.S.R.	7	10.12
1976	P. Karppinen	Finland	7	29.03

double sculls

			min	sec
1904	U.S.		10	03.2
1920	U.S.		7	09.0
1924	U.S.		7	45.0
1928	U.S.		6	41.4
1932	U.S.		7	17.4
1936	Gt.Brit.		7	20.8
1948	Gt.Brit.		6	51.3
1952	Arg.		7	32.2
1956	U.S.S.R.		7	24.0
1960	Czech.		6	47.50
1964	U.S.S.R.		7	10.66
1968	U.S.S.R.		6	51.82
1972	U.S.S.R.		7	01.77
1976	Norway		7	13.20

pairs (without coxswain)

			min	sec
1904	U.S.		10	57.0
1908	Gt.Brit.		9	41.0
1920	Italy		7	56.0
1924	Neth.		8	19.4
1928	Germany		7	06.4
1932	Gt.Brit.		8	00.0
1936	Germany		8	16.1
1948	Gt.Brit.		7	21.1
1952	U.S.		8	20.7
1956	U.S.		7	55.4
1960	U.S.S.R.		7	02.0
1964	Canada		7	32.9
1968	E.Ger.		7	26.5
1972	E.Ger.		6	53.1
1976	E.Ger.		7	23.31

pairs (with coxswain)

			min	sec
1900	Neth.		7	34.2
1924	Switz.		8	39.0
1928	Switz.		7	42.6
1932	U.S.		8	25.8
1936	Ger.		8	36.9
1948	Den.		8	00.5
1952	France		8	28.6
1956	U.S.		8	26.1
1960	Ger.†		7	29.14
1964	U.S.		8	21.33
1968	Italy		8	04.81
1972	E.Ger.		7	17.25
1976	E.Ger.		7	58.99

fours sculls

			min	sec
1976	E.Ger.		6	18.65

fours (without coxswain)

1900	France	7	11.0
1904	U.S.	9	53.8
1908	Gt.Brit.	8	34.0
1920	Gt.Brit.	7	08.6
1928	Gt.Brit.	6	36.0
1932	Gt.Brit.	6	58.2
1936	Germany	7	01.8
1948	Italy	6	39.0
1952	Yugos.	7	16.0
1956	Canada	7	08.8
1960	U.S.	6	26.26
1964	Denmark	6	59.30
1968	E.Ger.	6	39.18
1972	E.Ger.	6	24.27
1976	E.Ger.	6	37.42

fours (with coxswain)

			min	sec
1900	Ger.	5	59.0
1912	Ger.	6	59.4
1920	Switz.	6	54.0
1924	Switz.	7	18.4
1928	Italy	6	47.8
1932	Ger.	7	19.0
1936	Ger.	7	16.2
1948	U.S.	6	50.3
1952	Czech.	7	33.4
1956	Italy	7	19.4
1960	Ger.†	6	39.12
1964	Ger.†	7	00.44
1968	N.Z.	6	45.62
1972	W.Ger.	6	31.85
1976	U.S.S.R.	6	40.22

eights

			min	sec
1900	U.S.		6	09.8
1904	U.S.		7	50.0
1908	Gt.Brit.		7	52.0
1912	Gt.Brit.		6	15.0
1920	U.S.		6	02.6
1924	U.S.		6	33.4
1928	U.S.		6	03.2
1932	U.S.		6	37.6
1936	U.S.		6	25.4
1948	U.S.		5	56.7
1952	U.S.		6	25.9
1956	U.S.		6	35.2
1960	Ger.†		5	57.18
1964	U.S.		6	18.23
1968	W.Ger.		6	07.00
1972	N.Z.		6	08.94
1976	E.Ger.		5	58.29

Rowing (women)

single sculls

			min	sec
1976	C. Scheiblich	E.Ger.	4	05.56

double sculls

1976	Bulgaria	3	44.36

pairs (without coxswain)

1976	Bulgaria	4	01.22

fours sculls

1976	E.Ger.	3	29.99

fours (with coxswain)

1976	E.Ger.	3	45.08

eights (with coxswain)

1976	E.Ger.	3	33.32

Rowing events no longer included

1904

single sculls (association seniors)	D. Duffield	U.S.
single sculls (intermediate)	F. Shepard	U.S.

1912

fours, inriggers (with coxswain)	Denmark	

Shooting

free pistol

1936	T. Ullmann	Sweden
1948	E. Vásquez Cam	Peru
1952	H. Benner	U.S.
1956	P. Linnosvuo	Finland
1960	A. Gushchin	U.S.S.R.
1964	V. Markkanen	Finland
1968	G. Kosykh	U.S.S.R.
1972	R. Skanaker	Sweden
1976	U. Potteck	E.Ger.

rapid fire pistol

1948	K. Takács	Hung.
1952	K. Takács	Hung.
1956	S. Petrescu	Romania
1960	W. McMillan	U.S.
1964	P. Linnosvoup	Finland
1968	J. Zapedzki	Poland
1972	J. Zapedzki	Poland
1976	N. Klaar	E.Ger.

free rifle‡

1908	A. Helgerud	Nor.
1912	P. Colas	France
1920	M. Fisher	U.S.
1924	M. Fisher	U.S.
1948	E. Grünig	Switz.
1952	A. Bogdanov	U.S.S.R.
1956	V. Borisov	U.S.S.R.
1960	H. Hammerer	Austria
1964	G. Anderson	U.S.
1968	G. Anderson	U.S.
1972	L. Wigger	U.S.

small-bore rifle (prone)

1900	A. Carnell	Gt.Brit.
1908	A. Carnell	Gt.Brit.
1912	F. Hird	U.S.
1920	L. Nuesslein	U.S.
1924	C. Coquelin de Lisle	France
1932	B. Rönnmark	Sweden
1936	W. Rögeberg	Norway
1948	A. Cock	U.S.
1952	I. Sarbu	Romania
1956	G. Quellette	Canada
1960	P. Kohnke	Germany†
1964	L. Hammerl	Hung.
1968	J. Kurka	Czech.
1972	Ho Jun Li	N.Korea
1976	K. Smieszek	W.Ger.

small-bore (three positions)

1952	E. Kongshaug	Norway
1956	A. Bogdanov	U.S.S.R.
1960	V. Shamburkin	U.S.S.R.
1964	L. Wigger	U.S.
1968	B. Klinger	W.Ger.
1972	J. Writer	U.S.
1976	L. Bassham	U.S.

clay pigeon (trapshooting)

1900	W. Ewing	Canada
1908	W. Ewing	Canada
1912	J. Graham	U.S.
1920	M. Arie	U.S.
1924	G. Halasy	Hung.
1952	G. Généreux	Canada
1956	G. Rossini	Italy
1960	I. Dumitrescu	Romania
1964	E. Mattarelli	Italy
1968	J. Braithwaite	Gt.Brit.
1972	A. Scalzone	Italy
1976	D. Haldeman	U.S.

skeet shooting

1968	E. Petrov	U.S.S.R.
1972	K. Wirnheir	W.Ger.
1976	J. Panacek	Czech.

moving target (running boar)

1900	L. Debray	France
1972	L. Zhelezniak	U.S.S.R.
1976	A. Gazov	U.S.S.R.

*The distances in the men's rowing events have varied from time to time. In 1904 it was 2 miles; in 1908, 1.5 miles; from 1912 to 1936, 2,000 metres; in 1908, 1 mile 350 yards; and since 1952, 2,000 metres (1 mile 427 yards). The distance in women's rowing events is 1,000 metres. †Joint East–West German team. ‡Not held in 1976.

Olympic Champions, 1896–1976 (continued)

Shooting events no longer included

1896 events

army gun (200 metres)	P. Karaseudas	Greece
army gun (300 metres)	G. Orphanidis	Greece
pistol (25 metres)	J. Phrangudis	Greece
pistol (30 metres)	S. Paine	U.S.
service revolver	J. Paine	U.S.

1900 individual events

army gun (300 metres)	A. Helgerad	U.S.
army gun (all-around 4,000 yards)	J. Millner	U.S.
full-bore rifle (300 metres standing)	L. Madsen	Den.
full-bore rifle (300 metres kneeling)	K. Staeheli	Switz.
full-bore rifle (300 metres prone)	A. Paroche	France
full-bore rifle (300 metres)	E. Kellenberger	Switz.
6-millimetre small gun (open rear sight)	C. Grosett	France
small-bore rifle (vanishing target)	W. Styles	Gt.Brit.

1900 individual events

small-bore rifle (moving target)	A. Fleming	Gt.Brit.
service revolver	M. Larrouy	France
free revolver	R. Röderer	Switz.
revolver and pistol	P. Van Asbrock	Belg.
running deer (single shot)	O. Swahn	Swed.
running deer (double shot)	W. Winans	U.S.
live pigeon	L. Bon de Lunden	Belg.
live pigeon (hunting gun)	R. de Barbarin	France

1900 team events

army gun (300 metres)	Norway
army gun (all-around)	U.S.
full-bore rifle (300 metres)	Switz.
small-bore rifle	Gt.Brit.
revolver	Switz.
revolver and pistol	U.S.
running deer	Sweden
clay pigeon	Gt.Brit.

1908 individual

army gun (1,000 yards)	J. Millner	Gt.Brit.
small-bore rifle (vanishing target)	W. Styles	Gt.Brit.
small-bore rifle (moving target)	A. Fleming	Gt.Brit.
revolver and pistol	P. Van Asbrock	Belgium
running deer (single shot)	O. Swahn	Sweden
running deer (double shot)	W. Winans	U.S.

1908 team

army gun (all-around)	U.S.
free rifle (300 metres)	Norway
small-bore rifle	Gt.Brit.
revolver and pistol	U.S.
running deer	Swed.
clay pigeon	Gt.Brit.

1912 individual

army gun (300 metres)	S. Prokopp	Hung.
army gun (100 metres)	P. Colas	France
free rifle (300 metres)	P. Colas	France
small-bore rifle	F. Hird	U.S.
small-bore rifle (vanishing target)	V. Carlberg	Sweden
revolver and pistol	A. Lane	U.S.
dueling pistol	A. Lane	U.S.
running deer (single shot)	A. Swahn	Sweden
running deer (double shot)	A. Lundeberg	Sweden

1912 team

army gun (all-around)	U.S.
free rifle (300 metres)	Sweden
small-bore (vanishing target)	Sweden
revolver and pistol	U.S.
dueling pistol	Sweden
running deer	Sweden
clay pigeon	U.S.

1920 individual

rifle (300-metre 2 position)	M. Fisher	U.S.
rifle (300-metre standing)	C. Osburn	U.S.
rifle (300-metre prone)	O. Olsen	Nor.
rifle (600-metre prone)	H. Johansson	Swed.
pistol	C. Frederick	U.S.
revolver	C. Paraense	Braz.
running deer (single shot)	O. Olsen	Nor.
running deer (double shot)	O. Lilloe-Olsen	Nor.

1920 team

rifle (300-metre 2 position)	U.S.
rifle (300-metre standing)	Den.
rifle (300-metre prone)	U.S.
rifle (600-metre prone)	U.S.
rifle (all-around)	U.S.
small-bore rifle	U.S.
pistol	U.S.
pistol and revolver	U.S.
running deer (single shot)	Nor.
running deer (double shot)	Nor.
clay pigeon	U.S.

1924 individual

pistol	H. Bailey	U.S.
running deer (single shot)	J. Boles	U.S.
running deer (double shot)	O. Lilloe-Olsen	Norway

1924 team

rifle (all-around)	U.S.
small-bore rifle	France
pistol	U.S.
running deer (single shot)	Norway
running deer (double shot)	Gt.Brit.
clay pigeon	U.S.

1932 individual

pistol or revolver	R. Morigi	Italy

1936 individual

pistol	C. van Oyen	Germany

1952 individual

running deer	J. Larsen	Norway

1956 individual

running deer	V. Romanenko	U.S.S.R.

Swimming (men)

100-metre freestyle

			min	sec
1896	A. Hajós	Hungary	1	22.2
1904	Z. Halmay	Hungary	1	02.8*
1908	C. Daniels	U.S.	1	05.6
1912	D. Kahanamoku	U.S.	1	03.4
1920	D. Kahanamoku	U.S.	1	00.4
1924	J. Weissmuller	U.S.		59.0
1928	J. Weissmuller	U.S.		58.6
1932	Y. Miyazaki	Japan		58.2
1936	F. Csik	Hungary		57.6
1948	W. Ris	U.S.		57.3
1952	C. Scholes	U.S.		57.4
1956	J. Henricks	Australia		55.4
1960	J. Devitt	Australia		55.2
1964	D. Schollander	U.S.		53.4
1968	M. Wenden	Australia		52.2
1972	M. Spitz	U.S.		51.22†
1976	J. Montgomery	U.S.		49.99

200-metre freestyle

			min	sec
1900	F. Lane	Australia	2	25.2
1904	C. Daniels	U.S.	2	44.2‡
1968	M. Wenden	Australia	1	55.2
1972	M. Spitz	U.S.	1	52.78†
1976	B. Furniss	U.S.	1	50.29

400-metre freestyle

			min	sec
1904	C. Daniels	U.S.	6	16.2§
1908	H. Taylor	Gt.Brit.	5	36.8
1912	G. Hodgson	Canada	5	24.4
1920	N. Ross	U.S.	5	26.8
1924	J. Weissmuller	U.S.	5	04.2
1928	A. Zorilla	Arg.	5	01.6
1932	C. Crabbe	U.S.	4	48.4
1936	J. Medica	U.S.	4	44.5
1948	W. Smith	U.S.	4	41.0
1952	J. Boiteux	France	4	30.7
1956	M. Rose	Australia	4	27.3
1960	M. Rose	Australia	4	18.3
1964	D. Schollander	U.S.	4	12.2
1968	M. Burton	U.S.	4	09.0
1972	B. Cooper	Australia	4	00.26†
1976	B. Goodell	U.S.	3	51.93

1,500-metre freestyle

			min	sec
1904	E. Rausch	Germany	27	18.2‖
1908	H. Taylor	Gt.Brit.	22	48.4
1912	G. Hodgson	Canada	22	00.0
1920	N. Ross	U.S.	22	23.2
1924	A. Charlton	Australia	20	06.6
1928	A. Borg	Sweden	19	51.8
1932	K. Kitamura	Japan	19	12.4
1936	N. Terada	Japan	19	13.7
1948	J. McLane	U.S.	19	18.5
1952	F. Konno	U.S.	18	30.0
1956	M. Rose	Australia	17	58.9
1960	J. Konrads	Australia	17	19.6
1964	R. Windlo	Australia	17	01.7
1968	M. Burton	U.S.	16	38.9
1972	M. Burton	U.S.	15	52.6
1976	B. Goodell	U.S.	15	02.4

100-metre butterfly

			min	sec
1968	D. Russell	U.S.		55.9
1972	M. Spitz	U.S.		54.27†
1976	M. Vogel	U.S.		54.35

200-metre butterfly

			min	sec
1956	W. Yorzyk	U.S.	2	19.3
1960	M. Troy	U.S.	2	12.8
1964	K. Berry	Australia	2	06.6
1968	C. Robie	U.S.	2	08.7
1972	M. Spitz	U.S.	2	00.70†
1976	M. Bruner	U.S.	1	59.23

100-metre backstroke

			min	sec
1904	W. Brack	Germany	1	16.8¶
1908	A. Bieberstein	Germany	1	24.6
1912	H. Hebner	U.S.	1	21.2
1920	W. Kealoha	U.S.	1	15.2
1924	W. Kealoha	U.S.	1	13.2
1928	G. Kojac	U.S.	1	08.2
1932	M. Kiyokawa	Japan	1	08.6
1936	A. Kiefer	U.S.	1	05.9
1948	A. Stack	U.S.	1	06.4
1952	Y. Oyakawa	Japan	1	05.4
1956	D. Theile	Australia	1	02.2
1960	D. Theile	Australia	1	01.9
1968	R. Matthes	E.Ger.		58.7
1972	R. Matthes	E.Ger.		56.58†
1976	J. Naber	U.S.		55.49

*100 yards. †Race first timed in hundredths of a second. ‡220 yards. §440 yards. ‖One mile. ¶100 yards.

Olympic Champions, 1896–1976 (continued)

Swimming (men) (continued)

200-metre backstroke

			min	sec
1900	E. Hoppenberg	Germany	2	47.0
1964	J. Graef	U.S.	2	10.3
1968	R. Matthes	E.Ger.	2	09.6
1972	R. Matthes	E.Ger.	2	02.82*
1976	J. Naber	U.S.	1	59.19

100-metre breaststroke

			min	sec
1968	D. McKenzie	U.S.	1	07.7
1972	N. Tagushi	Japan	1	04.94*
1976	J. Hencken	U.S.	1	03.11

200-metre breaststroke

			min	sec
1908	F. Holman	Gt.Brit.	3	09.2
1912	W. Bathe	Germany	3	01.8
1920	H. Malmroth	Sweden	3	04.4
1924	R. Skelton	U.S.	2	56.6
1928	Y. Tsuruta	Japan	2	48.8
1932	Y. Tsuruta	Japan	2	45.4
1936	T. Hamuro	Japan	2	42.5
1948	J. Verdeur	U.S.	2	39.3
1952	J. Davies	Australia	2	34.4
1956	M. Furukawa	Japan	2	34.7
1960	W. Mulliken	U.S.	2	37.4
1964	I. O'Brien	Australia	2	27.8
1968	F. Muñoz	Mexico	2	28.7
1972	J. Hencken	U.S.	2	21.55*
1976	D. Wilkie	Gt.Brit.	2	15.11

200-metre medley†

			min	sec
1968	C. Hickcox	U.S.	2	12.0
1972	G. Larsson	Swed.	2	07.17*

400-metre medley

			min	sec
1964	R. Roth	U.S.	4	45.4
1968	C. Hickcox	U.S.	4	48.4
1972	G. Larsson	Swed.	4	31.98*
1976	R. Strachan	U.S.	4	23.68

400-metre freestyle relay†

		min	sec
1964	U.S.	3	33.2
1968	U.S.	3	31.7
1972	U.S.	3	26.42*

400-metre medley relay

		min	sec
1960	U.S.	4	05.4
1964	U.S.	3	58.4
1968	U.S.	3	54.9
1972	U.S.	3	48.16*
1976	U.S.	3	42.22

Swimming (women)

800-metre freestyle relay

		min	sec
1908	Gt.Brit.	10	55.6
1912	Australia	10	11.2
1920	U.S.	10	04.4
1924	U.S.	9	53.4
1928	U.S.	9	36.2
1932	Japan	8	58.4
1936	Japan	8	51.5
1948	U.S.	8	46.0
1952	U.S.	8	31.1
1956	Australia	8	23.6
1960	U.S.	8	10.2
1964	U.S.	7	52.1
1968	U.S.	7	52.3
1972	U.S.	7	35.8
1976	U.S.	7	23.2

100-metre freestyle

			min	sec
1912	F. Durack	Australia	1	22.2
1920	E. Bleibtrey	U.S.	1	13.6
1924	E. Lackie	U.S.	1	12.4
1928	A. Osipowich	U.S.	1	15.0
1932	H. Madison	U.S.	1	06.8
1936	H. Mastenbroek	Neth.	1	05.9
1948	G. Andersen	Denmark	1	06.3
1952	K. Szöke	Hungary	1	06.8
1956	D. Fraser	Australia	1	02.0
1960	D. Fraser	Australia	1	01.2
1964	D. Fraser	Australia		59.5
1968	J. Henne	U.S.	1	00.0
1972	S. Neilson	U.S.		58.59*
1976	K. Ender	E.Ger.		55.65

200-metre freestyle

			min	sec
1968	D. Meyer	U.S.	2	10.5
1972	S. Gould	Australia	2	03.56*
1976	K. Ender	E.Ger.	1	59.26

400-metre freestyle

			min	sec
1924	M. Norelius	U.S.	6	02.2
1928	M. Norelius	U.S.	5	42.8
1932	H. Madison	U.S.	5	28.5
1936	H. Mastenbroek	Neth.	5	26.4
1948	A. Curtis	U.S.	5	17.8
1952	V. Gyenge	Hungary	5	12.1
1956	L. Crapp	Australia	4	54.6
1960	C. von Saltza	U.S.	4	50.6
1964	V. Duenkel	U.S.	4	43.3
1968	D. Meyer	U.S.	4	31.8
1972	S. Gould	Australia	4	19.04*
1976	P. Thümer	E.Ger.	4	09.89

800-metre freestyle

			min	sec
1968	D. Meyer	U.S.	9	24.0
1972	K. Rothhammer	U.S.	8	53.7
1976	P. Thümer	E.Ger.	8	37.1

100-metre butterfly

			min	sec
1956	S. Mann	U.S.	1	11.0
1960	C. Schuler	U.S.	1	09.5
1964	S. Stouder	U.S.	1	04.7
1968	L. McClements	Australia	1	05.5
1972	M. Aoki	Japan	1	03.34*
1976	K. Ender	E.Ger.	1	00.13

200-metre butterfly

			min	sec
1968	A. Kok	Neth.	2	24.7
1972	K. Moe	U.S.	2	15.57*
1976	A. Pollack	E.Ger.	2	11.41

100-metre backstroke

			min	sec
1924	S. Bauer	U.S.	1	23.2
1928	M. Braun	Neth.	1	22.0
1932	E. Holm	U.S.	1	19.4
1936	D. Senff	Neth.	1	18.9
1948	K. Harup	Den.	1	14.4
1952	J. Harrison	S.Af.	1	14.3
1956	J. Grinham	Gt.Brit.	1	12.9
1960	L. Burke	U.S.	1	09.3
1964	C. Ferguson	U.S.	1	07.7
1968	K. Hall	U.S.	1	06.2
1972	M. Belote	U.S.	1	05.78*
1976	U. Richter	E.Ger.	1	01.83

200-metre backstroke

			min	sec
1968	L. Watson	U.S.	2	24.8
1972	M. Belote	U.S.	2	19.19*
1976	U. Richter	E.Ger.	2	13.43

100-metre breaststroke

			min	sec
1968	B. Bjedov	Yugos.	1	15.8
1972	C. Carr	U.S.	1	13.58*
1976	H. Anke	E.Ger.	1	11.16

200-metre breaststroke

			min	sec
1924	L. Morton	Gt.Brit.	3	33.2
1928	H. Schrader	Germany	3	12.6
1932	C. Dennis	Australia	3	06.3
1936	H. Maehata	Japan	3	03.6
1948	H. van Vliet	Neth.	2	57.2
1952	E. Székecy	Hungary	2	51.7
1956	U. Happe	Germany‡	2	53.1
1960	A. Lonsbrough	Gt.Brit.	2	49.5
1964	G. Prozumenshchikova	U.S.S.R.	2	46.4
1968	S. Wichman	U.S.	2	44.4
1972	B. Whitfield	Australia	2	41.71*
1976	M. Koshevaia	U.S.S.R.	2	33.35

200-metre medley†

			min	sec
1968	C. Kolb	U.S.	2	24.7
1972	S. Gould	Australia	2	23.07*

400-metre medley

			min	sec
1964	D. De Varona	U.S.	5	18.7
1968	C. Kolb	U.S.	5	08.5
1972	G. Neall	Australia	5	02.97*
1976	U. Tauber	E.Ger.	4	42.77

400-metre freestyle relay

		min	sec
1912	Gt.Brit.	5	52.8
1920	U.S.	5	11.6
1924	U.S.	4	58.8
1928	U.S.	4	47.6
1932	U.S.	4	38.0
1936	Neth.	4	36.0
1948	U.S.	4	29.2
1952	Hung.	4	24.4
1956	Australia	4	17.1
1960	U.S.	4	08.9
1964	U.S.	4	03.8
1968	U.S.	4	02.5
1972	U.S.	3	55.19*
1976	U.S.	3	44.82

400-metre medley relay

		min	sec
1960	U.S.	4	41.1
1964	U.S.	4	33.9
1968	U.S.	4	28.3
1972	U.S.	4	20.75*
1976	E.Ger.	4	07.95

Diving (men)

springboard

1908	A. Zürner	Ger.
1912	P. Günther	Ger.
1920	L. Kuehn	U.S.
1924	A. White	U.S.
1928	P. Desjardins	U.S.
1932	M. Galitzen	U.S.
1936	R. Degener	U.S.
1948	B. Harlan	U.S.
1952	D. Browning	U.S.
1956	R. Clotworthy	U.S.
1960	G. Tobian	U.S.
1964	K. Sitzberger	U.S.
1968	R. Wrightson	U.S.
1972	V. Vasin	U.S.S.R.
1976	P. Boggs	U.S.

platform (high) diving

1908	H. Johansson	Swed.
1912	E. Adlerz	Swed.
1920	C. Pinkston	U.S.
1924	A. White	U.S.
1928	P. Desjardins	U.S.
1932	H. Smith	U.S.
1936	M. Wayne	U.S.
1948	S. Lee	U.S.
1952	S. Lee	U.S.
1956	J. Capilla	Mex.
1960	R. Webster	U.S.
1964	R. Webster	U.S.
1968	K. DiBiasi	Italy
1972	K. DiBiasi	Italy
1976	K. DiBiasi	Italy

Diving (women)

springboard

1920	A. Riggin	U.S.
1924	E. Becker	U.S.
1928	H. Meany	U.S.
1932	G. Coleman	U.S.
1936	M. Gestring	U.S.
1948	V. Draves	U.S.
1952	P. McCormick	U.S.
1956	P. McCormick	U.S.
1960	I. Kramer	Ger.‡
1964	I. Engel-Kramer	Ger.‡
1968	S. Gossick	U.S.
1972	M. King	U.S.
1976	J. Chandler	U.S.

platform (high) diving

1912	G. Johansson	Swed.
1920	S. Fryland Clausen	Den.
1924	C. Smith	U.S.
1928	E. Becker Pinkston	U.S.
1932	D. Poynton	U.S.
1936	D. Poynton-Hill	U.S.
1948	V. Draves	U.S.
1952	P. McCormick	U.S.
1956	P. McCormick	U.S.
1960	I. Kramer	Ger.‡
1964	L. Bush	U.S.
1968	M. Duchkova	Czech.
1972	U. Knape	Swed.
1976	E. Vaytsekhovskaia	U.S.S.R.

*Race first timed in hundredths of a second. †Not held in 1976. ‡Joint East–West German team.

Olympic Champions, 1896–1976 (continued)

Swimming and diving events no longer included

1896

		min	sec
100 metres for sailors			
J. Malokinis	Greece	2	20.4
400-metre freestyle			
P. Neumann	Austria	8	12.6
1,000-metre freestyle			
A. Hajós	Hung.	18	22.2

1900

		min	sec
1,000-metre freestyle			
J. Jarvis	Gt.Brit.	13	40.2
4,000-metre freestyle			
J. Jarvis	Gt.Brit.	58	24.0
200-metre obstacle			
F. Lane	Australia	2	38.4
60-metre underwater			
W. de Vaudeville	Germany	1	53.4

1904

		min	sec
880-yard freestyle			
E. Rausch	Ger.	13	11.4
1-mile freestyle			
E. Rausch	Ger.	27	18.2
440-yard breaststroke			
G. Zacharias	Ger.	7	23.6
200-yard relay			
U.S.		2	04.6
100-yard backstroke			
W. Brack	Ger.	1	16.8
combined springboard and high diving			
G. Sheldon	U.S.		
plunge for distance			
W. Dickey	U.S.	62 ft	6 in.

1912

		min	sec
400-metre breaststroke			
W. Bathe	Ger.	6	29.6
diving from 5- and 10-metre boards			
E. Adlerz	Sweden		

1920

		min	sec
400-metre breaststroke			
H. Malmroth	Sweden	6	31.8
300-metre freestyle (women)			
E. Bleibtrey	U.S.	4	34.0
diving from 5- and 10-metre boards			
A. Wallman	Sweden		

1924

artistic diving	
R. Eve	Australia

Team Sports

*football (association; soccer)**

1900	Gt.Brit.
1904	Canada
1908	Gt.Brit.
1912	Gt.Brit.
1920	Belgium
1924	Uruguay
1928	Uruguay
1936	Italy
1948	Sweden
1952	Hung.
1956	U.S.S.R.
1960	Yugos.
1964	Hung.
1968	Hung.
1972	Poland
1976	E.Ger.

basketball (men)

1936	U.S.
1948	U.S.
1952	U.S.
1956	U.S.
1960	U.S.
1964	U.S.
1968	U.S.
1972	U.S.S.R.
1976	U.S.

basketball (women)

1976	U.S.S.R.

hockey (field)

1908	Gt.Brit.
1920	Gt.Brit.
1928	India
1932	India
1936	India
1948	India
1952	India
1956	India
1960	Pak.
1964	India
1968	Pak.
1972	W.Ger.
1976	New Zealand

handball (team) (men)

1936	Germany†
1952	Sweden‡
1972	Yugos.§
1976	U.S.S.R.§

handball (team) (women)

1976	U.S.S.R.§

volleyball (men)

1964	U.S.S.R.
1968	U.S.S.R.
1972	Japan
1976	Poland

volleyball (women)

1964	Japan
1968	U.S.S.R.
1972	U.S.S.R.
1976	Japan

water polo

1900	Gt.Brit.
1904	U.S.
1908	Gt.Brit.
1912	Gt.Brit.
1920	Gt.Brit.
1924	France
1928	Germany
1932	Hung.
1936	Hung.
1948	Italy
1952	Hung.
1956	Hung.
1960	Italy
1964	Hung.
1968	Yugos.
1972	U.S.S.R.
1976	Hung.

Weight lifting‖

bantamweight

			kg
1948	J. de Pietro	U.S.	307.5
1952	I. Udodov	U.S.S.R.	315.0
1956	C. Vinci	U.S.	342.5
1960	C. Vinci	U.S.	345.0
1964	A. Vakhonin	U.S.S.R.	357.5
1968	M. Nassiri	Iran	367.5
1972	I. Foeldi	Hung.	377.5
1976	N. Nurikian	Bulgaria	262.5

flyweight

			kg
1972	Z. Smalcerz	Poland	337.5
1976	A. Varonin	U.S.S.R.	242.5

featherweight

			kg
1920	F. de Haes	Belg.	220.0
1924	P. Gabetti	Italy	402.5¶
1928	F. Andrysek	Austria	287.5
1932	R. Suvigny	France	287.5
1936	A. Terlazzo	U.S.	312.5
1948	M. Fayad	Egypt	332.5
1952	R. Chimishkyan	U.S.S.R.	337.5
1956	I. Berger	U.S.	352.5
1960	Ye. Minayev	U.S.S.R.	372.5
1964	Y. Miyake	Japan	397.5
1968	Y. Miyake	Japan	392.5
1972	N. Nurikian	Bulgaria	402.5
1976	N. Kolesnikov	U.S.S.R.	285.0

lightweight

			kg
1920	A. Neyland	Estonia	257.5
1924	E. Décottignies	France	440.0¶
1928	K. Helbig	Ger.	322.5
	H. Haas	Austria	
1932	R. Duverger	France	325.0
1936	M. Mesbah	Egypt	342.5
	R. Fein	Austria	
1948	I. Shams	Egypt	360.0
1952	T. Kono	U.S.	362.5
1956	I. Rybak	U.S.S.R.	380.0
1960	V. Bushuyev	U.S.S.R.	397.5
1964	W. Baszanowski	Poland	432.5
1968	W. Baszanowski	Poland	437.5
1972	M. Kirzhinov	U.S.S.R.	460.0
1976	Disqualified		

middleweight

			kg
1920	B. Gance	France	245.0
1924	C. Galimberti	Italy	492.5¶
1928	F. Roger	France	335.0
1932	R. Ismayr	Germany	345.0
1936	K. el Thouni	Egypt	387.5
1948	F. Spellman	U.S.	390.0
1952	P. George	U.S.	400.0
1956	F. Bogdanovsky	U.S.S.R.	420.0
1960	A. Kurynov	U.S.S.R.	437.5
1964	H. Zdrazila	Czech.	445.0
1968	V. Kurentsov	U.S.S.R.	475.0
1972	Y. Bikov	Bulgaria	485.0
1976	Y. Mitkov	Bulgaria	335.0

light heavyweight

			kg
1920	E. Cadine	France	290.0
1924	C. Rigoulot	France	502.5¶
1928	S. Nosseir	Egypt	355.0
1932	L. Hostin	France	365.0
1936	L. Hostin	France	372.5
1948	S. Stanczyk	U.S.	417.5
1952	T. Lomakin	U.S.S.R.	417.5
1956	T. Kono	U.S.	447.5
1960	I. Palinski	Poland	442.5
1964	R. Plyukfelder	U.S.S.R.	475.0
1968	B. Selitsky	U.S.S.R.	485.0
	V. Belayev	U.S.S.R.	
1972	L. Jenssen	Norway	507.5
1976	V. Shary	U.S.S.R.	365.0

*The competitions in 1900 and 1904 are said to be unofficial. †Outdoors (field handball). ‡Demonstration. §Indoors.
‖In 1976, the press lift was removed, weights given being the total for the clean and press and the snatch. ¶Total of five lifts.

Olympic Champions, 1896–1976 (continued)

Weight lifting (continued)

middle heavyweight

			kg
1952	N. Schemansky	U.S.	445.0
1956	A. Vorobev	U.S.S.R.	462.5
1960	A. Vorobev	U.S.S.R.	472.5
1964	V. Golovanov	U.S.S.R.	487.5
1968	I. Kangasniemi	Finland	517.5
1972	A. Nikolov	Bulg.	525.0
1976	D. Rigert	U.S.S.R.	382.5

heavyweight

			kg
1896	V. Jensen	Den.	111.5
1904	P. Kakousis	Greece	111.58
1920	F. Bottino	Italy	270.0
1924	G. Tonani	Italy	517.5*
1928	J. Strassberger	Ger.	372.5
1932	J. Skobia	Czech.	380.0
1936	J. Manger	Ger.	410.0
1948	J. Davis	U.S.	452.5
1952	J. Davis	U.S.	460.0
1956	P. Anderson	U.S.	500.0
1960	Y. Vlasov	U.S.S.R.	537.5
1964	L. Zhabotinsky	U.S.S.R.	572.5
1968	L. Zhabotinsky	U.S.S.R.	572.5
1972	Y. Talts	U.S.S.R.	580.0
1976	Disqualified		

super-heavyweight

			kg
1972	V. Alekseyev	U.S.S.R.	640.0
1976	V. Alekseyev	U.S.S.R.	440.0

Weight lifting events no longer included

1896			kg
weight lifting, one hand	L. Elliot	Gt.Brit.	71.0
1904			
dumbell competition	O. Osthoff	U.S.	48 points

Wrestling†

freestyle light flyweight

1972	R. Dmitriev	U.S.S.R.
1976	K. Issaev	Bulgaria

flyweight

1904	R. Curry	U.S.
1948	L. Viitala	Finland
1952	H. Gemici	Turkey
1956	M. Zalkalamanidze	U.S.S.R.
1960	A. Bilek	Turkey
1964	Y. Yoshida	Japan
1968	S. Nakata	Japan
1972	K. Kato	Japan
1976	Y. Takada	Japan

bantamweight

1900	G. Mehnert	U.S.
1908	G. Mehnert	U.S.
1924	K. Pihlajamäki	Fin.
1928	K. Maakinen	Fin.
1932	R. Pearce	U.S.
1936	O. Zombori	Hung.
1948	N. Akar	Turkey
1952	I. Ishii	Japan
1956	M. Dagistanli	Turkey
1960	T. McCann	U.S.
1964	Y. Uetake	Japan
1968	Y. Uetake	Japan
1972	H. Yanagida	Japan
1976	V. Umin	U.S.S.R.

featherweight

1904	I. Niflot	U.S.
1908	G. Dole	U.S.
1920	C. Ackerly	U.S.
1924	R. Reed	U.S.
1928	A. Morrison	U.S.
1932	H. Pihlajamäki	Fin.
1936	K. Pihlajamäki	Fin.
1948	G. Bilge	Turkey
1952	B. Sit	Turkey
1956	S. Sasahara	Japan
1960	M. Dagistanli	Turkey
1964	O. Watanabe	Japan
1968	M. Kaneko	Japan
1972	Z. Abdulbekov	U.S.S.R.
1976	J.-M. Yang	S.Kor.

lightweight

1904	B. Bradshaw	U.S.
1908	G. Relwyskow	Gt.Brit.
1920	R. Anttila	Finland
1924	R. Vis	U.S.
1928	O. Käpp	Estonia
1932	C. Pacome	France
1936	K. Kárpáti	Hung.
1948	C. Atik	Turkey
1952	O. Anderberg	Sweden
1956	E. Habibi	Iran
1960	S. Wilson	U.S.
1964	E. Dimov	Bulg.
1968	A. Movahed	Iran
1972	D. Gable	U.S.
1976	P. Pinigin	U.S.S.R.

welterweight

1904	O. Roem	U.S.
1924	H. Gehri	Switz.
1928	A. Haavisto	Fin.
1932	J. Van Bebber	U.S.
1936	F. Lewis	U.S.
1948	J. Dogu	Turkey
1952	W. Smith	U.S.
1956	M. Ikeda	Japan
1960	D. Blubaugh	U.S.
1964	I. Ogan	Turkey
1968	M. Atalay	Turkey
1972	W. Wells	U.S.
1976	J. Date	Japan

middleweight

1904	C. Erickson	U.S.
1908	S. Bacon	Gt.Brit.
1920	E. Leino	Finland
1924	F. Haggmann	Switz.
1928	E. Kyburz	Switz.
1932	I. Johansson	Swed.
1936	E. Poilvé	France
1948	G. Brand	U.S.
1952	D. Cimakuridze	U.S.S.R.
1956	N. Stanchev	Bulg.
1960	H. Güngor	Turkey
1964	P. Gardchev	Bulg.
1968	B. Gurevich	U.S.S.R.
1972	L. Tediashvili	U.S.S.R.
1976	J. Peterson	U.S.

light heavyweight

1920	A. Larsson	Sweden
1924	J. Spellman	U.S.
1928	T. Sjöstedt	Sweden
1932	P. Mehringer	U.S.
1936	K. Fridell	Sweden
1948	H. Wittenberg	U.S.
1952	W. Palm	Sweden
1956	G.-R. Takhti	Iran
1960	I. Atli	Turkey
1964	A. Medved	U.S.S.R.
1968	A. Ayuk	Turkey
1972	B. Peterson	U.S.
1976	L. Tediashvili	U.S.S.R.

heavyweight

1896	K. Schumann	Germany
1904	B. Hansen	U.S.
1908	G. O'Kelly	Gt.Brit.
1920	R. Rothe	Switz.
1924	H. Steele	U.S.
1928	J. Richthoff	Sweden
1932	J. Richthoff	Sweden
1936	K. Palusalu	Estonia
1948	G. Bóbis	Hung.
1952	A. Mekokishvili	U.S.S.R.
1956	H. Kaplan	Turkey
1960	W. Dietrich	Germany‡
1964	A. Ivanitsky	U.S.S.R.
1968	A. Medved	U.S.S.R.
1972	I. Yarygin	U.S.S.R.
1976	I. Yarygin	U.S.S.R.

super-heavyweight

1972	A. Medved	U.S.S.R.
1976	A. Andiev	U.S.S.R.

Greco-Roman light flyweight

1972	G. Berceanu	Romania
1976	A. Shumakov	U.S.S.R.

flyweight

1948	P. Lombardi	Italy
1952	B. Gurevich	U.S.S.R.
1956	N. Solovyev	U.S.S.R.
1960	D. Pirvulescu	Romania
1964	T. Hanahara	Japan
1968	P. Kirov	Bulg.
1972	P. Kirov	Bulg.
1976	V. Konstantinov	U.S.S.R.

bantamweight

1924	E. Pütsep	Estonia
1928	C. Leucht	Germany
1932	J. Brendel	Germany
1936	M. Lórincz	Hung.
1948	K. Pettersen	Sweden
1952	I. Hódos	Hung.
1956	K. Vyrupayev	U.S.S.R.
1960	O. Karavayev	U.S.S.R.
1964	M. Ichiguchi	Japan
1968	J. Varga	Hung.
1972	R. Kazakov	U.S.S.R.
1976	P. Ukkola	Finland

featherweight

1912	K. Koskelo	Finland
1920	O. Friman	Finland
1924	K. Anttila	Finland
1928	V. Väli	Estonia
1932	G. Gozzi	Italy
1936	Y. Erkan	Turkey
1948	M. Oktav	Turkey
1952	Y. Punkin	U.S.S.R.
1956	R. Mäkinen	Finland
1960	M. Sille	Turkey
1964	I. Polyak	Hung.
1968	R. Rurua	U.S.S.R.
1972	G. Markov	Bulg.
1976	K. Lipien	Poland

lightweight

1908	E. Porro	Italy
1912	E. Väre	Finland
1920	E. Väre	Finland
1924	O. Friman	Finland
1928	L. Keresztes	Hungary
1932	E. Malmberg	Sweden
1936	L. Koskela	Finland
1948	G. Freij	Sweden
1952	C. Safin	U.S.S.R.
1956	K. Lehtonen	Finland
1960	A. Koridze	U.S.S.R.
1964	K. Ayvaz	Turkey
1968	M. Mumemura	Japan
1972	S. Khisamutdinov	U.S.S.R.
1976	S. Nalbandyan	U.S.S.R.

welterweight

1924	E. Verterlund	Finland
1928	V. Kokkinen	Finland
1932	I. Johansson	Sweden
1936	R. Svedberg	Sweden
1948	G. Andersson	Sweden
1952	M. Szilvási	Hung.
1956	M. Bayrak	Turkey
1960	M. Bayrak	Turkey
1964	A. Kolesov	U.S.S.R.
1968	R. Vesper	E.Ger.
1972	V. Macha	Czech.
1976	A. Bykov	U.S.S.R.

middleweight

1908	F. Martenson	Sweden
1912	C. Johansson	Sweden
1920	C. Westergren	Sweden
1928	I. Moustafa	Egypt
1932	V. Kokkinen	Finland
1936	I. Johansson	Sweden
1948	A. Grönberg	Sweden
1952	A. Grönberg	Sweden
1956	G. Kartoziya	U.S.S.R.
1960	D. Dobrev	Bulg.
1964	B. Simic	Yugos.
1968	L. Metz	E.Ger.
1972	C. Hegedus	Hung.
1976	M. Petkovic	Yugos.

light heavyweight

1908	V. Weckman	Finland
1912	A. Ahlgren	Sweden
	I. Bohling	Finland
1920	C. Johansson	Sweden
1924	C. Westergren	Sweden
1932	R. Svensson	Sweden
1936	A. Cadier	Sweden
1948	K. Nilsson	Sweden
1952	K. Gröndahl	Finland
1956	V. Nikolayev	U.S.S.R.
1960	T. Kis	Turkey
1964	B. Radev	Bulg.
1968	B. Radev	Bulg.
1972	V. Rezantsev	U.S.S.R.
1976	V. Rezantsev	U.S.S.R.

heavweight

1908	R. Weisz	Hung.
1912	Y. Saarela	Finland
1920	A. Lindfors	Finland
1924	J. Deglane	France
1928	R. Svensson	Sweden
1932	C. Westergren	Sweden
1936	K. Palusalu	Estonia
1948	A. Kirecci	Turkey
1952	J. Kotkas	U.S.S.R.
1956	A. Parfenov	U.S.S.R.
1960	I. Bogdan	U.S.S.R.
1964	I. Kozma	Hung.
1968	I. Kozma	Hung.
1972	N. Martinescu	Rom.
1976	N. Bolboshin	U.S.S.R.

super-heavyweight

1972	A. Roshin	U.S.S.R.
1976	A. Kolchinski	U.S.S.R.

*Total of five lifts. †The weights have varied from time to time, and classifications differ in various reports. ‡Joint East–West German team.

Olympic Champions, 1896–1976 (continued)

Yachting

*Dragon**		*Star**		*Flying Dutchman*		*Finn monotype*			*Soling*	
1948	Nor.	1932	U.S.	1960	Norway	1952	P. Elvström	Den.	1972	U.S.
1952	Nor.	1936	Ger.	1964	N.Z.	1956	P. Elvström	Den.	1976	Den.
1956	Sweden	1948	U.S.	1968	Gt.Brit.	1960	P. Elvström	Den.		
1960	Greece	1952	Italy	1972	Gt.Brit.	1964	W. Kuhweide	Germany†		
1964	Den.	1956	U.S.	1976	W.Ger.	1968	V. Mankin	U.S.S.R.	*Tornado*	
1968	U.S.	1960	U.S.S.R.			1972	S. Maury	France	1976	U.K.
1972	Australia	1964	Bahamas	*Tempest*		1976	J. Shumann	E.Ger.		
		1968	U.S.	1972	U.S.S.R.				*470*	
		1972	Australia	1976	Sweden				1976	W.Ger.

Events no longer included

1900		1908		1912		1920		1920	
over-10-metre class	France	over-10-metre class (12-metre boat)	Gt.Brit.	over-10-metre class (12-metre boat)	Norway	40-metre class	Sweden	8-metre (new)	Norway
10-metre class	Germany	8-metre class	Gt.Brit.	10-metre class	Sweden	30-metre class	Sweden	7-metre (old)	Gt.Brit.
8-metre class (3-ton boat)	Gt.Brit.	7-metre class	Gt.Brit.	8-metre class	Norway	12-metre class (old)	Norway	6.5-metre class (new)	Neth.
6-metre class (2-ton boat)	Switz.	6-metre class	Gt.Brit.	6-metre class	France	12-metre class (new)	Norway	6-metre class (old)	Belg.
						10-metre class (old)	Norway	6-metre class (new)	Norway
						10-metre class (new)	Norway	12-foot centreboard boat	Neth.
						8-metre (old)	Norway	18-foot centreboard boat	Gt.Brit.

1924			1936			1956	
8-metre class (new)		Norway	8-metre class (new)		Italy	sharpie class	N.Z.
6-metre class (new)		Norway	6-metre class (new)		Gt.Brit.	5.5-metre class	Swed.
12-foot centreboard boat	L. Huybrechts	Belgium	monotype class	"Nürnberg"	Neth.		
						1960	
1928			1948			5.5-metre class	U.S.
8-metre class (new)		France	6-metre class		U.S.		
6-metre class (new)		Norway	Swallow class		Gt.Brit.	1964	
12-foot dinghy	S. Thørell	Sweden	Firefly class	P. Elvström	Den.	5.5-metre class	Austr.
1932			1952			1968	
8-metre class (new)		U.S.	6-metre class		U.S.	5.5-metre class	Swed.
6-metre class (new)		Sweden	5.5-metre class		U.S.		
monotype class	J. Lebrun	France					

Winter sports (men)

skiing

30-kilometre cross-country			hr	min	sec
1956	V. Hakulinen	Fin.	1	44	06.0
1960	S. Jernberg	Swed.	1	51	03.9
1964	E. Mäntyranta	Fin.	1	30	50.7
1968	F. Nones	Italy	1	35	39.2
1972	V. Vedenin	U.S.S.R.	1	36	31.2
1976	S. Saveliev	U.S.S.R.	1	30	29.4

15-kilometre cross-country			hr	min	sec
1924‡	T. Haug	Nor.	1	14	31.0
1928‡	J. Gröttumsbraaten	Nor.	1	37	01.0
1932‡	S. Utterström	Swed.	1	23	07.0
1936‡	E.-A. Larsson	Swed.	1	14	38.0
1948‡	M. Lundstrom	Swed.	1	13	50.0
1952‡	H. Brenden	Nor.	1	01	34.0
1956	H. Brenden	Nor.		49	39.0
1960	H. Brusveen	Nor.		51	55.5
1964	E. Mäntyranta	Fin.		50	54.1
1968	H. Groenningen	Nor.		47	54.2
1972	S.-A. Lundback	Swed.		45	28.2
1976	N. Bajukov	U.S.S.R.		43	58.5

50-kilometre cross-country			hr	min	sec
1924	T. Haug	Nor.	3	44	32.0
1928	P. Hedlund	Swed.	4	52	03.3
1932	V. Saarinen	Fin.	4	28	00.0
1936	E. Viklund	Swed.	3	30	11.0
1948	N. Karlsson	Swed.	3	47	48.0
1952	V. Hakulinen	Fin.	3	33	33.0
1956	S. Jernberg	Swed.	2	50	27.0
1960	K. Hamalainen	Fin.	2	59	06.3
1964	S. Jernberg	Swed.	2	43	52.6
1968	O. Ellefsaeter	Nor.	2	28	45.8
1972	P. Tyldum	Nor.	2	43	14.8
1976	I. Formo	Nor.	2	37	30.1

relay races (4 × 10 kilometres)			hr	min	sec
1936	Finland		2	41	33.0
1948	Sweden		2	32	08.0
1952	Finland		2	20	16.0
1956	U.S.S.R.		2	15	30.0
1960	Finland		2	18	45.6
1964	Sweden		2	18	34.6
1968	Norway		2	08	33.5
1972	U.S.S.R.		2	04	47.9
1976	Finland		2	07	59.7

ski jumping§		
1924	J. Tullin Thams	Norway
1928	A. Andersen	Norway
1932	B. Ruud	Norway
1936	B. Ruud	Norway
1948	P. Hugsted	Norway
1952	A. Bergmann	Norway
1956	A. Hyvärinen	Finland
1960	H. Recknagel	Germany†
1964 (80)	T. Engan	Norway
1964 (70)	V. Kankkonen	Finland
1968 (90)	V. Beloussov	U.S.S.R.
1968 (70)	J. Raška	Czech.
1972 (90)	W. Fortuna	Poland
1972 (70)	Y. Kasaya	Japan
1976 (90)	K. Schnabl	Austria
1976 (70)	H.-G. Aschenbach	E.Ger.

Nordic combined 15 kilometres and jumping		
1924	T. Haug	Norway
1928	J. Gröttumsbraaten	Norway
1932	J. Gröttumsbraaten	Norway
1936	O. Hagen	Norway
1948	H. Hasu	Norway
1952	S. Slättvik	Norway
1956	S. Stenersen	Norway
1960	G. Thoma	Germany†
1964	T. Knutsen	Norway
1968	F. Keller	W.Ger.
1972	U. Wehling	E.Ger.
1976	U. Wehling	E.Ger.

slalom			min	sec
1948	E. Reinalter	Switz.	2	10.3
1952	O. Schneider	Austria	2	00.0
1956	A. Sailer	Austria	3	14.7
1960	E. Hinterseer	Austria	2	08.9
1964	J. Stiegler	Austria	2	21.13
1968	J. Killy	France	1	39.73
1972	F. Ochoa	Spain	1	49.27
1976	P. Gros	Italy	2	03.29

giant slalom			min	sec
1952	S. Eriksen	Norway	2	25.0
1956	A. Sailer	Austria	3	00.1
1960	R. Staub	Switz.	1	48.3
1964	F. Boulieu	France	1	46.71
1968	J. Killy	France	3	29.28
1972	G. Thoeni	Italy	3	09.62
1976	H. Hemmi	Switz.	3	26.97

downhill			min	sec
1948	H. Oreiller	France	2	55.0
1952	Z. Colo	Italy	2	30.8
1956	A. Sailer	Austria	2	52.2
1960	J. Vuarnet	France	2	06.0
1964	E. Zimmermann	Austria	2	18.16
1968	J. Killy	France	1	59.85
1972	B. Russi	Switz.	1	51.43
1976	F. Klammer	Austria	1	45.73

alpine combined		
1936	F. Pfnür	Germany
1948	H. Oreiller	France
1972	G. Thoeni	Italy
1976	G. Thoeni	Italy

*Not held in 1976. †Joint East–West German team. ‡1924–52, 18 kilometres. §From 1924–60 the jumping was held on one hill; in 1964 there were two events, one on a 70-metre and the other on an 80-metre hill, and in 1968, 1972, and 1976 there were 70- and 90-metre events.

Olympic Champions, 1896–1976 (continued)

Winter sports (men) (continued)

biathlon

			hr	min	sec
1960	K Lestander	Sweden	1	33	21.6
1964	V. Melanin	U.S.S.R.	1	20	26.8
1968	M. Solberg	Norway	1	13	45.9
1972	M. Solberg	Norway	1	15	55.5
1976	N. Kruglov	U.S.S.R.	1	14	12.3

biathlon relay

1968		U.S.S.R.	2	13	02.4
1972		U.S.S.R.	1	51	44.9
1976		U.S.S.R.	1	57	55.6

luge

			min	sec
1964	T. Köhler	Ger.*	3	26.77
1968	M. Schmid	Austria	2	52.48
1972	W. Scheidel	E.Ger.	3	27.58
1976	D. Guenther	E.Ger.	3	27.69

luge (pairs)

1964		Austria	1	41.62
1968		E.Ger.	1	35.85
1972		Italy	1	28.35
1976		E.Ger.	1	25.60

two-man bobsled

			min	sec
1932	U.S.		8	14.74
1936	U.S.		5	29.29
1948	Switz.		5	29.2
1952	W.Ger.		5	24.54
1956	Italy		5	30.14
1964	Gt.Brit.		4	21.90
1968	Italy		4	41.54
1972	W.Ger.		4	57.07
1976	E.Ger.		3	44.42

four-man bobsled

			min	sec
1924	Switz.		5	45.54
1928	U.S.		3	20.5†
1932	U.S.		7	53.68
1936	Switz.		5	19.85
1948	U.S.		5	20.1
1952	W.Ger.		5	07.84
1956	Switz.		5	10.44
1964	Can.		4	14.46
1968	Italy		2	17.39
1972	Switz.		4	43.07
1976	E.Ger.		3	40.43

500-metre skating

			sec
1924	C. Jewtraw	U.S.	44.0
1928	C. Thunberg	Finland	43.4
	B. Evensen	Norway	
1932	J. Shea	U.S.	43.4
1936	I. Ballangrud	Norway	43.4
1948	F. Helgesen	Norway	43.1
1952	K. Henry	U.S.	43.2
1956	Ye. Grishin	U.S.S.R.	40.2
1960	Ye. Grishin	U.S.S.R.	40.2
1964	R. McDermott	U.S.	40.1
1968	E. Keller	W.Ger.	40.3
1972	E. Keller	W.Ger.	39.4
1976	Y. Kulikov	U.S.S.R.	39.2

1,000-metre skating

			min	sec
1976	P. Mueller	U.S.	1	19.3

1,500-metre skating

			min	sec
1924	C. Thunberg	Finland	2	20.8
1928	C. Thunberg	Finland	2	21.1
1932	J. Shea	U.S.	2	57.5
1936	C. Mathisen	Norway	2	19.2
1948	S. Farstad	Norway	2	17.6
1952	H. Andersen	Norway	2	20.4
1956	Yu. Mikhaylov	U.S.S.R.	2	08.6
	Ye. Grishin			
1960	Ye. Grishin	U.S.S.R.	2	10.4
	R. Aas	Norway		
1964	A. Antson	U.S.S.R.	2	10.3
1968	C. Verkerk	Neth.	2	03.4
1972	A. Schenk	Neth.	2	03.0
1976	J. E. Storholt	Norway	1	59.4

5,000-metre skating

			min	sec
1924	C. Thunberg	Finland	8	39
1928	I. Ballangrud	Norway	8	50.5
1932	I. Jaffee	U.S.	9	40.8
1936	I. Ballangrud	Norway	8	19.6
1948	R. Liaklev	Norway	8	29.4
1952	H. Andersen	Norway	8	10.6
1956	B. Shilkov	U.S.S.R.	7	48.7
1960	V. Kosichkin	U.S.S.R.	7	51.3
1964	K. Johannesen	Norway	7	38.4
1968	F. Maier	Norway	7	22.4
1972	A. Schenk	Neth.	7	23.6
1976	S. Stensen	Norway	7	24.5

10,000-metre skating

			min	sec
1924	J. Skutnabb	Fin.	18	4.8
1932	I. Jaffee	U.S.	19	13.6
1936	I. Ballangrud	Nor.	17	24.3
1948	A. Seyffarth	Swed.	17	26.3
1952	H. Andersen	Nor.	16	45.8
1956	A. Ericsson	Swed.	16	35.9
1960	K. Johannesen	Nor.	15	46.6
1964	J. Nilsson	Swed.	15	50.1
1968	J. Hoeglin	Swed.	15	23.6
1972	A. Schenk	Neth.	15	01.3
1976	P. Kleine	Neth.	14	50.6

figure skating

1908	U. Salchow	Sweden
1920	G. Grafström	Sweden
1924	G. Grafström	Sweden
1928	G. Grafström	Sweden
1932	K. Schäfer	Austria
1936	K. Schäfer	Austria
1948	R. Button	U.S.
1952	R. Button	U.S.
1956	H. Jenkins	U.S.
1960	D. Jenkins	U.S.
1964	M. Schnelldorfer	Ger.*
1968	W. Schwarz	Austria
1972	O. Nepela	Czech.
1976	J. Curry	Gt.Brit.

ice hockey

1920	Canada	
1924	Canada	
1928	Canada	
1932	Canada	
1936	Gt.Brit.	
1948	Canada	
1952	Canada	
1956	U.S.S.R.	
1960	U.S.	
1964	U.S.S.R.	
1968	U.S.S.R.	
1972	U.S.S.R.	
1976	U.S.S.R.	

figure skating (pairs—men and women)

1908	Anna Hübler and H. Burger	Germany
1920	Ludovika and W. Jakobsson	Finland
1924	Hellne Engelmann and A. Berger	Austria
1928	Andrée Joly and P. Brunet	France
1932	Andrée and P. Brunet	France
1936	Maxie Herber and E. Baier	Germany
1948	Micheline Lannoy and P. Baugniet	Belg.
1952	Ria and B. Falk	W.Ger.
1956	Elisabeth Schwarz and K. Oppelt	Austria
1960	Barbara Wagner and R. Paul	Canada
1964	Ludmilla Belousova and O. Protopopov	U.S.S.R.
1968	Ludmilla Belousova and O. Protopopov	U.S.S.R.
1972	Irina Rodnina and A. Ulanov	U.S.S.R.
1976	Irina Rodnina and A. Zaitsev	U.S.S.R.

figure skating (dance)

1976	Ludmila Pakhomova and A. Gorshkov	U.S.S.R.

Winter sports (women)

5-kilometre cross-country

			min	sec
1964	K. Boyarskikh	U.S.S.R.	17	50.5
1968	T. Gustafsson	Sweden	16	45.2
1972	G. Kulakova	U.S.S.R.	17	00.5
1976	H. Takalo	Finland	15	48.7

10-kilometre cross country

1952	L. Wideman	Finland	41	40.0
1956	L. Kozyreva	U.S.S.R.	38	11.0
1960	M. Gusakova	U.S.S.R.	39	46.6
1964	K. Boyarskikh	U.S.S.R.	40	24.3
1968	T. Gustafsson	Sweden	36	46.5
1972	G. Kulakova	U.S.S.R.	34	17.8
1976	R. Smetanina	U.S.S.R.	30	13.4

3 x 5 kilometre relay‡

1956	Finland	1	09	01.0
1960	Sweden	1	04	21.4
1964	U.S.S.R.		59	20.2
1968	Norway		57	30.0
1972	U.S.S.R.		48	46.2
1976	U.S.S.R.	1	07	49.8

slalom

			min	sec
1948	C. Fraser	U.S.	1	57.2
1952	A. Lawrence	U.S.	2	10.6
1956	R. Colliard	Switz.	1	52.3
1960	A. Heggtveit	Canada	1	49.6
1964	C. Goitschel	France	1	29.86
1968	M. Goitschel	France	1	59.85
1972	B. Cochran	U.S.	1	31.24
1976	R. Mittermaier	W.Ger.	1	30.54

giant slalom

			min	sec
1952	A. Lawrence	U.S.	2	06.8
1956	O. Reichert	Ger.*	1	56.5
1960	Y. Ruegg	Switz.	1	39.9
1964	M. Goitschel	France	1	52.24
1968	N. Greene	Canada	1	51.97
1972	M.-T. Nadig	Switz.	1	29.90
1976	K. Kreiner	Canada	1	29.13

downhill

			min	sec
1948	H. Schlunegger	Switz	2	28.2
1952	T. Jochom-Beiser	Austria	1	47.1
1956	M. Berthod	Switz.	1	40.7
1960	H. Beibl	Germany*	1	37.6
1964	C. Haas	Austria	1	55.39
1968	O. Pall	Austria	1	40.87
1972	M.-T. Nadig	Switz.	1	36.68
1976	R. Mittermaier	W.Ger.	1	46.16

alpine combined

1936	C. Cranz	Ger.		
1948	T. Beiser	Austria		
1972	A. Proell	Austria		
1976	R. Mittermaier	W.Ger.		

luge

			min	sec
1964	O. Enderlein	Ger.*	3	24.67
1968	E. Lechner	Italy	2	29.37
1972	A. Muller	E.Ger.	2	59.18
1976	M. Schumann	E.Ger.	2	50.62

500-metre skating

			min	sec
1960	H. Haase	Germany*		45.9
1964	L. Skoblikova	U.S.S.R.		45.0
1968	L. Titova	U.S.S.R.		46.1
1972	A. Henning	U.S.		43.3
1976	S. Young	U.S.		42.8

1,000-metre skating

1960	K. Guseva	U.S.S.R.	1	34.1
1964	L. Skoblikova	U.S.S.R.	1	32.6
1968	C. Geijssen	Neth.	1	32.6
1972	M. Pflug	W.Ger.	1	31.4
1976	T. Averina	U.S.S.R.	1	28.4

1,500-metre skating

			min	sec
1960	L. Skoblikova	U.S.S.R.	2	25.2
1964	L. Skoblikova	U.S.S.R.	2	22.6
1968	K. Mustonen	Finland	2	22.4
1972	D. Holum	U.S.	2	20.8
1976	G. Stepanskaya	U.S.S.R.	2	16.6

3,000-metre skating

1960	L. Skoblikova	U.S.S.R.	5	14.3
1964	L. Skoblikova	U.S.S.R.	5	14.9
1968	J. Schut	Neth.	4	56.2
1972	S. Baas-Kaiser	Neth.	4	52.1
1976	T. Averina	U.S.S.R.	4	45.2

figure skating

1908	M. Syers	Gt.Brit.
1920	M. Mauroy	Sweden
1924	H. Planck-Szabo	Austria
1928	S. Henie	Norway
1932	S. Henie	Norway
1936	S. Henie	Norway
1948	B. Scott	Canada
1952	J. Altwegg	Gt.Brit.
1956	T. Albright	U.S.
1964	S. Dijkstra	Neth.
1968	P. Fleming	U.S.
1972	T. Schuba	Austria
1976	D. Hamill	U.S.

*Joint East–West German team. †Five men. ‡4×5 kilometre relay in 1976.

Olympic Champions, 1896–1976 (continued)

Sports events no longer included

Archery

1900

game shooting	Mackintosh	Australia
cordon doré (50 metres)	H. Herouin	France
chapelet	E. Mougin	France
cordon doré (33 metres)	H. van Innis	Belgium
la perche à la herse	E. Foulon	France
au chapelet (33 metres)	H. van Innis	Belgium
la perche à la pyramide	E. Grumiaux	France

1904

individual events (men)

American round	H. Taylor	U.S.
York round	P. Bryant	U.S.

individual events (women)

Columbia round	M. Howell	U.S.
national round	M. Howell	U.S.

team events

men	U.S.
women	U.S.

1908

men's events

York round	W. Dod	Gt.Brit.
Continental style	E. Grizot	France

women's event

national round	Q. Newall	Gt.Brit.

1920

individual events (men)

fixed target (small)	E. van Meer	Belg.
fixed target (large)	E. Clostens	Belg.
moving target (28m)	H. van Innis	Belg.
moving target (33m)	H. van Innis	Belg.
moving target (50m)	L. Brulé	France

Archery (continued)

1920

women's events

individual competition	O. Newal	Gt.Brit.

team events

fixed target (2 events)	Belg.
moving target (28m)	Neth.
moving target (33m)	Belg.
moving target (50m)	Belg.

Baseball

1912 U.S.
1936 U.S.
1952 Fin. (Finnish baseball)

Gliding

1936 H. Schreiber Ger.

Glima (Icelandic wrestling)

1912 Iceland

Golf

men

1900	C. Sands	U.S.
1904	G. Lyon	Canada

women

1900	M. Abbot	U.S.

Gymnastics

rope climbing

1896	N. Andriakopoulos	Greece
1904	G. Eyser	U.S.
1924	B. Supcik	Czech.
1932	R. Bass	U.S.

Swedish exercises (team)

1912 Sweden
1920 Sweden

Gymnastics (continued)

optional exercises (team)

1912 Norway
1920 Denmark
1932 U.S.

parallel bars (team)

1896 Ger.

horizontal bar (team)

1896 Ger.

Indian clubs

1904 E. Hennig U.S.
1932 G. Roth U.S.

tumbling

1932 R. Wolfe U.S.

combined competition (9 events)

1904 A. Spinnier Ger.

triathlon

1904 M. Emmerich U.S.

seven-day event

1904 A. Heida U.S.

prescribed apparatus (team)

1904 U.S.
1908 Swed.
1912 Italy

mass exercises (team)

1952 Finland

women hand apparatus (team)

1952 Sweden

woman team drill

1956 Hungary

side horse (vaults)

1924 A. Séguin Fr.
1932 I. Pelle Hung.

Jeu de paume (Royal tennis)

1908 J. Gould U.S.

Lacrosse

1904 Canada
1908 Canada
1948 U.S. and Gt.Brit.

Motor boat racing (1908)

Class A	E. Thubron's	"Camille"	France
Class B	T. Thorneycroft's	"Cyrinus"	Gt.Brit.
Class C	T. Thorneycroft's	"Cyrinus"	Gt.Brit.

Mountaineering

1932 F. Schmidt Ger.
T. Schmidt

Polo

1908 Gt.Brit.
1920 Gt.Brit.
1924 Arg.
1936 Arg.

Rackets

1908	*singles*	E. Noel	Gt.Brit.
	doubles	V. Pennel	Gt.Brit.
		J. Astor	

Roque

1904 C. Jacobus U.S.

Rugby football

1900 France
1908 Australasia
1920 U.S.
1924 U.S.

Tennis (covered courts) (indoor tennis)

men's singles

1908 A. Gore Gt.Brit.
1912 A. Gobert France

women's singles

1908 G. Eastlake-Smith Gt.Brit.
1912 E. Hannam Gt.Brit.

Tennis (covered courts) (continued)

men's doubles

1908 A. Gore, H. Roper-Barrett Gt.Brit.
1912 M. Germot, A. Gobert France

mixed doubles

1912 E. Hannam, C. Dixon Gt.Brit.

Tennis (lawn)

men's singles

1896	J. Boland	Gt.Brit.
1900	L. Doherty	Gt.Brit.
1904	B. Wright	U.S.
1908	M. Ritchie	Gt.Brit.
1912	C. Winslow	S.Af.
1920	L. Raymond	S.Af.
1924	V. Richards	U.S.

women's singles

1900	C. Cooper	Gt.Brit.
1908	D. Chambers-Lambert	Gt.Brit.
1912	M. Broquedis	France
1920	S. Lenglen	France
1924	H. Wills	U.S.

Tennis (lawn) (continued)

men's doubles

1896	J. Boland	Gt.Brit.
	F. Thraun	Ger.
1900	H. and R. Doherty	Gt.Brit.
1904	E. Leonard, B. Wright	U.S.
1908	G. Hillyard, R. Doherty	Gt.Brit.
1912	C. Kitson, C. Winslow	S.Af.
1920	O. Turnbull, M. Woosnam	Gt.Brit.
1924	F. Hunter, V. Richards	U.S.

women's doubles

1920	H. McNair, K. McKane	Gt.Brit.
1924	H. Wills, H. Wightman	U.S.

mixed doubles

1900	C. Cooper, R. Doherty	Gt.Brit.
1912	D. Köring, H. Schomburgk	Ger.
1920	S. Lenglen, M. Décugis	France
1924	H. Wightman, R. Williams	U.S.

Winter sports

1922

combined speed skating (men)
C. Thunberg Finland
military ski patrol
Switzerland
curling Gt.Brit.

1928

military ski patrol Norway
one-man bobsled J. Heaton U.S.

Winter sports (continued)

1932

sled-dog race E. Goddard Canada
curling Canada

1936

military ski patrol	Italy	
ice shooting (team)	Austria	
distance shooting	G. Edenhauser	Austria
target shooting	I. Reiterer	Austria

1948

skeleton bobsled	N. Bibbia	Italy
military ski patrol	Switzerland	
winter pentathlon	G. Lindh	Sweden

1964

Austrian curling Austria

BIBLIOGRAPHY

History of ancient games: H.A. HARRIS, *Greek Athletes and Athletics* (1964); E.N. GARDINER, *Greek Athletic Sports and Festivals* (1910), *Athletics of the Ancient World* (1930).

Revival of modern Olympics: BARON PIERRE DE COUBERTIN, *Mémoire Olympique* (1931), *Une Campagne de vingt-et-un ans* (1887–1908); W. HENRY, *An Approved History of the Olympic Games* (1948); J. KIERAN and A. DALEY, *The Story of the Olympic Games, 776 B.C. to 1968*, rev. ed. (1969).

Reports of games: There are full Official Reports of every celebration since 1904. P. LOVESEY and T. MCNAB, *The Guide to British Track and Field Literature from 1275–1968* (1969), contains information about 900 works, including foreign works and in particular of 50 works on the Olympic Games.

(H.M.A.)

Atlantic Ocean

The Atlantic Ocean is the name given to that vast stretch of world ocean that separates the continents of Europe and Africa from those of North and South America. The term, derived from Greek mythology, means the Sea of Atlas. It is second in size only to the Pacific Ocean (*q.v.*).

The Atlantic is, generally speaking, S-shaped and narrow in relation to its length. The area of the Atlantic without its dependent seas is 31,800,000 square miles (82,400,000 square kilometres), and with them, 41,100,000 square miles (106,400,000 square kilometres).

The ocean's drainage area

Although not the largest of the world's oceans, the Atlantic has by far the largest drainage area. The continents on both of its sides tend to slope toward it, so that it receives the waters of a large proportion of the great rivers of the world, including the St. Lawrence, the Mississippi, the Orinoco, the Amazon, the Río de La Plata, the Congo, the Niger, the Loire, the Rhine, the Elbe, and the great rivers of the Mediterranean and the Baltic. The total area of land draining to the Atlantic and the Arctic Sea is nearly 16,691,000 square miles (43,229,700 square kilometres), almost four times the area draining to the Pacific Ocean, and almost exactly four times the area draining to the Indian Ocean (*q.v.*).

In the north, between the island of Spitsbergen and Greenland, the waters of the Atlantic merge with those of the Arctic Ocean; in the southeast, between the southern tip of Africa and Antarctica they merge with those of the Indian Ocean; in the southwest, Drake Passage (*q.v.*) connects the South Atlantic and South Pacific oceans. The Atlantic and Pacific are also connected by the Panama Canal (*q.v.*) about 10° north of the Equator, as well as via the difficult and virtually unused Northwest Passage (*q.v.*), which runs through the Arctic Ocean to the north of North America.

A large proportion of the world's fishing grounds are located in the Atlantic. Vast mineral reserves have been located on the ocean floor as well as on the floor of the continental shelves on either side of the ocean. At the present time the greater part of the nonbiological resources being exploited consists of petroleum and gas. The North Atlantic is the most heavily travelled seaway in the world. (For associated articles, see CONTINENTAL DRIFT; CONTINENTAL SHELF AND SLOPE; OCEAN BASINS; OCEAN CURRENTS; OCEANIC RIDGES; OCEANS AND SEAS; OCEANS, DEVELOPMENT OF; for associated physical features, see BAFFIN BAY; BALTIC SEA; BARENTS SEA; BISCAY, BAY OF; BLACK SEA; CARIBBEAN SEA; DRAKE PASSAGE; ENGLISH CHANNEL; FUNDY, BAY OF; GUINEA, GULF OF; GULF STREAM; HUDSON BAY; IRISH SEA; MEDITERRANEAN SEA; MEXICO, GULF OF; NORTH SEA; NORTHWEST PASSAGE; SAINT LAWRENCE, GULF OF; SCOTIA SEA; WEDDELL SEA.)

PHYSIOGRAPHY

Extent. The Arctic Basin, which stretches from the Bering Strait across the North Pole to Spitsbergen and Greenland, connects with the Atlantic Ocean by way of the narrow, but deep, straits between Spitsbergen and Greenland. In the south, the Atlantic extends to the shores of the Antarctic continent. The Atlantic Ocean has, therefore, a share in both the seas of ice.

In contrast to the South Atlantic, the North Atlantic is rich in islands, in the variety of its coastline, and in tributary seas. The latter include the Caribbean Sea, the Gulf of Mexico, the Gulf of St. Lawrence, Hudson Bay, and Baffin Bay on the west, and the Mediterranean Sea, Black Sea, North Sea, and Baltic Sea on the east. Between Spitsbergen and Novaya Zemlya, on the one hand, and the Murmansk coast, on the other, lies the Barents Sea; between Greenland, Iceland, the Faeroes, Shetlands, Norway, and Spitsbergen lies the Norwegian–Greenland Sea. Hudson Strait in the northwest is 64 miles wide; Davis Strait to the north of it 200 miles; and Denmark Strait, between east Greenland and Iceland, 159 miles across. The passage between Iceland and northern Scotland is 518 miles wide.

Tributary seas

In the South Atlantic, on the other hand, between Cape Horn and South Africa, the ocean approaches Antarctica on a 3,965-mile front, and is much colder and rawer than the North Atlantic.

From east to west, the ocean's breadth varies considerably. From Newfoundland to Ireland it is 2,059 miles; further south it widens to over 3,000 miles, before narrowing again so that the distance from Cabo São Roque, Brazil, to Cape Palmas, Liberia, is only 1,769 miles. Southward, it again becomes broader and is bordered by simple coasts almost without islands.

Relief of the ocean floor. The foundations of knowledge of the Atlantic floor were laid in the last century when the practical needs of telegraphic engineers and the curiosity of natural scientists led to the first explorations of the abyss. Though soundings were taken by lead line or wire, real comprehension of sea floor topography was not possible until the development of continuously recording precision echo sounders after World War II. The detailed profiles produced by these instruments allowed a realistic appraisal of the ocean floor, which made the earlier probings seem vague and almost dream-like.

The Mid-Atlantic Ridge

The outstanding feature of the Atlantic floor is the Mid-Atlantic Ridge, an immense median mountain range extending throughout the length of the Atlantic, claiming the centre third of the ocean bed, and reaching 1,000 miles in breadth. This feature, though of tremendous proportions, is but the Atlantic portion of a world-encircling ridge, the Mid-Oceanic Ridge, which stretches 45,000 miles around the entire globe.

In some places the Mid-Atlantic Ridge reaches above sea level. The Azores, Ascension, St. Helena, Tristan da Cunha, Gough, and Bouvet—all volcanic islands—rise from its flank; Iceland, which rises from its crest, is rent by an extension of the median rift valley. East and west of the ridge, 12,000 to 18,000 feet below sea level, lie basins which seem on first inspection to present a rather even profile. Study reveals, however, that parts of the basin floor are as mountainous as the Mid-Atlantic Ridge, while other parts are extremely smooth. The former are rocky abyssal hills; the latter are the abyssal plains that form the upper surface of great ponds of mud which fill many of the broad depressions. Large ancient volcanoes are found singly or in rows in the basins; these rise to form seamounts and occasionally islands. As the continents are approached and the rugged Mid-Atlantic Ridge is left behind, the echo-sounding profile of a survey vessel will first reveal an abyssal plain and then the smooth, undulating surface of the continental rise.

These broad embankments, which lie at depths of 8,000 to 15,000 feet at the foot of continents, reach in widths of over 300 miles off Northwest Africa, Angola, Argentina, and the eastern seaboard of the United States. In other areas they are exceedingly narrow. The continental rise is in fact an immense refuse heap of sediments eroded from the continents through geologic time. Here, in all probability—within the 10,000- to 50,000-foot-thick accumulations—lie the largest potential reserves of petroleum on Earth.

The continental shelf

The continents have steep sides and bevelled edges; the continental slope is cut by canyons that funnel sediments from continental sources to the continental rise (consisting of debris, which has accumulated at the foot of the continental slope) below. The continental shelf is the wetted perimeter of the continent, which has only re-

cently been claimed by the sea. A few thousand years ago the ocean level was lower, and the shelf belonged to the continent; almost certainly, it will again, in the future. It is this nonoceanic part of the ocean with which man is most concerned, for here fish and minerals are exploited, and here shoals create hazards to mariners.

The Caribbean Islands and the South Sandwich Islands form great unstable island arcs, where the greatest depths of the Atlantic are found in steep-sided, narrow gashes that drop to over 30,000 feet below sea level, and over 10,000 feet below the floors of adjacent basins. Earthquakes occur continuously along the arcs as the deformation of the Earth's crust actively proceeds in these unstable belts.

The North Polar Basin and the intervening Norwegian Basin are separated from the open Atlantic by a shallow ridge extending from Greenland to Scotland upon which Iceland and the Faeroe Islands rise above sea level. The maximum depths in Denmark Strait, between Greenland and Iceland, and over the Wyville-Thompson Ridge, between the Faeroes and Scotland, are only about 1,600 feet. Lowering the sea level about 1,600 feet would expose a land bridge from North America to Europe and would completely isolate the waters of the Polar Basin from both the Atlantic and the Pacific. In the North Polar Basin relatively few soundings were made until a Russian expedition (1937–38), which landed by plane on the ice within 60 miles of the pole. After that time numerous soundings were made from the ice floes and ice islands throughout the basin. The basin, roughly elliptical, is divided into two parts by the Lomonosov Ridge, having a sill depth of 5,000 feet, which runs from the continental shelf north of Ellesmere Island through a position of 89° N 180° W, then south near the meridian of 140° E toward the New Siberian Islands. The depression on the right looking north from Ellesmere Island is smaller, but deeper—over 16,000 feet—whereas, depths in the larger basin to the left (toward Alaska) approach 13,000 feet. There are two lobes in the larger basin and some evidence of a second ridge roughly parallel to the Lomonosov Ridge. Between Greenland and Spitsbergen the sill depth is about 5,000 feet, and in the Norwegian Basin the greatest depth is about 12,000 feet.

Depths greater than 13,000 feet occur in the Caribbean Basin and in the Mediterranean Sea. The former has numerous shallow and several deep connections with the open ocean, but the Mediterrean communicates with the Atlantic only through the Strait of Gibraltar, which is about 12 miles wide and where the maximum depth on the sill is only 1,000 feet. The partial isolation of the large adjacent seas has a profound effect on the conditions in the seas and also upon those in the open ocean.

(B.C.H.)

Oceanic islands

Islands. Among purely oceanic islands without any foundation of continental rock, usually the result of volcanic action, there are Jan Mayen, Iceland, the Azores, Ascension, St. Helena, Tristan de Cunha, and Bouvet (latitude 54° 26′ S), which rise from the Mid-Atlantic Ridge, and the Canaries, Madeira, the Cape Verdes, and Fernando de Noronha (near Cabo São Roque), rising from the continental margins of Africa and South America. Volcanic islands of a different sort are those of the two great arcs, the Lesser Antilles and the South Sandwich. Partly continental and partly oceanic are the Greater Antilles in the Caribbean, and South Georgia and the South Orkneys in the Scotia Sea. Purely continental are Spitsbergen and the Bear Islands, the British Isles, Newfoundland, and the Falkland Islands.

Ice. The area of ice-covered ocean is an important variable feature of the Earth. Variations in ice extent influence, and are influenced by, large-scale circulation of atmosphere and ocean. Seasonal variations and longer term variations associated with global climatic trends are much larger in the Antarctic than in the Arctic. In the Northern Hemisphere, maximum sea ice extent occurs in April and is about 5 percent of the area of the hemisphere. The September minimum is about 25 percent less. In the Southern Hemisphere, the maximum in October, is about 8 percent of the area, while the March minimum is about 75 percent less. Typical Antarctic pack ice is under one year old and three to seven feet thick, while typical Arctic pack ice is several years old and ten to 13 feet thick, melting about three feet at the surface during summer and growing about three feet at the bottom during winter.

Pack ice is in constant motion, mainly under the influence of wind, with open leads and fracture patterns giving the appearance of a giant jigsaw puzzle. Most of the pack ice leaves the Arctic Ocean east of Greenland where the southerly drift reaches 20 miles per day. Roughly 720 cubic miles (3,000 cubic kilometres) of ice enter the Greenland Sea yearly, where most of it melts before reaching the latitude of Iceland. Large variations in ice extent are associated with global climatic trends. Following a long warming trend, there was an advance of sea ice in the 1960s, comparable to that of the last century, which threatened the economy of Iceland. The causes of these changes are unknown.

Icebergs

Icebergs are another form of ice found especially in the North and South Atlantic. In the Northern Hemisphere, the main source of icebergs is the southwest coast of Greenland, between 69° and 73° N, where the six fastest moving glaciers on Earth release over 5,000 icebergs a year into Baffin Bay. Altogether, the Labrador Current every year carries 7,000 to 8,000 icebergs south. Most of these melt before reaching the Grand Bank, but about 400 drift south of Newfoundland and about 50 reach the southern tip of the Grand Bank.

The glacial flow from Antarctica is mostly in broad streams which form ice shelves hundreds of feet thick. Fragments from the outer edges form the tabular icebergs, sometimes over 60 miles long, found in the southern ocean. Tabular icebergs are formed in the Arctic only rarely, on a smaller scale, and only in a small region of Ellesmere Island and the Franz Josef Islands. In 1972 seven manned, drifting Arctic scientific stations were situated on such ice islands.

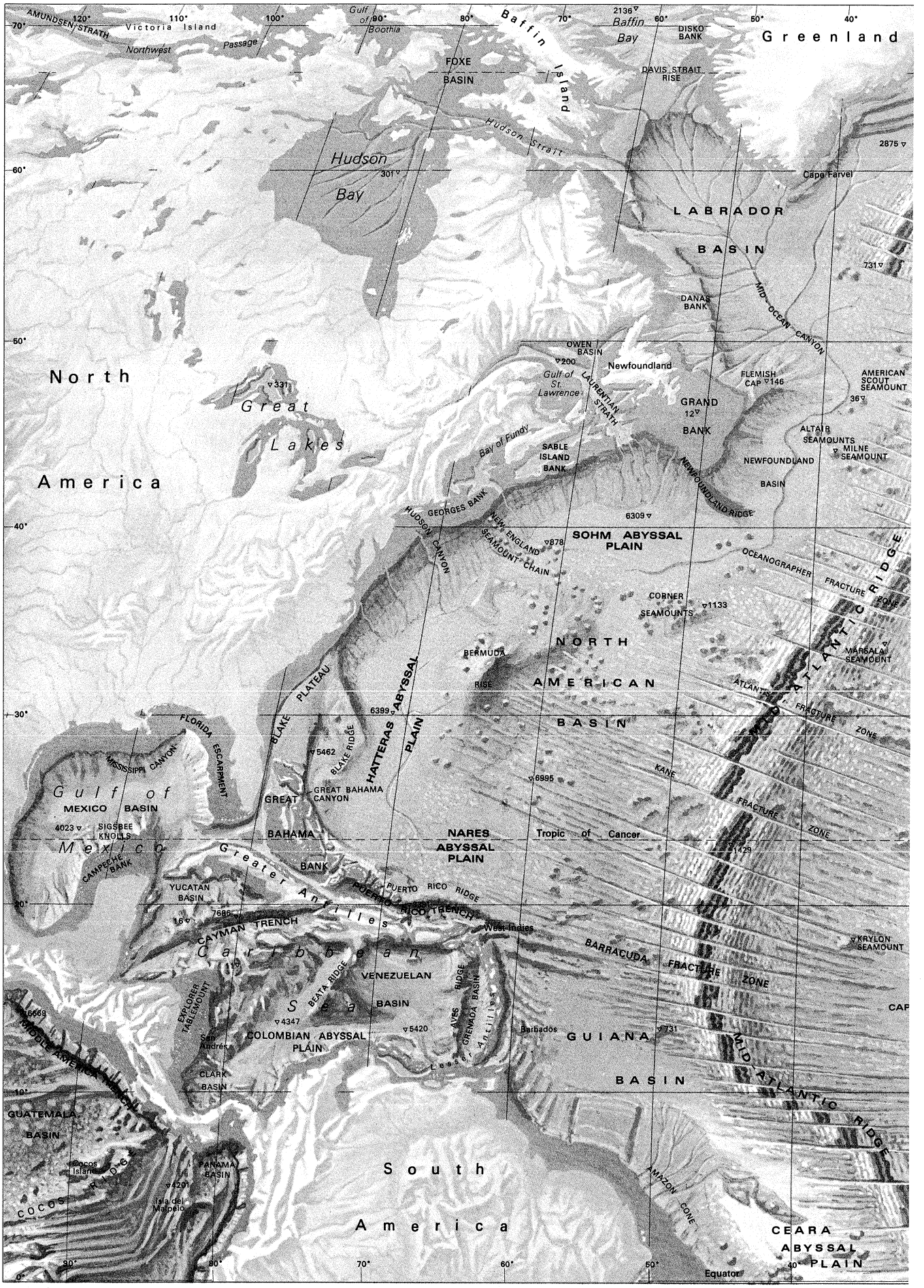
(J.O.Fl.)

GEOLOGY

The Atlantic is the youngest of the oceans. Its origin and development are now accounted for by the theory of continental drift, according to which a vast "proto-continent," Pangaea, which once "floated" on the Earth's liquid mantle, broke up in the late Mesozoic Era about 150,000,000 years ago; the rifting caused the landmasses of the Western and Eastern hemispheres to separate, opening up the Atlantic Ocean Basin. As can be seen on the map, the continental coastlines of North America and Europe, and of South America and Africa almost seem to match. If the edges of the continental shelves are matched, the fit is almost perfect. Other geological and paleontological similarities on both sides of the Atlantic have been found to substantiate the theory of continental drift and, thus, to help explain the evolution of the Atlantic.

Perhaps the most conclusive evidence bearing out this theory of origin is to be found in the existence of the Mid-Atlantic Ridge. The ridge is, in effect, a long rift zone of mountains, volcanoes, and faulted plateaus. A high-heat flow, which is associated with the extrusion of magna due to sea-floor spreading, exists in the rift zone. The crustal material on either side of the Ridge is notably younger than that on the corresponding plateaus, indicating an uprising of material from the Earth's mantle onto the crest of the ridge. The newer rock is mainly composed of gabbro (a coarse-grained rock formed deep down under heat and pressure), basalt (a rock which originally poured out at the surface in molten form), and serpentine (a common rock-forming mineral). Consequent movement of the ocean floor and of the continents in opposite directions outward from the ridge is resulting in an increasing widening of the Atlantic Basin at an estimated rate of from less than half an inch to about four inches (one to ten centimetres) a year. A corresponding spreading is occurring at an even faster pace in the Pacific Ocean; in the Atlantic, however, the slower rate of spreading causes the flanks of the ridge to be built up steeply by accumulating lava. The physiography and ge-

Movement of the ocean floor

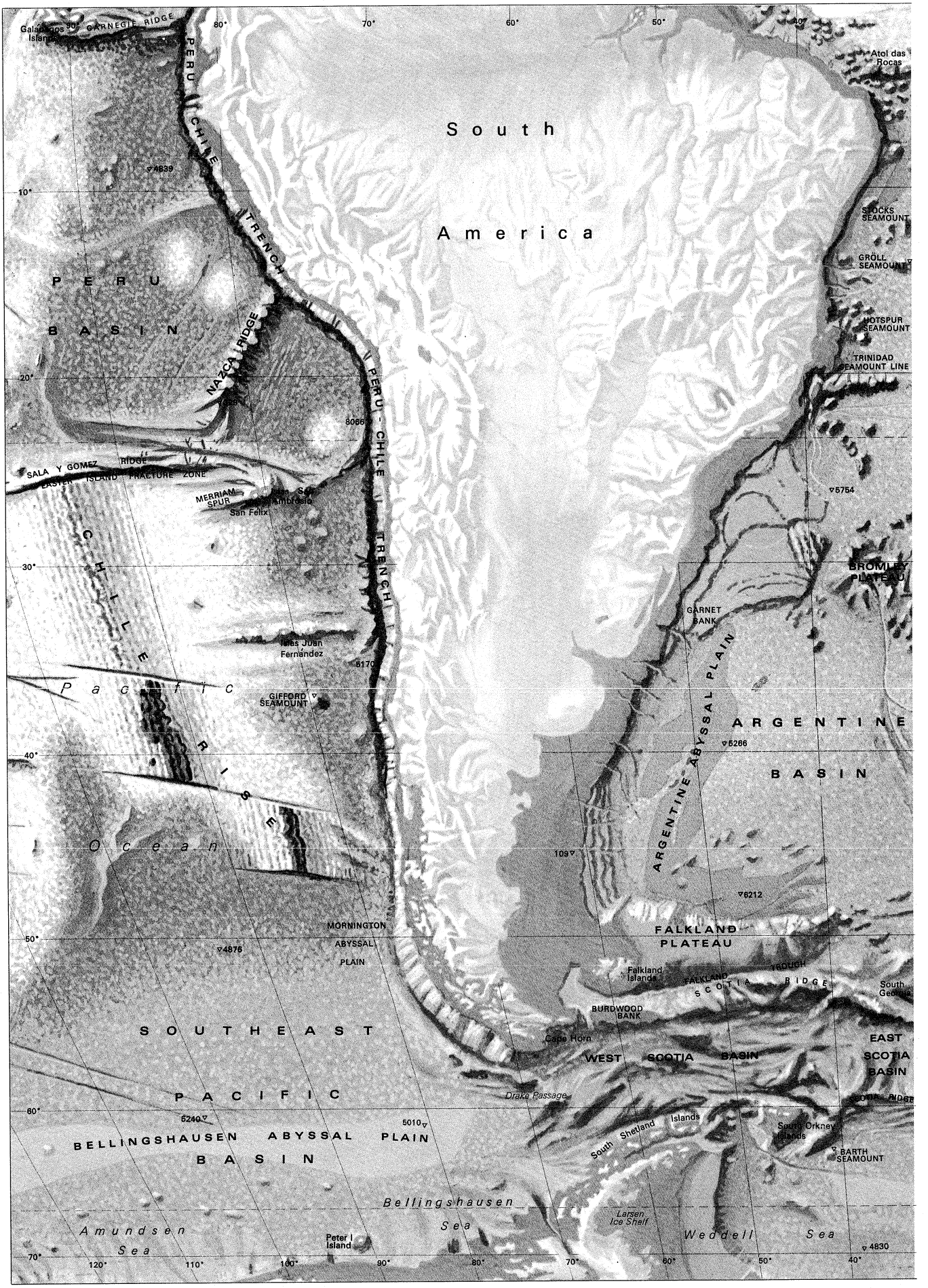


NORTHERN ATLANTIC OCEAN

Depths in metres

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


Mayen
LOFOTEN BASIN
Novaya Zemlya
Barents Sea
NANSEN CORDILLERA
SOUTHERN RIDGE
NORWEGIAN BASIN
GREENLAND-ICELAND RISE
Arctic Circle
Iceland
Norwegian Sea
FAEROE-ICELAND RIDGE
Faeroe Islands
Gulf of Bothnia
REYKJANES RIDGE
Shetland Islands
WYVILLE THOMPSON RIDGE
NORWEGIAN TRENCH
Rockall
ROCKALL RISE
ROCKALL TROUGH
North Sea
Baltic Sea
DEVILS HOLE
DOGGER BANK
British Isles
GIBBS FRACTURE ZONE
PORCUPINE BANK
HECATE SEAMOUNT
FARADAY SEAMOUNT GROUP
GREAT SOLE BANK
English Channel
Europe
WEST EUROPEAN BASIN
BISCAY ABYSSAL PLAIN
CELTIC ESCARPMENT
IBERIAN ABYSSAL PLAIN
Black Sea
ADRIATIC ABYSSAL PLAIN
ALGERIAN BASIN
TYRRHENIAN ABYSSAL PLAIN
AZORES
Azores PLATEAU
AZORES-GIBRALTAR RIDGE
Mediterranean Sea
Aegean Sea
Strait of Gibraltar
MALTA RISE
MEDITERRANEAN RIDGE
TAMPERE SEAMOUNT
ATLANTIS SEAMOUNT
Madeira
MESSINA ABYSSAL PLAIN
LEVANTINE BASIN
CANARY BASIN
CRUISER TABLEMOUNT
GREAT METEOR TABLEMOUNT
Canary Islands
CANARY RIDGE
CAPE VERDE BASIN
ECHO SEAMOUNT
TROPIC SEAMOUNT
CAPE VERDE TERRACE
Cape Verde Islands
Cape Verde
Africa
VERDE BASIN
GUINEA FRACTURE ZONE
SIERRA LEONE RISE
NIGER FAN
Rochedos São Pedro São Paulo
ROMANCHE FRACTURE ZONE
GUINEA BASIN
São Tomé
© Rand McNally & Co.
A-513800-957 -1 -1 -1 -1
0 200 400 600 800 1000 km
0 200 400 600 800 mi



SOUTHERN ATLANTIC OCEAN Depths in metres

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


Fernando de Noronha
PERNAMBUCO ABYSSAL PLAIN
6537
CHAIN FRACTURE ZONE
GUINEA RISE
7758
5759
2330
CONGO CANYON
Ascension Island
6119
BRAZIL BASIN
5574
ANGOLA BASIN
Africa
St. Helena
6050
Ilhas Martim Vaz
SEAMOUNT
MID-ATLANTIC RIDGE
Tropic of Capricorn
WALVIS RIDGE
EWING SEAMOUNT
VALDIVIA SEAMOUNT
VEMA SEAMOUNT
5457
WÜST SEAMOUNT
CAPE BASIN
AGULHAS BANK
Tristan da Cunha
WYANDOT SEAMOUNT
AFRICANA SEAMOUNT
NATAL SEAMOUNT
SCHMIDT-OTT SEAMOUNT
AGULHAS PLATEAU
2310
Gough Island
ENKER SEAMOUNT
411
DISCOVERY TABLEMOUNT
5636
CAPE AGULHAS BASIN
HERDMAN SEAMOUNT
MOZAMBIQUE FRACTURE ZONE
319
Prince Edward Islands
FALKLAND FRACTURE ZONE
2310
560
METEOR SEAMOUNT
MERZ SEAMOUNT
PRINCE EDWARD FRACTURE ZONE
ATLANTIC-INDIAN RIDGE
OB TABLEMOUNT
247
LENA TABLEMOUNT
SPIESS SEAMOUNT
413
Bouvet Island
SOUTH SANDWICH TRENCH
METEOR DEEP
8428
South Sandwich Islands
5865
6972
ENDERBY ABYSSAL PLAIN
WEST ATLANTIC-INDIAN BASIN
WEDDELL ABYSSAL PLAIN
ATLANTIC-INDIAN BASIN
MAUD SEAMOUNT
840
GUINERUS BANK
Antarctic Circle
Antarctica
677
30°
20°
10°
0°
10°
20°
30°
40°
50°
60°
70°
© Rand McNally & Co.
A-513900-957 -1 -1 -1-2
0 200 400 600 800 1000 km
0 200 400 600 800 mi

ology of the Mid-Atlantic Ridge are now the subject of much scientific study, as is the geologic study of the Atlantic as a whole. Such investigations were undertaken only comparatively recently, since the theories of continental drift and of sea-floor spreading were not generally accepted until the later 1960s. (I.M.P.)

BOTTOM DEPOSITS

The greater part of the bottom of the Atlantic between the Arctic and the Antarctic circles is covered with globigerina ooze (a calcareous deposit consisting mainly of the shells of dead one-celled animals). At depths greater than 16,400 feet the calcium carbonate content decreases, and the calcareous deposits give way to red clay. On submarine ridges the finer material is lacking, and the shells of pteropod gastropods (mollusks of the gastropod class comprising the snails) may be sufficiently abundant to characterize the deposits as pteropod ooze. Diatom ooze (formed from microscopic unicellular algae having cell walls consisting of or resembling silica) is the most widespread deposit in the high southern latitudes but, contrary to conditions in the Pacific, is not found in northern latitudes. The bottom itself is covered with mud (oozes, globigerina, etc.), 60 percent; sand, 25 percent; and rock, gravel, and shells, 15 percent. Airborne material is abundant off the west coast of Africa where dry offshore winds carry material from the desert regions. In high latitudes ice-rafted detritus, including rock fragment, sometimes showing the effect of glacial abrasion, is an important component. The calcium carbonate content of the sediments of the Atlantic is notably greater than that at comparable depths and latitudes in the Pacific.

Turbidity currents

After World War II thousands of sediment cores, some exceeding 66 feet in length, were collected in the North and South Atlantic by means of a piston coring tube. These cores revealed the importance of turbidity currents—that is, occasional catastrophic torrents of sediment-laden, hence denser, water flowing downslope under clear water—as carriers of great quantities of sediment to the greatest depths in the Atlantic. Since the last ice age, turbidity currents have been relatively infrequent, with the consequence that the characteristic deposits laid down by them are as a rule covered by a few inches of normal pelagic sediment. Study of the shells of planktonic foraminifera (microscopic, floating, unicellular organisms) in these cores shows that the climatic changes, ice ages, and interglacial ages of the last 2,000,000 years have been recorded in the sediments as alternations of species adapted to cold or to warm water. In the 1960s the Joint Oceanographic Institutions for Deep Earth Sampling deep-drilling project penetrated the entire thickness of sediment in the Atlantic. Apparently the oldest sediments in the Atlantic Basin accumulated during the Mesozoic Era (from 65,000,000 to 225,000,000 years ago). Dating of sediment layers by radioactive decay or by examination of the traces in rocks of reversals of the Earth's magnetic poles (which occur every few million years) shows that the rate of accumulation of pelagic sediment in the Atlantic is between two-fifths and four-fifths of an inch per 1,000 years. Locally, however, the rate is much faster because of deposition by turbidity currents. (D.B.E.)

CLIMATE

Weather over the North Atlantic is largely determined by large-scale wind currents and air masses emanating from North America. In winter the Prevailing Westerly Winds at levels of 10,000 to 40,000 feet over North America meander in such a way that a northward bulge (ridge) is generated by and over the Rockies and a southward bulge (trough) over the eastern half of the continent. This geographically forced flow pattern sets the stage for the frequent deployment of cold air masses from Canada and Alaska to the Atlantic seaboard. Large temperature contrasts are thereby frequently set up between the polar outbreaks and mild air from the Pacific or tropical air from the Gulf of Mexico or Gulf Stream. Along these zones of contrast, which are called fronts, wave cyclones (low pressure areas) are formed, and these develop into strong vortices as they move toward Newfoundland and Iceland. Their growth rate largely depends on the temperature contrast, so that winter storms usually become appreciably stronger than summer storms in terms of low pressure, wind, and severe weather.

Storm patterns

These cyclonic storms carry heat, moisture, and momentum northward from the tropics and thereby siphon off the excess heat constantly generated by solar heating in the tropics. They also contribute a large share of the energy required to maintain the Prevailing Westerlies of midlatitudes, which are found to be one-half as strong and about 10° further north over the North Atlantic in summer than in winter.

Since even in winter the temperatures of air masses along the eastern seaboard vary considerably from one week to another, the number of coastal storms, their growth rate, and even their paths may vary. Thus, despite the underlying fixed geography, the North Atlantic average pressure distribution, on which the prevailing winds depend, may show large differences from one January to another. In some winters, Iceland may be dominated by prevailing high pressure in contrast to the normally low pressure, and in this case storms leaving the North American coast are blocked and shunted into the Davis Strait and to the Azores. When this happens, warm maritime air masses that normally flow into Europe and account for its mild climate are replaced by cold air from the European Arctic and from Siberia.

Thus, it is not surprising that in winter tremendous amounts of heat are extracted from the western North Atlantic by overflowing cold air masses. Although the transfer of real (sensible) heat is large, the transfer of heat by evaporative losses into the cold, dry air is about three times larger. The oceanic heat losses are soon restored by the flow of warm water associated with the Gulf Stream and other currents. The net effect of the increase in heat and moisture off the east coast is to further stimulate the growth of cyclonic storms. In the United States, an observer looking eastward from a high point over Boston may frequently see banks of large cumulus clouds (clouds showing great vertical development in the form of heaps or piles of cloud) at sea as a manifestation of rapid offshore heating when a cold air mass flows out over the warmer seas.

In latitudes 15° N to 30° N the North Atlantic is characterized by prevailing high pressure with an attendant lack of intense storms and severe weather. These high-pressure areas are part of a globe-encircling belt in which air from the westerlies to the north and from the tropics to the south sinks about 900 feet a day, and is warmed by compression so that the weather is often sunny and rainless. South of this North Atlantic high-pressure zone the Northeast Trade Winds blow with characteristic steadiness.

Although low latitudes of the North Atlantic are usually storm free, there are notable exceptions during later summer and early fall, when wavy patterns in the east winds occur and occasionally develop into tropical storm vortices or hurricanes. The hurricanes grow by the liberation of vast amounts of heat released when vapour evaporated from the warm ocean is lifted and condensed to bands of heavy showers. Hurricanes may live more than a week, travelling as severe wind vortices steered by upper air currents. Thus they frequently move clockwise around the periphery of the North Atlantic high-pressure belt and into the Prevailing Westerlies, often ending up in the Icelandic area. They have, however, occasionally struck England, and even the Azores, in modified form with abnormal upper-air wind patterns.

Hurricanes

Over the South Atlantic the belt of Prevailing Westerlies extends from about 40° S almost to Antarctica, and the South Atlantic high-pressure area is centred around 30° S. This anticyclone (circulation of winds around a central region of high atmospheric pressure) leads to Southeast Trade Winds on its north side, since the rotation of wind around the high-pressure area is opposite to that in the Northern Hemisphere, due to the Coriolis force (the effect caused by the Earth's rotation). The Southeast Trades meet the Northeast Trades in the zone

roughly centring on the Equator called the doldrums or intertropical convergence. Here heavy showers result from ascending warm, moist air that is being continually replaced by moistened trade wind air.

As in the North Atlantic, the weather is usually settled and fine in the latitudes of the high pressure but is unsettled and stormy in the higher latitudes of the Westerlies. The great storminess of the Southern Hemisphere Westerlies largely derives from the temperature contrast set up by the cold Antarctic continent and the adjacent open sea, rather than the west–east contrast described in connection with the North Atlantic storms.

While many regional weather peculiarities may be found over the Atlantic, one of the most interesting, especially to the jet-age traveller, is the large amount and variety of cloud in the Westerlies. These clouds are continually being generated by the large cyclonic storm systems, by warm, moist air masses condensing while moving northward over colder water (in advance of storms), and by rapid vertical ascent (convection) produced by cold air streaming over warm water. Extensive fog banks may frequently be seen in summer off the Grand Bank, when heated air from the continent is forced to flow over the cold Labrador Current. (J.Na.)

HYDROGRAPHY

Surface currents. The surface currents of the Atlantic Ocean primarily correspond to the system of prevailing winds with such modifications as are imposed upon the movement of the water by land boundaries. Other factors that influence the currents are regional excesses of evaporation or precipitation, regional differences in cooling or heating, friction, and the Earth's rotation. In the North Atlantic the Trade Winds maintain a fairly steady current from east to west, partly by the direct action of the wind and partly by maintaining an accumulation of warm water on the northern side of the current. A great bulk of water carried by this current continues into the Caribbean Sea and through the Strait of Yucatán into the Gulf of Mexico from which it flows out as a warm and swift current through the Strait of Florida. This current, reinforced by water which has flowed on the eastern side of the Antilles, forms the Gulf Stream (*q.v.*) off the American east coast. The Gulf Stream follows the coast closely as far as Cape Hatteras; continues at some distance from the coast; and turns more and more toward the east, flowing due east to the south of the Grand Banks of Newfoundland in latitude 40° N. In its further course, the Gulf Stream loses its identity as a well-defined current. The warm surface waters turn to the right and form part of the big eddy circulating around the Sargasso Sea (an area of the North Atlantic between the West Indies and the Azores, characterized by relatively still waters and containing gulfweed [*sargassum*] and specialized marine animals). Somewhat colder water continues toward the European coast as the North Atlantic Current. One diffuse branch turns south, and another branch turns north and splits up still more. One part, the Irminger Current, turns northwest, washes the southeast coast of Iceland and continues past the southern cape of Greenland. The waters of this current become gradually mixed with cold low-salinity water from the Arctic Ocean, but the last traces of Gulf Stream water are still found in latitude 65° N off western Greenland. Another branch of the Gulf Stream system enters the Norwegian Sea to the north of Scotland. One small portion turns south into the North Sea, but the major part follows the coast of Norway to North Cape and continues to Spitsbergen, sending minor branches into the Barents Sea. North of Spitsbergen the current submerges below the less saline waters of the Arctic and continues as a subsurface current clear across the Arctic Ocean where traces of Gulf Stream water of temperature slightly above 32° F (0° C) are found to the north of the New Siberian Islands (north of Siberia, separating the Laptev and East Siberian seas).

The Gulf Stream

The surface layer of the Arctic Ocean, throughout the year, is at a temperature close to the freezing point (at 29° F, or −1.6° C) but is of relatively low density because the salinity has been reduced by runoff from the great Siberian rivers. This cold, low-salinity water flows out from the Arctic Ocean along the east coast of Greenland where it is gradually mixed with Atlantic water. It continues around the south cape of Greenland, Cape Farvel; flows north along the west coast of Greenland; turns around again; and, after addition of cold water from Baffin Bay, flows south as the cold Labrador Current. To the south of the Grand Bank of Newfoundland, where this cold water meets the warm waters of the Gulf Stream, it is deflected toward the east and mixes with the Atlantic water. In winter this mixed water, with a salinity of almost 35 parts per thousand, is cooled to a temperature of nearly 37° F (3° C), whereby it attains a density high enough to make it sink to the bottom and spread to the south. On an average for the whole year, 5,230,000 cubic yards of water sink every second (about 86 cubic miles per day). Similarly, bottom water is formed in winter to the north of Iceland, but this has a considerably lower temperature, about 30° F (−1° C). It fills the deep basin of the Norwegian Sea but is prevented from returning directly into the Atlantic Ocean by the submarine ridge which extends from Scotland to Iceland and from Iceland to Greenland; after some intermixing, the Norwegian Basin water eventually crosses the ridge to the Atlantic.

In the southeast part of the North Atlantic, surface water flows into the Mediterranean and high-salinity Mediterranean deep water flows out along the bottom of the strait and spreads over wide areas. Along the west coast of northwest Africa, the Canary Current flows to the southwest and continues across the southern part of the North Atlantic as part of the North Equatorial Current. Low temperatures prevail on the African coast.

The currents of the South Atlantic correspond in many respects to those of the North Atlantic. The Southeast Trade Winds maintain the South Equatorial Current that flows toward the west where it divides into two branches, one that continues to the Northern Hemisphere and enters the Caribbean together with water from the North Equatorial Current, and one which turns south as the Brazil Current, a weak counterpart of the Gulf Stream. Between the Equatorial currents, the Equatorial Countercurrent flows toward the east and is particularly well-developed off Ghana, where it is known as the Guinea Current. To the south of the high-pressure area of the South Atlantic the current flows to the east and turns toward the Equator when reaching the African coast. There the Benguela Current is more pronounced than its northern counterpart, the Canary Current, and is characterized by lower temperatures near the coast, caused by more intense upwelling. Further south the Antarctic Circumpolar Current enters the Atlantic Ocean through Drake Passage, sending one branch, the Falkland Current, a counterpart of the Labrador Current, along the east coast of Argentina. The major branch of the Antarctic Circumpolar Current continues to the east into the Indian Ocean, sending another branch to the south and feeding a large clockwise eddy in the Weddell Sea.

African coastal currents

Deepwater currents. The deep and bottom water of the North Atlantic, as already stated, consists of surface water sinking between Iceland and Greenland and in the Labrador Sea, from where it spreads to the south. At depths between 3,000 feet and 6,500 feet the water that flows out from the Mediterranean spreads and can be recognized by an intermediate salinity maximum. With increasing distance from the Mediterranean the salinity decreases because of mixing with other water masses, but traces of Mediterranean water are found as far south as latitude 40° S.

In the Antarctic, bottom water with a temperature of about 31° F (−0.6° C), and salinity of 34.65 parts per thousand, is formed by sinking of water from the continental shelf. The temperature of this water is so low that its density is higher than that of the North Atlantic deep water. This water flows toward the north and can be traced as bottom water to 40° N. Surface water sinks at the Antarctic Convergence in about latitude 50° S and spreads to the north as low-salinity water. This Antarctic

intermediate water also crosses the Equator and can be traced to about 20° N. Large amounts of the Antarctic bottom water and intermediate water mix with the North Atlantic deep water, return to the south, and rise toward the surface between 50° and 60° S latitude. In rising, the deep water brings quantities of plant nutrients, phosphates, and nitrates to the surface layers, and the oceanic circulation therefore accounts for the high productivity of the Antarctic waters.

The deep and bottom waters of the Atlantic are characterized by a high oxygen content because there exists a fairly rapid circulation. The waters have sunk from the surface where they became saturated with oxygen by contact with the air.

TEMPERATURE AND SALINITY

Temperature. The distribution of the sea-surface temperature is closely related to the character of the currents. The waters of the North Equatorial Currents spread to the north and to the south when reaching the east coasts of North and South America and, correspondingly, the region of high surface temperature is wide off the American east coasts but narrow off the African coast, where the Canary Current and the Benguela Current carry cold water toward the Equator. Therefore, in latitudes about 10° S to 30° S and 10° N to 30° N the sea surface is warmer off the eastern coast than off the western, but poleward of 30° this feature is reversed. This reversal is barely evident in the South Atlantic, where the Falkland Current carries cold water up to about latitude 30° S (in August to 25° S), but is conspicuous in the North Atlantic. There the Labrador Current brings cold water to latitude 40° N, whereas the extreme branches of the Gulf Stream system carry warm water along the coast of Norway where ports remain ice-free even in latitude 71° N. The contrast between the South and the North Atlantic is related to the surface currents, which in turn reflect the action of the prevailing winds and the effect of the shape of the coasts. Where the Falkland Current meets the Brazil Current, and where the Labrador Current meets the Gulf Stream, the surface temperature changes rapidly within a very short distance. The change is particularly striking when passing from the Gulf Stream to the Labrador Current where the phenomena is called "the cold wall."

The Labrador Current

In the Tropics the surface temperature is controlled by climatological factors to such an extent that it is nearly uniform, and differences related to currents do not appear. Such differences are very marked, however, at a depth of about 650 feet, where in latitude 6° to 7° N the temperature is 50° F (10° C), whereas it is 68° F (20° C) in latitude 20° N. The existence of the cold water at shallow depths to the north of the Equator should not be interpreted as showing that deep water rises to the surface. The temperature distribution is directly related to the existence of Equatorial currents—which flow toward the west, and within which the warm water must be to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.

The distribution of temperature at greater depths has already been touched upon when discussing the deep water circulation. In the North Atlantic, the temperature decreases slowly toward the bottom from a value of about 41° F (5° C) at 3,000 ft. to about 36.5° F (2.5° C) at the bottom. In the South Atlantic, up to latitude 40° S, the temperature first decreases to a minimum between 3,000 feet and 4,000 feet, increases again and reaches a maximum of 36° to 39° F (2° to 4° C) at about 6,500 feet, indicating the flow of North Atlantic deep water, and then decreases to less than 34° F (1° C) at the bottom where Antarctic bottom water is encountered. To the south of 40° S low temperatures prevail throughout, and near Antarctica a large body of water has a temperature below 32° F (0° C).

Salinity. The surface waters of the North Atlantic Ocean have a higher salinity than those of any other ocean, reaching values exceeding 37 parts per thousand in latitudes 20° to 30° N. The salinity distribution is also related to the currents but is greatly influenced by evaporation and precipitation. It has been shown that for each ocean the average surface salinity can be taken as equal to a constant basic value plus a correction that is directly proportional to the difference between evaporation and precipitation. The basic salinity value differs from one ocean to another and is highest for the North Atlantic. It is 35.5 parts per thousand for the North Atlantic and 34.5 for the South Atlantic. This difference can be explained as the effect of the intense evaporation in the Mediterranean and the outflow from that sea of high-salinity water that maintains the salinity of the North Atlantic at a higher level than that characteristic of any other ocean. On an average for every latitude range, say 0° to 5° N, and so on, the deviations from the basic value are proportional to the difference between evaporation and precipitation. Near the Equator precipitation dominates, and surface salinities of about 35 parts per thousand are encountered; but in latitudes 20° to 25° N and about 20° S evaporation greatly exceeds precipitation, and over large areas the surface salinity is above 37 parts per thousand. Proceeding poleward precipitation again becomes greater than evaporation, and, correspondingly, the surface salinity decreases, in large areas to values below 34 parts per thousand. Superimposed upon these general features are the effects of currents, which again are more striking in the North Atlantic where Atlantic water of salinity exceeding 35 parts per thousand is carried as far north as Spitsbergen in latitude 78° N, and Arctic water of salinity below 34 parts per thousand is carried south to nearly 45° N off Newfoundland. North of 40° N the sea-surface isohalines (lines of equal salinity) run nearly in a north–south direction, whereas south of 45° S they run east–west. In adjacent seas the salinity depends also upon the runoff from rivers. In the Mediterranean and the Red Sea, where the runoff is small and evaporation is great, high salinities prevail; in the Black Sea and in the Baltic, where large rivers empty, the salinity is low. In the inner part of the Gulf of Bothnia between Sweden and Finland, the water is nearly fresh. The surface water of the Polar Sea has a salinity of 30 to 33 parts per thousand because of admixture of freshwater from the great Siberian rivers.

Evaporation and precipitation

(H.U.S./R.H.Fl./C.A.B.)

MARINE LIFE

The Atlantic is divided into five water layers or zones, each harbouring a unique type of marine life. From top to bottom these are: the littoral benthic, euphotic, mesopelagic, bathypelagic, and benthic zones.

The five ecological zones

The littoral (tidal) zone, which has a maximum depth of 200 feet, contains the near-shore varieties of animals such as the oysters, clams, mussels, and burrowing animals, as well as most types of attached seaweed. Sea moss, kelp (a large brown seaweed), fucus (an olive-brown seaweed) and brown algae are the most common.

The euphotic zone (lighted surface layers), which has a maximum depth of 600 feet, is the sunlit portion of open water which harbours the organisms that manufacture 90 percent of the life-sustaining nutrients in the sea. These organisms—zooplankton (microscopic sea animals) and phytoplankton (single-celled green sea plants)—are both part of a mass assemblage of floating organisms. The vast quantities of plankton are consumed by other sea dwellers, including herring, mackerel, tuna, sharks, and baleen whales.

Below this zone is the mesopelagic (literally mid-oceanic) realm at a depth of from 660 to 3,000 feet; it is the habitat of sperm whales and of giant squids, which feed upon the organic debris that sinks down from the euphotic zone.

The bathypelagic (deep ocean) zone, which occurs at a depth of from 3,000 to 13,200 feet is permanently dark; it is here that the phenomenon of bioluminescence is found. Bioluminescent animals are those which are able to produce their own light; almost two-thirds of the fish in this zone have this characteristic. These include the anglerfish, lanternfish, and viperfish. The most common explanation for the development of bioluminescence is its use as a lure to attract prey.

The bottom, or benthic zone, which occurs at an average depth of 12,000 feet contains animals of elementary structure, including the large attached invertebrates such as the crinoids (sea lilies) and glass sponges. Certain brachiopods (bottom-dwelling invertebrates), once thought to be extinct, exist here in large numbers. These three varieties of animals consume organic matter from the upper zones. Fish are uncommon, except for poor swimmers such as the grenadier fish (related to cod) and the tripod fish (an abyssal fish which awaits its prey resting on its fins and tail). The upwelling of cold waters returns nutrients to the littoral benthic and euphotic zones, sustaining the animal life there, and so completing the ocean's life cycle.

ECONOMIC RESOURCES

Biological resources. The Atlantic Ocean contains six of the world's 14 major fishing grounds. They are roughly situated in the northwest, northeast, west central, east central, southwest, and southeast Atlantic.

Principal fishing grounds

Regions of high fertility, supporting large quantities of marine life, lie along the Atlantic seaboard of the North and South American continents. The waters of the Grand Bank off Newfoundland are among the most populous fishing grounds in the world. Here, where the waters of the Gulf Stream and Labrador Current merge, herring and menhaden (a fish resembling a shad but with a more compressed body) thrive in great numbers. The northeast Atlantic and North Sea have long been traditional fishing grounds, but are in danger of being overfished. Since the 1950s, fishing has increased south of the Equator, with large quantities of tuna, hake, and herring being caught. The waters off the Florida Keys, especially Key West, abound in shrimp and sponges. The calm Sargasso Sea, bounded on the east by the Canary Current and on the west by the Gulf Stream, is the breeding ground of all the eels of both Europe and North America. Clams, lobsters, crabs, and octopus are other important biological resources.

Mineral resources. Ever since the British "Challenger" expedition conducted a pioneering oceanographic survey from 1872 to 1876, it has been known that the Atlantic Ocean contains large stores of manganese nodules. These nodules are distributed most abundantly in the red clay bottom deposits throughout 10 percent of the pelagic (oceanic) area. Their composition is roughly 24 percent manganese and 14 percent iron; the nodules are formed of concentric layers, anywhere from several microns (a micron is a millionth part of a metre) in size to large slabs.

Large diamond deposits have also been located off the coast of South West Africa; the diamonds were washed into the ocean millions of years ago. The diamondiferous sediments vary in thickness from six to 30 feet; the stones are almost all of gem quality, and few industrial diamonds are found among them. The richest area so far located is along the southwest African coastline between Hottentot's Bay and Oranjemund.

Petroleum, natural gas, and sulfur are found in the Gulf of Mexico and off the west central coast of Africa. Petroleum and gas represent 90 percent of the Atlantic's available mineral resources; they occur in the sediments of the continental rise.

Petroleum and gas reserves

Smaller deposits of coal and tin are found in sediments off the British Isles. Salt occurs along most of the Atlantic seaboard, and 70 percent of the world's bromine is extracted from the ocean waters. The water of the ocean itself, in desalinated form, constitutes a potential, though still extremely expensive, resource.

Resource exploitation. Exploitation in the second half of the 20th century has produced a dangerous shortage of many species of fish in several of the traditional fishing grounds. The Atlantic at the present time yields 40 percent of the world's catch of fish. The situation is most serious in the North Atlantic, where several varieties of fish, including flounders, ocean perch, cod, hake, herring, and tuna, are being overfished. South of the Equator, the yield, except for tuna and pilchard, is not yet in danger. Tuna has been the most exploited species, and, as a result, yields are becoming more and more meagre not only in the Atlantic but also in other oceans.

No comparable problem, however, faces the mining industry. A single diamond dredge, for instance, sweeps up by suction $420,000 worth of diamonds off the coast of South West Africa every day, and the resources are just beginning to be tapped. There seems to be an almost unlimited reserve of manganese nodules, but they have as yet been little exploited because it is still far more expensive to reclaim them from the sea than to mine manganese on land. Petroleum and natural gas reserves are being exploited off the coast of southern United States in the Gulf of Mexico, and petroleum resources are also being tapped in European and African territorial waters.

HISTORY

Exploration. The story of the surface exploration of the Atlantic is well known. Centuries before the age of Columbus, the Vikings in their 11th-century wooden ships charted the waters around Greenland, Iceland, and northeastern North America. After the Portuguese discovery of Madeira and the Azores in the 15th century, and the transatlantic voyage of Christopher Columbus (*q.v.*) in 1492 to 1493, Europeans of many nationalities joined in the exploration of the lands bordering the Atlantic's western shores. Not much scientific attention, however, was paid to the Atlantic until 1842, when a pioneering oceanographic study of the Atlantic and the Pacific was made by Lieutenant Matthew Fontaine Maury of the United States Navy. Maury compiled charts on winds and currents, collected other data, and prepared an extensive treatise on the Gulf Stream, thus paving the way for the further studies of the Atlantic. A Scottish naturalist, Sir Wyville Thompson, discovered the Mid-Atlantic Ridge on the "Challenger" expedition already mentioned, using temperature variations as indicators that a vast barrier existed below the surface. His findings were substantiated by a German expedition of 1925–1927, which verified the presence of a distinct mountain range.

Exploration was not, however, conducted solely from the surface. As interest in oceanography increased, such deep-sea vehicles as Trieste I, a deep-sea bathyscaph built for Jacques Piccard, a member of a well-known Swiss family of scientists and explorers, were constructed, and used to obtain further information about the ocean's depths. Many different types of submarine have also been used to gather information.

New methods of underwater research

The "Glomar Challenger," a United States deep-sea drilling ship, has made extensive surveys of Atlantic and adjacent waters; in 1970 it discovered a vast oil field 12,000 feet below the Gulf of Mexico. Yet another technique is to establish an underwater habitat in order to study marine environments for an extended period. One such habitat, the United States Sealab, has housed four aquanauts for three weeks in waters off Bermuda, enabling them to study marine life in its natural environment.

Naval predominance. Following early Portuguese and other explorations, Spain became the dominant naval power in the Atlantic in the 16th century. Later, Spain's naval supremacy was challenged by French, British, and Dutch ships. In the 19th century, Britain established a naval supremacy that endured until the first half of the 20th century, during which United States seapower developed. At the present time both the United States and the Soviet Union have powerful naval forces in Atlantic waters, and are competing with each other in oceanographic research and its applications.

NAVIGATION

The areas of congested sea traffic all lie in the North Atlantic. As ice conditions occur at certain times of the year, the International Convention for the Safety of Life at Sea was adopted in 1948 in order to systematize and regulate use of shipping lanes. Each lane is desginated for use at a certain time of year. Countries are also assigned their own routes in order to relieve traffic congestion; for instance, the United States is assigned a route far to the

Shipping lanes

south only when extreme ice conditions occur. An International Ice Patrol, formed and operated by the United States Coast Guard, maintains surveillance of icebergs. The United States Navy and the Royal Canadian Air Force fly air reconnaissance flights over both the Atlantic and Arctic oceans for the same purpose, while certain European nations maintain surveillance in the Baltic and North seas. The Soviet northern sea route, where many icebergs are to be found, is also kept under close observation. Hardly any regular ice reconnaissance flights are flown south of the Equator, but satellite observations of Antarctic ice limits were being recorded in the 1970s.

Among the many harbours on the Atlantic, the port of Amsterdam, used by ships from all over the world, is perhaps the busiest. Punta Arenas in Chile is the southernmost port in the South Atlantic, and Archangel in the Soviet Union is the northernmost North Atlantic port.

BIBLIOGRAPHY. J. BARDACH, *Harvest of the Sea* (1968), a general, elementary text on the resources of the oceans; E. BULLARD, "The Origin of the Oceans," *Scient. Am.*, 221:66–75 (1969); J.S. DOUGLAS, *The Story of the Oceans* (1952), a general, historical text on the exploration of the oceans; R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography* (1966), an excellent technical publication on the various aspects of oceanography; J.D. ISAACS, "The Nature of Oceanic Life," *Scient. Am.* 221:146–160 (1969), a consideration of the major life zones in the oceans; E.J. LONG, *New Worlds of Oceanography* (1965), text covering major discoveries such as fishing grounds and mining potentials, with chapters on submersibles and research vessels; H.W. MENARD, "The Deep-Ocean Floor," *Scient. Am.* 221:126–132 (1969), a report on the composition and resources of the abyssal zone; E. OROWAN, "The Origin of the Oceanic Ridges," *Scient. Am.*, 221:18, 102–108 (1969), a somewhat technical paper on the geological origin of the oceanic ridges, presenting several theories; J. SCOFIELD, "The Lower Keys, Florida's Out Islands," *Natn. Geogr. Mag.*, 139:72–93 (1971); U.S. HYDROGRAPHIC OFFICE, *Table of Distances Between Ports Via the Shortest Navigation Routes* (1943), an official table of sea lanes and instructions on their usage; E. WENK, "Physical Resources of the Oceans," *Scient. Am.*, 221:166–176 (1969), a paper on new resource discoveries in the Atlantic, Pacific, and Indian Oceans; G. YOUNG, "Everything's Coming Up Diamonds," *Sci. Dig.*, 68:68–73 (1970), an article on diamond dredging off South West Africa.

(I.M.P.)

Atlas Mountains

The Atlas Mountains form the geological backbone of the countries of the Maghrib region—Morocco, Algeria, and Tunisia. They extend for more than 1,200 miles (2,000 kilometres), from the Moroccan port of Agadir in the southwest, to the city of Tunis in the northeast. Their thick rim rises to form a high sill separating the Mediterranean basin from the Sahara, thus constituting a barrier that hinders, without completely preventing, communication between north and south. Across the mountains filter both air masses and human migrations. It is, however, only in the east–west direction that the Atlas Mountains facilitate movement. These are the conditions that create at the same time both the individuality and the homogeneity of the Atlas countries. (For coverage of the three countries of the Maghrib, see the articles ALGERIA; MOROCCO; and TUNISIA; for coverage of associated physical features, see the articles MEDITERRANEAN SEA; and SAHARA.)

The mountains. The Atlas mountain system takes the shape of an extended oblong, enclosing within its ranges a whole complex of plains and plateaus.

The northern section is formed by the Atlas Tellien (Tell Atlas), so-called because it receives enough rainfall to bear fine forests. From west to east several massifs (mountainous masses) occur. The first of these is er-Rif, which forms a half-moon-shaped arc between Ceuta and Melilla; its crest line is over 5,000 feet (1,500 metres) at several points, reaching 8,058 feet (2,456 metres) at Jebel Tidirhine. Beyond the Moulouya gap the Algerian ranges begin, among which the rugged bastion of Ouarsenis (which reaches a height of 6,510 feet [1,984 metres]), the Grande Kabylie Massif which reaches 7,500 feet (2,307 metres) at the peak of Lalla Kredidja, and the mountains of Kroumirie in Tunisia are all prominent.

The southern section, which is subject to desert influences, is deservedly called the Atlas Saharien. It includes in the centre a palisade formed by shorter ranges, such as the Ksour and Ouled Naïl mountains, grouped into massifs between two mighty ranges—the Moroccan Haut Atlas to the west, and the Aurès mountains to the east. The Haut Atlas culminates in Jebel Toubkal 13,665 feet (4,165 metres), which is surrounded by high snow-capped peaks; the Aurès is formed of long parallel folds, which reach a height of 7,638 feet (2,328 metres) at Djebel Chélia.

The Atlas Tellien and Atlas Saharien merge in the west into the long folds of the Moyen Atlas and in the east join together in the Tébessa and Medjerda mountains.

Geology. If the relief of the Atlas region is relatively simple, its geology is complex. In essence, the two Atlases comprise two different structural regions.

Formation of the Atlas Tellien

The Atlas Tellien originally arose out of a basin filled with sediment, which was dominated to the north by a marginal rim, of which the massifs of Tizi-Ouzou, Collo, and Edough are the remnants. Its elevation took place during a lengthy mountain-building process, which was marked by upheavals in the Tertiary Period (which lasted from 65,000,000 to 2,500,000 years ago); over the cluster of folds that was uplifted from the rift valley were spread sheets of flysch (deposits of sandstones and clays), which were carried down from the north over the top of the marginal rim. Thus the Atlas Tellien represents an example of a young folded mountain range still in the process of formation, as is shown by the earth tremors to which it is still subject.

To the south, the Atlas Saharien belongs to another structural grouping, that of the vast plateaus of the African continent, which form part of the ancient base rock largely covered by sediments deposited by shallow seas and by alluvial deposits. The Atlas Saharien results either from the mighty folding of the substructure that raised up fragments of the base rock, such as the horst (uplifted block of the earth's crust), which constitutes the Moroccan Haut Atlas, or else from the crumpling into folds of the earth's crust during the Jurassic Period (from 190,000,000 to 136,000,000 years ago) and the Cretaceous Period (from 136,000,000 to 65,000,000 years ago).

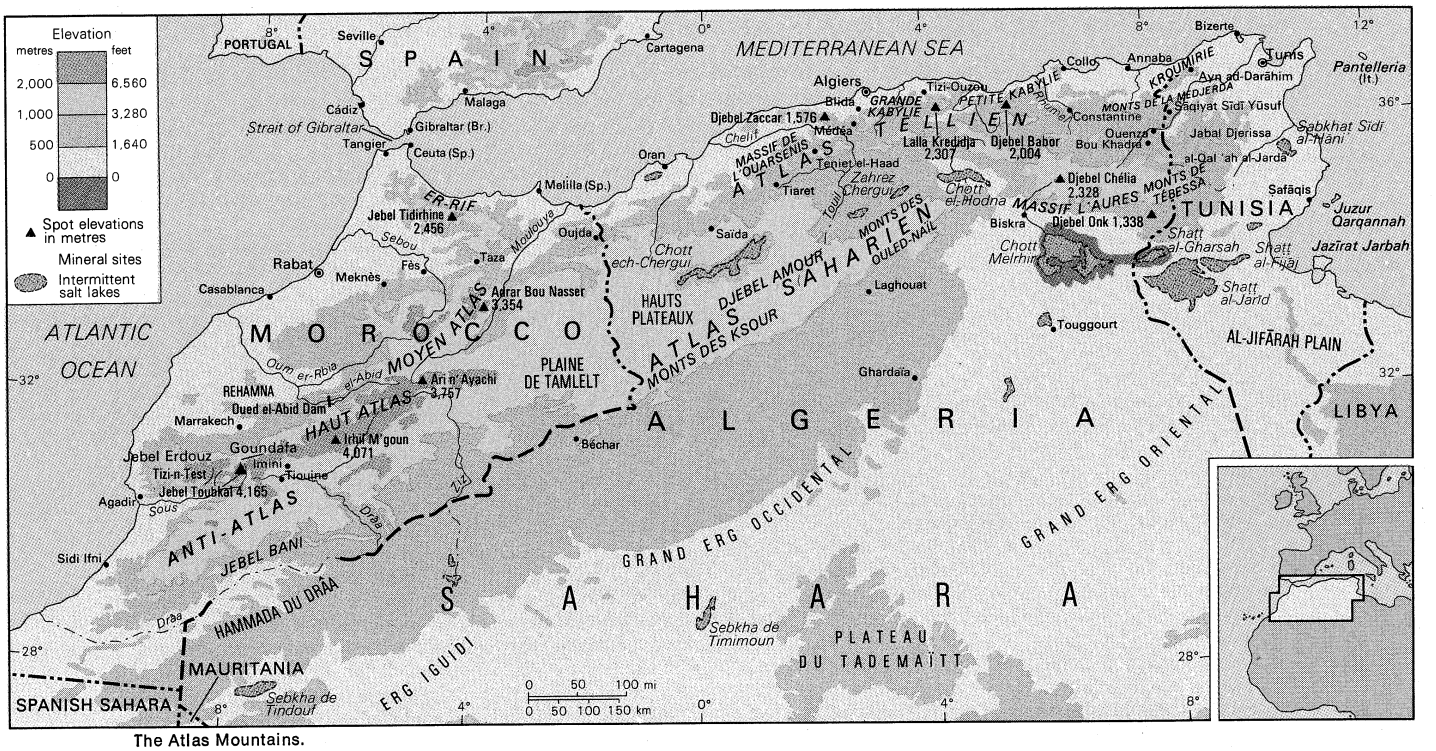
Climate. The Atlas Mountains are the meeting place of two different kinds of air masses—the humid and cold polar air masses that come from the north, and the hot and dry tropical air masses that move up from the south. To the influence of altitude and of latitude, there must therefore also be added that of aspect or exposure.

Rain is more plentiful in the Atlas Tellien than in the Atlas Saharien, and more so to the northeast than to the southwest: thus the highest rainfall is recorded in the east of the Atlas Tellien. ʿAyn ad-Darāhim in the Kroumirie mountains receives 60 inches a year; nowhere in the Anti-Atlas Mountains, which are to the south of the Haut Atlas, is the total rainfall more than 17 inches (432 millimetres) a year. In a single massif the slopes exposed to the north receive more rainfall than those exposed to the south; mountain dwellers distinguish between slopes that in the Berber language are called *amalou* (humid), and those that are dry or *assammeur*.

With increased altitude the temperature drops rapidly; despite the proximity of the sea, the coastal massifs are cold regions. At 6,574 feet (2,004 metres) the summits of Djebel Babor in the Petite Kabylie region are covered with snow for four or five months, while the Moroccan Haut Atlas retains its snows until the height of summer. Winter in the Atlas is hard, imposing severe living conditions upon the inhabitants.

Hydrology. The seasonal character of the rains, which fall in torrents, determines the characteristics of drainage in the Atlas: the runoff feeds streams that are of great erosive capacity, and that have cut their way down through the thickness of accumulated layers of sediment to form deep narrow gorges difficult to cross. The fortress of Cirta, now called Constantine, stands on a rock sculptured out by one such stream, the winding Rhumel River.

The great Maghribian wadis, or oueds, (channels of watercourses that are dry except during periods of rain)



The Atlas Mountains.

issue from the Atlas mountain ranges. Among them are the Moulouya that rises from the Moyen Atlas and the Chelif that rises from Djebel Amour. Destructive of the soils of their headstreams, they deposit their loads of silt at the foot of the mountain ranges or else leave a long line of conical deposits locally known as *dirs* (hills).

Soils. Soil worthy of the name is rare at higher altitudes in the Atlas region. Most often nothing is to be found but bare rock, debris, and fallen materials incessantly renewed by landslides. Two materials predominate—limestone, which forms ledges that are half-buried in rough debris, and marls (chalky clays) cut by erosion into a maze of ravines and crumbling gullies. The rarer sandstones favour forest growth. The best soils are the alluvia found on the terraced slopes and on the valley bottoms.

Vegetation. Erosion is aggravated by the sparseness of the vegetation covering the landscape; only 8,000,000 hectares (20,000,000 acres) of land are forested. On er-Rif, Kabylie, and Kroumirie ranges, which experience rainfall, moist forests of cork oaks cover an undergrowth of arbutus (cane apple) and heather shrub, and carpets of rockroses and lavender. When the total rainfall is below 31 inches (790 millimetres) and limestone is present, green oak and arbor vitae (a species of pine tree) cover the soil, forming light, dry forests with a thin and bushy undergrowth. The cedar predominates at higher altitudes. On the dry summits of the Atlas Saharien the vegetation is reduced to scattered stands of green oak and juniper trees.

The clearance of land for agriculture has resulted in the disappearance of the forest in the Atlas ranges; the animal life of the mountains is also in retreat. There only remain a few bands of jackals, some tribes of monkeys, and some herds of wild boar in the oak woods.

The inhabitants. The mountains, with their inhospitable environment, have provided a refuge for the original inhabitants, who have fled successive invasions. Here the Berber race has survived, preserving its own language, traditions, and beliefs, while at the same time accepting Islām to some extent. Village communities still live according to a code of customary law, known as *kanun*, which deals with all questions of property and persons. The family unit traces its descent from a single ancestor, preserving its cohesion by the sense of solidarity that unites its members; an injury to the honour of one affects the group as a whole and demands vengeance.

Villages high on mountains

The concern of Berber society to preserve its individuality is evident in the choice of habitat. Villages, which are fortified, are generally perched high up on mountain crests. Small in size, such villages are composed of the dwellings, a mosque, a threshing floor, and a place for the assembly of the elders, or *jemaa*, which governs the affairs of each community. Families live, each unit apart, in separate rooms that form a square around a closed interior courtyard.

Despite the fundamental homogeneity of Berber society, there is a considerable diversity in different mountain localities. The Shleuh (Chleuch) of the Haut Atlas in Morocco inhabit the river valleys that cut down deeply into the massif. Their villages, with populations of several hundred inhabitants in each, are often located at an altitude of more than 6,500 feet (2,000 metres). They consist of terraced houses, crowded one against the other. They are often dominated by a communal fortified threshing floor, or else are grouped around the threshing floor-plus-dwelling (*tirhamt*) of the most powerful family. The mountain slopes in the vicinity are divided up for pasturage and cultivation. In some fields dry (i.e., non-irrigated) farming is practiced for growing cereals. Land that is irrigated by diverting water from oueds yields two crops a year—cereals in winter and vegetables in summer. The Shleuh use manure from their cattle to fertilize the soil. Oxen and goats penned together on the ground floor of dwellings graze on stubble and on fallow lands around the villages. Sheep follow a pattern of transhumance (seasonal migration), grazing on low-lying land in winter and on the uplands in summer.

During the period of the French Protectorate in Morocco, profound changes occurred, transforming the way of life of the Moyen Atlas populations. The dominant pattern of transhumance gave way to the practice of agriculture. The winter descent to the *azarhar* (plains) pasture is practically a thing of the past, since the land is now under cultivation. The ascent to high pastures in summer, however, still continues. Stock farming is increasingly practiced in one place. Lumbering also brings in an appreciable income.

Where the mountain and the plain meet, the *dir* lands offer rich potentialities, thanks to a light soil and abundant water. Grouped together in large villages, the *diara* populations (*i.e.*, populations who live on the slope of the *dirs*) constitute prosperous agricultural communities.

The Rif of Morocco and the Kabyle of Algeria have

many points of resemblance. Both Berber tribes, they inhabit the same wet mountain slopes covered with oak forests, are similarly attached to a barren soil, and are both inclined to isolationism. In contrast to the way of life of the Berbers of the Haut Atlas and the Moyen Atlas, stock raising plays only a secondary role in their village life; they are not so much agriculturalists as arboriculturists, although they grow a little sorgo (a sorghum used for fodder), and their wives grow vegetables in small gardens adjoining their houses. It is, however, the fig and olive trees covering the mountain slopes they inhabit that constitute their principal resources. The Kabyles are also skilled craftsmen, working with wood, silver, and wool. Until recently they were also peddlers, selling carpets and jewelry to the people of the plains.

The Aurès Massif, standing alone in northeastern Algeria, is perhaps the most backward mountain region in the Maghrib. The Shawia (Chaouia) populations who inhabit it follow a seminomadic style of life, which is partly agricultural and partly pastoral. They live in terraced stone villages in which the houses are built in tiers, one above the other, the whole being dominated by a *guelaa*, or fortified granary. When winter comes, the inhabitants of the high valleys lead their flocks to the lowlands surrounding the massif, where they pitch tents or live in caves. Returning to the uplands in summer, they irrigate the land to grow sorghum (a cereal grass) and vegetables and maintain apricot and apple orchards, while shepherds take the sheep to pasture on the hilltops.

Despite precarious living conditions, the Atlas Mountains are densely populated—overpopulated, even, in certain localities. In the Tizi-Ouzou region of the Grande Kabylie, for example, densities reach about 780 persons per square mile. Emigration is a necessity: the mountain regions have become a human reservoir upon which the Maghribian countries draw to obtain the labour force needed for development. Commercial agriculture attracts large numbers of farm workers to the plains either on a seasonal or a permanent basis. The Mitidja Plain, for example, has been settled by Kabyles. In Morocco, the Shleuh of the Haut Atlas provide labour for the phosphate mines.

Urban growth has served to increase the volume of the migratory stream that flows down from the mountains; the cities of Algiers, Constantine, Oran, and Casablanca are to a great extent peopled by mountain folk. The shantytowns of Algiers contain numerous Kabyles, and those of Casablanca many Shleuh. Many of these urban immigrants find employment as labourers, while others become shopkeepers.

In Algeria the insecurity that became general in most mountain districts during the Nationalist uprising that preceded independence led to the departure of large numbers of people. The exodus from the mountains continued after independence, with many mountain dwellers moving into the plains to occupy houses abandoned by departing Europeans. Rural and urban activities, however, still did not provide employment for all, for many emigrants, mostly from Algeria, sought work in France. To a considerable extent the mountain populations subsist on money sent back by these migratory workers.

Transport and communications. The Atlas Mountains have their own internal system of communications. Villages are linked by paths that, avoiding the valley bottoms, follow the crest lines of the hills. Travel is on foot or by mule, although local bus transport is increasing.

The massifs constitute an obstacle to traffic; roads and railroads traverse them by means of tunnels and viaducts, which are costly to build. Traffic between Algiers and Constantine is obliged to cross the Kabylie Massif; the route runs through the Oued Isser gorges and crosses the Biban mountains at the Portes de Fer Pass. The Chiffa Gorge cuts across the route between Blida and Médéa.

The relative impenetrability of the mountains explains why they have been avoided by the main transportation routes, and why, consequently, they constitute strongholds of ancient traditionalism. Obstacles to communication should not, however, be exaggerated; the mountains also offer many natural connecting links, or passes, that facilitate movement. Such topographic accidents localize communication routes: between the desert and the plains the nomads use synclinal corridors (*i.e.*, corridors formed by folds in the rocks in which the strata dip inward from both sides toward the centre) that separate the ridges of the Atlas Saharien range. The Biskra Gap, situated between the Monts des Ouled-Naïl and the Aurès, provides a natural conduit for traffic between Constantine and Touggourt. Between Algeria and Morocco both the road and the railroad cross the Atlas at the Taza Pass, which breaks the continuity of the mountain system between er-Rif and the Moyen Atlas. Passes are natural routes across the mountain barriers, and thus strategic points. The focal point of communication in the Grande Kabylie, for example, is Tizi-Ouzou, at the Genêt Pass, which has become in effect the capital of the massif. To surmount the obstacle formed by the Ouarsenis, situated between Chelif Plain and the Sersou Plateaus, one must pass by way of Teniet el-Haad. The passes of the Moroccan Haut Atlas have also played a decisive role in the history of relations between Morocco and the vast region known as the Sudan to the south; the ancient caravan route from Marrakech to the Drâa Valley used the Tizi-n-Test Pass, which thus became of great commercial importance.

The mountain passes

Economic resources. It might be thought that the Atlas Mountains have no part to play in the future development of the Maghribian countries, but this is not so. The development of the resources of the Atlas will open new perspectives. The mountain massifs constitute catchment areas with considerable potentialities. The construction of reservoir dams not only would permit the storage of enormous amounts of water for irrigating the plains but would also generate much hydroelectric energy. In Morocco efforts are being made to exploit the potentialities of the mountain oueds. In addition to the dams across the Oued el-Abid (a wadi) on the northern slope of the Haut Atlas, dams on the southern face have been constructed across the Oued Drâa and Ziz watercourses. In Algeria, the Kabylie regions have been developed. Among the major projects completed are the Iril-Emda Dam and the underground power station at Darguinah on the Agrioun Wadi. The Namoussa Dam, which will irrigate the Annaba Plain, is now under construction.

The geological formations of the Atlas are rich in minerals. The Moroccan Haut Atlas in particular contains important deposits. Among these are the lead and zinc deposits associated with Jebel Erdouz, the copper and lead deposits now being mined at Goundafa, and the manganese deposits of Imini and of Tiouine, the output of which is transported to Marrakech by overhead cable cars. In Algeria, iron ore is extracted from the Seba Chioukh Mountains, and from Zaccar, Ouenza, and Bou Khadra, while phosphate is mined at Djebel Onk. In Tunisia, the Haut Tell mountains produce phosphate at al-qal'ah al-Jardā', iron ore from Jabal Djerissa, and lead from Sāqiyat Sīdī Yūsuf. These raw materials are not processed on the spot but are either exported or else are processed in the coastal towns. The iron ore from Ouenza, for example, supplies the iron-smelting industry of Annaba.

Among forest products, cork is more important than timber; production is centred in the Kabylies, notably on the Collo Massif.

The tourist industry is also being developed, particularly in the Haut Atlas region of Morocco. In the Moyen Atlas, long snow-covered slopes suitable for winter sports are located in the vicinity of major towns. In Algeria, the establishment of industry in mountain regions is being encouraged, so that employment for the mountain dwellers will be available on the spot. A textile factory has been constructed at Drâa ben Khedda in the Grande Kabylie, while at Constantine, the principal city of the mountain regions, several industries have been established. Despite these efforts, however, contrasts between the life styles of the mountains and those of the plains and cities of the Maghrib have by no means diminished, nor are they soon likely to do so.

BIBLIOGRAPHY. CHARLES E. DE FOUCAULD, *Reconnaissance au Maroc 1883–84* (1888 and 1934), an authoritative work

containing careful surveys, maps, and sketches; J. THOMSON, *Travels in the Atlas and Southern Morocco* (1889), includes important botanical and geological data; W.B. HARRIS, *Tafilet: The Narrative of a Journey of Exploration in the Atlas Mountains and the Oases of the North-West Sahara* (1895); G.H. BOUSQUET, *Les Berbères*, 3rd ed. (1967), describes the people, their institutions, customs, and religion; J. BERQUE, *Structures sociales du Haut-Atlas* (1955). For the Atlas Tellien, see X. YACONO, *La Colonisation des plaines du Chélif de Lavigerie confluent de la Mina*, 2 vol. (1955–56); and *Les Bureaux arabes et l'évolution des genres de vie indigènes dans l'ouest du Tell algérois* (1953), which covers Dahra, Chélif, Ouarsenis, and Sersou. For the Atlas Saharien, see R. FURON, *Le Sahara* (1957), for geological data; and R. CAPOT-REY, *Le Sahara français* (1953).

(H.Is.)

Atmosphere

The atmosphere that surrounds the Earth and is commonly called the air consists of layers of gases and mixtures of gases, as well as water vapour and solid and liquid particles. The mean pressure exerted by a vertical column of the atmosphere at sea level equals that of a column of mercury 760 millimetres (29.92 inches) in height. Such a column of air exerts a force, called atmospheric pressure, expressed as 1.033 kilograms per square centimetre (14.7 pounds per square inch).

Higher in the atmosphere the pressure decreases, and at an altitude of six kilometres (four miles) it is only one-half that at sea level. If this decrease in the pressure within the field of gravity were uniform throughout the depth of the atmosphere, the weight of a column of air at about 60 kilometres (40 miles) would be reduced to a thousandth (10^{-3}) that at sea level, to a thousandth of a millionth (10^{-9}) at about 180 kilometres (110 miles), and to 10^{-21} at about 420 kilometres (260 miles). The last value (10^{-21}) represents a density less than that of interstellar space. It is known, however, that the atmosphere of the Earth extends well beyond 400 kilometres (250 miles). Visual observations—of the auroras, for example—reveal the presence of luminous rays up to altitudes as great as 1,000 kilometres (600 miles). Furthermore, shortwave radio transmission has proved that the medium is sufficiently dense at an altitude of several hundred kilometres to produce enough electric charges (electrons) to reflect radio waves. Rocket probes and especially the drag encountered by artificial satellites at altitudes of several thousand kilometres have demonstrated that the terrestrial atmosphere extends to a very great distance. This extension to high altitudes occurs because of the occurrence of constituents of low mass and because of pressure and temperature conditions.

This article treats the physical and chemical properties of the several regions of the atmosphere. For further information on the interaction of the Earth's magnetic field and the atmosphere, see EARTH, MAGNETIC FIELD OF; IONOSPHERE; VAN ALLEN RADIATION BELTS; and AURORAS. See ATMOSPHERE, DEVELOPMENT OF for treatment of the origin and evolution of the present atmosphere; and WINDS AND STORMS; OCEANS AND SEAS; and CLIMATE for the role of the atmosphere at levels near the Earth and its interaction with the hydrosphere.

REGIONS OF THE ATMOSPHERE

Vertical temperature distribution

The principal regions of the atmosphere, each of which is characterized by the pattern of vertical distribution of temperatures, are the troposphere, the stratosphere, the mesosphere, the thermosphere, and the exosphere (Figure 1). In meteorological research it has long been customary to deal with two regions: the troposphere and the stratosphere.

The lower atmosphere. *The troposphere.* The region of the atmosphere in contact with the Earth's surface, the troposphere, is the realm of the clouds, rain, snow, etc., and is characterized in general by a decrease of temperature with increasing altitude. The upper limit of the troposphere, known as the tropopause, is at an altitude of about 17 kilometres (11 miles) at the Equator and only six to eight kilometres (four to five miles) at the poles. In the middle latitudes, the altitude of this limit varies with the atmospheric conditions. In high-pressure areas it is about 13 kilometres (eight miles), and it may be below seven kilometres (four miles) under low-pressure conditions.

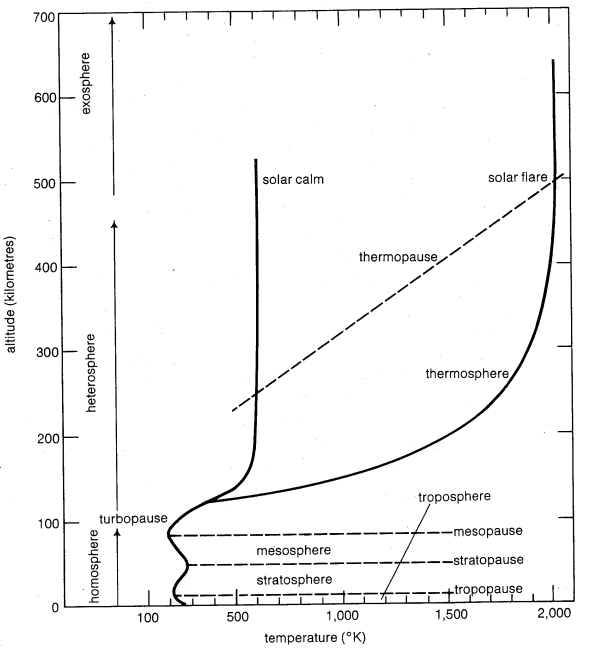


Figure 1: Regions of the atmosphere and their transitional zones.

The stratosphere and mesosphere. The second region of the atmosphere that has been closely studied, the stratosphere, is still within the reach of meteorological probes. In its lower section there is a slow, constant increase in temperature with altitude; the temperature rise becomes more rapid with increasing altitude, attaining a maximum of about 270° K (the Kelvin scale is an absolute Celsius scale; *i.e.*, 0° K equals −273° C [−460° F], which is absolute zero) at approximately 50 kilometres (30 miles), the upper limit of the stratosphere, which is known as the stratopause. The temperature of the stratopause varies less than that of the tropopause, which may decrease from 220° K at the pole to about 190° K at the Equator. There is some variation of the temperature of the stratopause with latitude and with the seasons. The troposphere and the stratosphere are clearly separated; the exchange between tropospheric air and stratospheric air requires several months, if not several years, indicating a difference in the circulation of the air within the two zones.

The middle atmosphere

Above 50 kilometres is the mesosphere or middle region, characterized by a rapid decrease of temperature to a minimum at about 85 kilometres (55 miles); this can be below 160° K in the summer at high latitudes. But, like the troposphere, the mesosphere is subject to strong seasonal variations of temperature at high latitudes. The prevailing level of the minimum temperature indicates another division in the atmosphere, known as the mesopause.

The upper atmosphere. Beyond the mesopause is an atmospheric region different in character from that of the lower regions. The first stratum of this higher region is the thermosphere, characterized by a continuous increase of temperature up to 500° K in the course of a night during minimum solar activity and to above 1,750° K in the course of a day during maximum solar activity. The altitude at which this increase of temperature ceases is the thermopause, which is at the base of an isothermal (constant-temperature) region that would extend into interplanetary space if the normal properties of a gas in hydrostatic (equal-sided) equilibrium continued to apply. In reality, the frequency of collisions between gas atoms becomes so low above a certain level, called the critical level, or baropause, or exobase, that the gas atoms can be considered to have their free-space trajectories. In this

highest region, the exosphere (outersphere), the study of the physical properties becomes the study of the movement of particles subject to gravity and capable of escaping from the atmosphere. In the exosphere, temperature no longer has the customary meaning.

Atmospheric mixtures. Of the regions described above in terms of their vertical temperature distribution, the first three—troposphere, stratosphere, and mesosphere—have the same general composition; that is, the tropospheric mixture of molecular nitrogen and oxygen is maintained in the stratosphere and the mesosphere. In combination the three regions comprise the homosphere. At the higher altitude of 100 kilometres (60 miles), in the thermosphere, molecular oxygen is strongly dissociated, and atomic oxygen becomes an important component of the atmosphere. The latter region is known as the heterosphere and is characterized also by the presence of light atoms, such as helium and hydrogen, at its higher altitudes. A relative increase of these light elements in comparison with heavier elements such as nitrogen and oxygen is a result of the absence of sufficient mixing of the air by turbulence; instead, its composition is dominated by gas diffusion in the field of gravity. The region of 100 to 110 kilometres (60 to 70 miles) is one of transition, called the turbopause; above it, an element can exhibit its own natural vertical distribution, rather than that characteristic of the general atmosphere.

Extraterrestrial effects. Above a certain level the atmosphere is subject to ultraviolet radiation, to X-rays, and to solar particles. These cause the production of electrically charged particles—that is, ions and electrons from various kinds of atoms and molecules—in the ionosphere (*q.v.*), which extends from the mesosphere to the outermost limits of the atmosphere. But the charged particles are affected by the magnetic field of the Earth and consequently behave differently from the neutral particles in the air. In regions where the pressure is high enough, as in the mesosphere and the greater part of the thermosphere, ionospheric conditions are dominated by the preponderant neutral atmosphere. But, when the numerical ratio of charged particles to neutral particles is no longer negligible, the ionosphere is characterized by conditions in which account must be taken of the electric field connecting the positively and negatively charged particles. The region in which charged particles have energies greater than those corresponding to thermal velocities and move among the lines of force of the terrestrial magnetic field is the magnetosphere. It is a vast region primarily related to the interplanetary space in which the charged particles coming from the Sun are propagated. Another important influence is an auroral zone called the auroral oval (because of its form), characterized by the occurrence of polar auroras resulting from an influx of charged particles. Within the auroral oval the solar protons produce a particular type of ionization. Thus, there is also a geographic division of the atmosphere, resulting from the presence of the terrestrial magnetic field.

The region of the ionosphere

Composition of the atmosphere. The air near the Earth's surface has a well-defined chemical composition, consisting of molecular nitrogen, N_2 (78.1 percent by volume), molecular oxygen, O_2 (21 percent), argon (0.9 percent), and a small amount of carbon dioxide, CO_2 (0.03 percent). Also contained in the atmosphere (see Figure 2) are small, variable amounts of water vapour, H_2O, and trace quantities of methane, CH_4, nitrous oxide, N_2O, carbon monoxide, CO, hydrogen, H_2, and ozone, O_3, and of helium, neon, krypton, and xenon.

Where there is sufficient mixing by turbulence, the air consists essentially of molecular nitrogen and oxygen. Only at the altitude where molecular diffusion outweighs eddy diffusion does there begin the relative impoverishment of O_2 and N_2 molecules in comparison with atomic oxygen, O (Figure 2). The mass of an oxygen atom is 16 in relative units (*i.e.*, as compared to the mass of the atom of the isotope carbon-12, which is taken as the standard unit, a mass of 12), whereas that of molecular oxygen is 32 and that of molecular nitrogen 28. The mean mass of the air of the homosphere, 29, is gradually altered above an altitude of about 100 kilometres (60 miles) to that resulting from an increasing proportion of atomic oxygen. In addition, the elements that are rare in the homosphere (for example, helium, $\frac{5}{1,000,000}$ of the air at sea level, and hydrogen, $\frac{1}{1,000,000}$ at sea level) become relatively more abundant at higher altitude. Because of their low masses (helium 4, hydrogen 1), the concentration of these atoms decreases much more slowly than that of oxygen and nitrogen in the heterosphere, where each element is independently subject to the field of gravity. Ascending into the heterosphere, the belts of molecular nitrogen and oxygen, of atomic oxygen, and of helium are encountered in that order. Finally, at its extreme limits, the atmosphere is composed of atoms of hydrogen.

Atomic mass and distribution

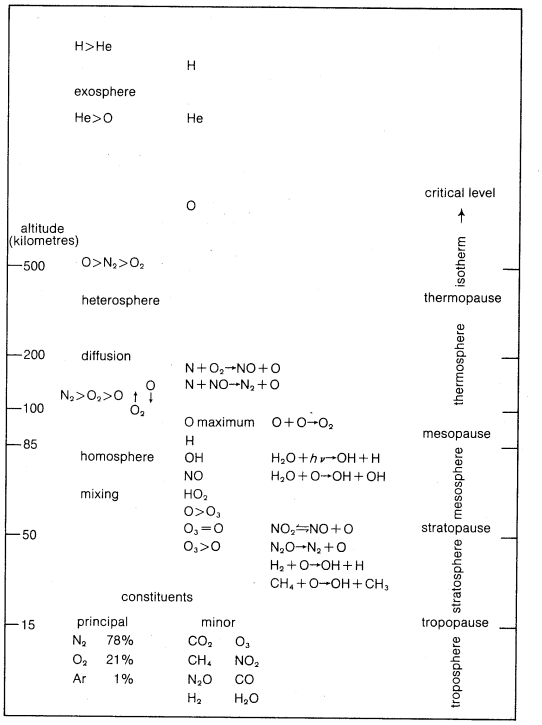


Figure 2: Distribution of principal and minor constituents of the homosphere, heterosphere, and exosphere.

THE HOMOSPHERE

It is not difficult to describe the homosphere in terms of only the concentration of its principal elements, molecular nitrogen and oxygen; however, the phenomena that occur under the influence of solar and cosmic radiation must be taken into account as well. Although most of the molecules of nitrogen and oxygen are not affected by radiations originating outside the atmosphere, the absorption of various radiations does lead to certain ionization and dissociation effects.

Galactic cosmic radiation produces a permanent ionization in the whole of the stratosphere and in the mesosphere. In the homosphere, however, ionization is strictly a secondary phenomenon; dissociation under the influence of solar ultraviolet radiation is more important.

The effects of ultraviolet radiation. Molecular oxygen, O_2, is photodissociated under the influence of radiation of wavelengths less than 2400 Å (angstroms; the angstrom is a unit of length equal to 10^{-8} centimetre) into atoms of oxygen, O. This process is the foundation of all photochemical reactions in the homosphere. Production of atoms of oxygen, O, in the stratosphere and the mesosphere leads immediately to the production of molecules of ozone, O_3, as is shown in the equation

$$O_2 + O + M \rightarrow O_3 + M,$$

in which a triple collision between a molecule of oxygen O_2, an atom of oxygen O, and a third particle, M, which may be a molecule of oxygen or of nitrogen, results in formation of a molecule of ozone, O_3.

In the homosphere, the ozone, O_3, and molecular oxy-

gen, O_2, molecules absorb all the solar ultraviolet radiation from 3000 to 1800 angstroms (ultraviolet ranges in wavelength from 100 to 3900 angstroms). Ozone absorbs primarily the radiation between 3000 and 2000 angstroms, and molecular oxygen that below 2000 angstroms. It follows that the absorption of ultraviolet radiation by ozone is the basis of the increase of temperature with altitude in the stratosphere; the ultraviolet heating is sufficient to develop a maximum temperature at the stratopause. Conversely, absorption by oxygen is very weak in the mesosphere, with the result that temperature decreases with altitude to the very low minimum found at the mesopause.

Dissociation of ozone. The presence of ozone in the stratosphere and mesosphere sets in motion a series of chemical processes. Under the influence of solar ultraviolet radiation of wavelengths below 3300 angstroms, ozone is photodissociated into molecules and atoms of oxygen in excited states. Production of an excited oxygen atom is important because it is a powerful oxidizer (*i.e.*, capable of producing chemical change by destroying normal molecules such as water vapour, methane, molecular hydrogen, and nitrous oxide).

When an excited oxygen atom is in the presence of, for example, stratospheric water vapour, H_2O, an immediate oxidation takes place, producing two hydroxyl radicals, OH (a radical is a group of atoms that occurs in the molecules of many compounds), expressed by the equation

$$O + H_2O \rightarrow OH + OH.$$

Formation of hydrogen compounds. With the dissociation of the water vapour, many compounds of hydrogen can be formed; for example, the hydroperoxy radical, HO_2, and hydrogen peroxide, H_2O_2. Furthermore, because the oxides of nitrogen NO and NO_2 are present, such compounds as nitric acid, HNO_3, can be formed. Analysis of these reactions shows that within the homosphere the existence of a mild degree of photodissociation of oxygen results in a remarkable series of reactions, which are important to the nature of the stratosphere and of the mesosphere. Accordingly, within the homosphere must be introduced the concept of the chemosphere, within which such chemical reactions occur.

Mechanisms of air mixing. One of the essential characteristics of the homosphere is that the turbulence of the air is always sufficient to maintain the same proportions of molecular nitrogen and oxygen in the mixture. Although molecular diffusion—that is, the tendency of gases of different molecular weights to separate—exists in the homosphere, it is always counterbalanced by eddy diffusion—the tendency to perfect mixture, in the case of the principal gases. But in the case, for example, of methane and of molecular hydrogen, which are oxidized as is water vapour, their removal from the air cannot be compensated by an influx of molecules from the troposphere. Molecular diffusion and turbulent diffusion are not sufficient to compensate for the loss resulting from oxidation. At the level of the stratopause, methane begins to disappear from the atmosphere. Water vapour is still present, however, because of its rapid recombination in the stratosphere and in the mesosphere.

The examples given illustrate the importance of the mechanisms of horizontal and vertical transport of the air in the stratosphere and in the mesophere. Analysis of the homosphere requires that the dynamic aspect be taken fully into account; further, the presence of helium and hydrogen in the heterosphere depends upon the conditions of transport in the homosphere. The helium produced in the Earth's crust enters the atmosphere and makes its way through the homosphere by the process of molecular diffusion until it reaches the thermosphere, where it plays an important role that will be discussed elsewhere in this article; atomic hydrogen is the product of the oxidation of hydrogen compounds, which occurs as determined by the properties of the homosphere.

THE HETEROSPHERE

At the beginning of the lower part of the thermosphere the change of composition and the increase in temperature are evident. Atomic oxygen becomes a factor in determining the mean molecular weight at an altitude of about 100 kilometres (60 miles). In addition to the atomic weight, 32, of the oxygen molecule, O_2, and 28, of the nitrogen molecule, N_2, the weight, 16, of atomic oxygen, O, must be considered. The decrease of the mean molecular weight essentially results from the photo-dissociation of the oxygen molecule under the influence of solar ultraviolet radiation of wavelengths below 1750 angstroms. Dissociation also occurs between 2000 and 2400 angstroms.

Oxygen dissociation in the lower thermosphere. An examination of thermosphere conditions reveals that equilibrium conditions cannot exist at each level between dissociation and reconstitution of the oxygen molecule. In other words, equality between the number of reactions that result in dissociation of the oxygen molecule and the number that result in recombination of the oxygen atoms is not possible under the conditions imposed by the altitude. At 100 kilometres the molecule can exist for more than ten days without being photo-dissociated, and accordingly its vertical distribution depends much more on the transport conditions than upon conditions of formation or destruction.

Variance in equilibrium conditions

There is, in fact, a departure from the conditions of photo-dissociation equilibrium and an approach to conditions of diffusion equilibrium. As a result, the following mechanisms are observed in the upper thermosphere: (1) when a molecule of oxygen is photo-dissociated by solar radiation, it yields two atoms of oxygen which do not recombine at the altitude at which they were produced, but are subject to a diffusion current descending to the altitude at which the atmosphere is dense enough to cause their disappearance; (2) the dissociated molecule is replaced by a molecule arriving from below, because O_2 survives long enough in the field of solar radiation to be transported upward by molecular diffusion.

In short, the dissociation state of the oxygen in the lower thermosphere is not determined by the equilibrium of photo-dissociation; instead, heavy O_2 molecules are carried upward, whereas the light atoms of oxygen must descend before they can recombine.

Effects of vertical transport. In the region in which the oxygen is subject to conditions of vertical transport, the transport processes also apply to other constituents of the atmosphere. At the altitude between 100 and 120 kilometres, eddy diffusion begins to lose its superiority over molecular diffusion. The transition results from the fact that the coefficient of molecular diffusion is inversely proportional to the density. Between sea level and 100 kilometres the density has diminished by a factor of at least 1,000,000,000; accordingly, molecular diffusion has increased in intensity by the same factor. Eddy diffusion having, on the other hand, gained no special advantage at 100 kilometres, it follows that molecular diffusion has become predominant at 120 kilometres. As a result, there is time for all of the atmospheric constituents that are not exceptionally reactive to become distributed in the gravity field according to their individual masses.

Atmospheric distribution by mass

Distribution of oxygen and nitrogen. Nitrogen, N_2, which is the most abundant molecule at 100 kilometres (of the order of 10^{13} molecules per cubic centimetre) and has a mass of 28, follows practically the vertical distribution of normal air, of which the mean molecular weight is 29. Molecular oxygen, O_2, with a mass of 32, decreases more rapidly with altitude than does molecular nitrogen (Figure 3).

Atomic oxygen, O, of mass 16, shows a relative increase with altitude in comparison with the oxygen and nitrogen molecules; it becomes more abundant than molecular oxygen, O_2, above 125 kilometres and more abundant than nitrogen, N_2, above 250 kilometres. As a result, the heterosphere is characterized by a thick layer of atomic oxygen extending for several hundreds of kilometres.

Distribution of helium and hydrogen. When subject to the diffusion process, helium and atomic hydrogen, whose proportions at 100 kilometres are of the order of $1/1{,}000{,}000$, can become the predominant elements because of their low masses (helium, 4; hydrogen, 1). It can be

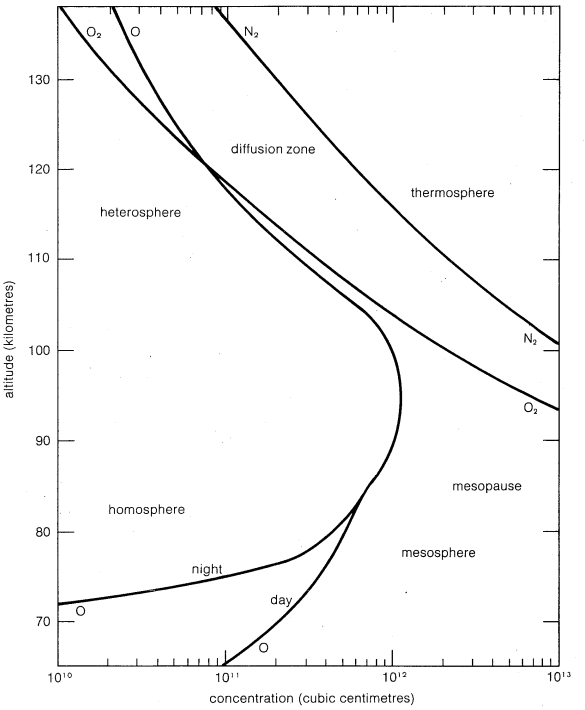


Figure 3: Concentration of oxygen and nitrogen in the atmosphere with respect to altitude.

shown that a belt of helium exists above the layer of atomic oxygen and also that atomic hydrogen forms the uppermost levels (corona) of the terrestrial atmosphere. The vertical distributions of atomic hydrogen, helium, atomic oxygen, molecular nitrogen, and molecular oxygen are presented in Figure 4, in which the density at a relatively low temperature of 600° K is given for each constituent.

Variations in atmospheric densities. Review of the data obtained from the drag on artificial satellites makes it evident that the density of the heterosphere is subject to considerable variations (Figure 5). Whereas at the maximum of solar activity that occurred in 1958 the density was extraordinarily high, it came down to a very low value during the 1964 minimum of solar activity. There is, thus, a close correlation with solar activity. It is the ultraviolet radiation, which penetrates into the heterosphere and is absorbed there, that exhibits variations closely linked to the Sun's activity. To the 11-year cycle of solar activity must be added the variation corresponding to the rotation of the Sun. Finally, it must be emphasized that above 200 kilometres there is a remarkable variation of density between day and night.

The rate of ultraviolet radiation

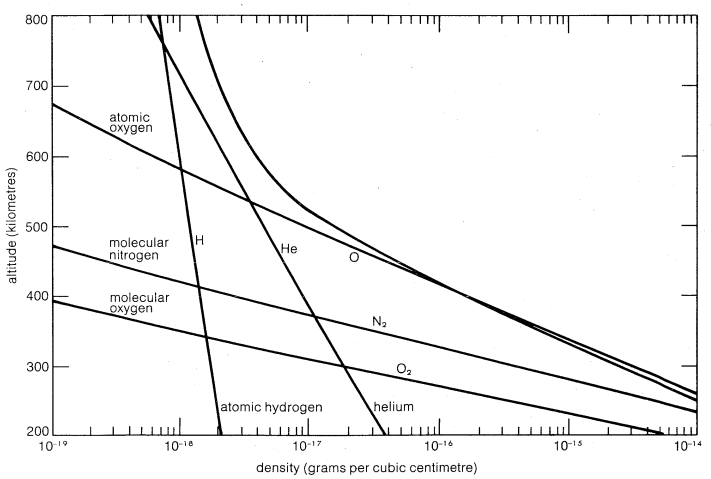


Figure 4: Density of atmospheric constituents in the heterosphere at a temperature of 600° K.

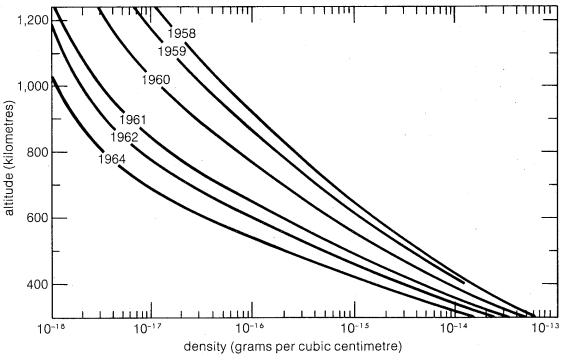


Figure 5: Relationship between mean atmospheric density and solar activity in the heterosphere.

Temperature fluctuations. To interpret the density variations in the thermosphere, it is important to consider that the thermospheric gas is subjected to strong variations of temperature. The thermal balance is evaluated using the assumption that an ultraviolet heating effect and a rapid recooling work against each other. Thermal conditions in the heterosphere evidently differ from those of the homosphere. In the lowest region of the homosphere, the troposphere, which is in contact with the ground, the source of heating is essentially the ground, which has absorbed the Sun's rays, and convection is an adequate means of transporting the heat. In the stratosphere and mesosphere, convection and infrared radiation from one direction meet ultraviolet heating from the other. But in the thermosphere the third means of heat transport—conduction—decisively outweighs the other means, radiation and convection. For that reason the thermal balance is such that very different conditions appear by day and by night.

During the day, the cooling by conduction is compensated by ultraviolet heating, whereas during the night the heat loss by vertical conduction is not compensated by horizontal transport. As a result, there is a very large daily variation of the temperature of the thermopause, which is manifest to the observer as a variation of the atmospheric density.

Cycles of solar activity. Furthermore, the fact that periods of 27 days can be detected in the variations of atmospheric density must be interpreted in terms of the rotation of the Sun. Ultraviolet emission does take place when the Sun is perfectly calm over the whole of its disk; but the brilliant regions visible as sunspots emit a greater ultraviolet radiation. An increase of solar activity in the hemisphere of the Sun that is visible from the Earth is an index of increased ultraviolet heating. Accordingly, the thermal balance in the thermosphere varies according to the amount of ultraviolet heating available to counter the conductive cooling.

Taking account of the variation in ultraviolet radiation during a cycle of solar activity, it is possible to predict the long-term variations of the thermospheric temperature. In Figure 6 the maximum variation of daytime temperature is shown; the mean temperature varies from a low of 750° K, with minimum solar activity, to a high of about 2,000° K.

Because the composition of the heterosphere varies as a function of temperature, the vertical distribution of the atmospheric constituents is variable (Figure 7). The proportion of helium is greatest at the highest temperatures and the proportion of hydrogen is greatest when the temperature is lowest.

The escape of helium and hydrogen. It must be remembered that the aeronomic problems regarding helium and atomic hydrogen are essentially different from those associated with the other elements. The continual introduction of helium into the atmosphere—at an average rate of about 1,000,000 atoms per square centimetre per second (as a result of its formation in the Earth's crust by the disintegration of thorium and uranium)—would lead in the course of some millions of years to a doubling of the total quantity of helium existing in the atmosphere.

Sources of airborne helium

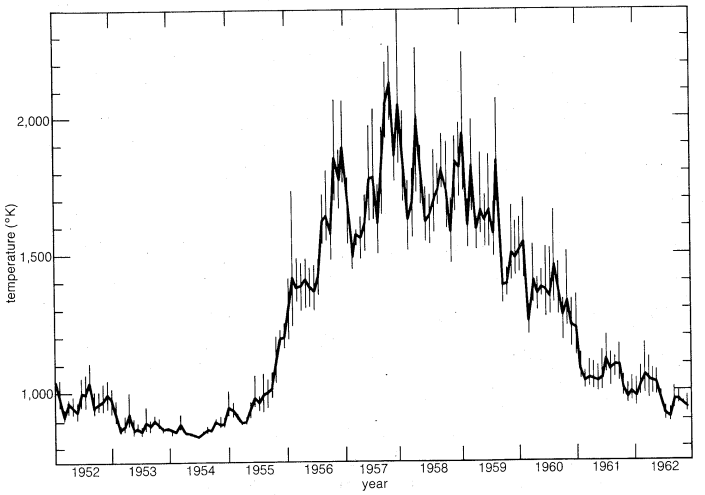


Figure 6: Mean temperatures at the thermopause for 27-day periods during 1952–1962 (see text).

Since the formation of the Earth about 4,600,000,000 years ago, an amount of helium equal to that introduced from the above sources must have escaped to avoid accumulation. Such escape can occur only at the top of the atmosphere by a transport process taking place in the homosphere and the heterosphere. The equations of diffusion reveal that the escape of 1,000,000 atoms of helium per square centimetre per second is perfectly possible, thus maintaining a helium distribution close to the equilibrium of diffusion in the heterosphere.

In the case of hydrogen, of which some 100,000,000 atoms escape, the limit of what diffusion will support is reached. For this reason, the vertical distribution of hydrogen is not that of a diffusion equilibrium but depends on a transport state in which the number of atoms moving toward the base of the heterosphere is equal to the number leaving it. It can be concluded that there are more atoms of hydrogen in the thermosphere at low temperature than at high temperature. The role of hydrogen in the thermosphere is therefore most evident at the minimum of solar activity. In the last analysis, the total mass of hydrogen atoms passing from the terrestrial atmosphere into interplanetary space is several hundred grams per second.

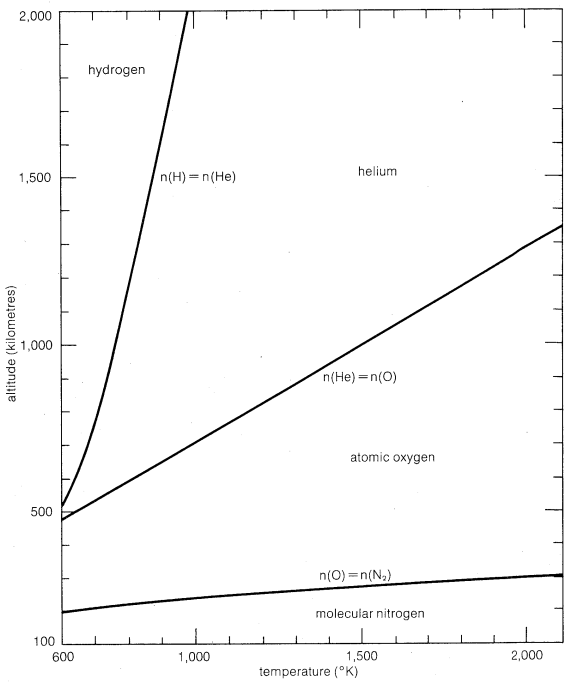


Figure 7: Atmospheric zones of common gases as a function of temperature and altitude.

It is possible to go further into the subject of the heterosphere with respect to perturbations such as the magnetic storms. It is clear that the contribution of energy due to certain particles can cause a general heating only if the energy is capable of penetrating across the lines of force of the terrestrial magnetic field. The heating can be local, as in the auroral oval, where there is direct precipitation of particles. It has been shown that the density of the thermospheric gas varies with the geomagnetic index, which can be explained by a variation of temperature. Such processes, however, require complicated explanations.

Ionization mechanisms. Because the aeronomic conditions of the ionization state depend simultaneously on the vertical distribution of the neutral elements nitrogen and oxygen and on the spectral distribution and the absorption of the solar radiation, it may be readily concluded that the amount of photo-ionization is closely related to the diurnal, seasonal, periodic, and 11-year variations in solar activity. In addition, account must be taken of the reactions occurring between the ions and the neutral particles and also of the processes that result in the disappearance of electrons and ions.

Ionization and solar activity

When an ionizing radiation causes the ejection of an electron from an atom or a particular molecule, a photoelectron appears, which possesses kinetic energy corresponding to the difference between the solar radiation energy and the energy required for the ionization. Such a photoelectron can give up part of its energy to ionize another atom if its energy is sufficiently high, as is the case with strongly energized electrons produced by X-rays. In general, any photoelectron loses part of its energy in exciting the most plentiful atoms and molecules; however, in the upper ionosphere, photoelectrons colliding particularly with the other electrons cause a general heating of the electronic gas.

The nitrogen ion. A photo-ion (atomic or molecular ion produced by photo-ionization) can collide with an electron, and it is possible that the two will reconstitute the neutral element. But it is more likely that the photo-ion will collide with an atom or molecule of another kind, exchange its charge, and return to the neutral state; for example, the nitrogen molecule—the element most plentiful below 200 kilometres—should furnish the most abundant ion. In reality, the photo-ion of molecular nitrogen very quickly comes into contact with a molecule of oxygen or with an atom of oxygen. The first case involves a transfer of charge; that is to say, the peripheral electron of the oxygen molecule, O_2, is transferred to the nitrogen ion, N_2^+, which reconstitutes the normal molecule, N_2, and forms the molecular ion, O_2^+: $N_2^+ + O_2 \rightarrow N_2 + O_2^+$.

The second case corresponds to a change of the atom and of the ion, in which the molecular nitrogen ion, N_2^+, colliding with the oxygen atom, O, dissociates to form an ion of nitric oxide, NO^+, and an atom of nitrogen, N: $N_2^+ + O \rightarrow NO^+ + N$. Thus, the second case results in an NO^+ ion that has been produced directly by the action of solar radiation. A molecule of nitric oxide appears in the heterosphere by way of the formation of the ion. In any case, the formation of the NO molecule in the thermosphere results in its descent toward the mesophere, where it is subject to the dissociation effect in that region.

Atomic-hydrogen ions. Although the hydrogen atom is photo-ionized by solar radiation, its ionization state is determined primarily by its reaction with atomic oxygen. There is an equilibrium of exchange of charges closely related to the neutral or ionized condition of the atmosphere. This equilibrium is written:

Equilibrium exchange of atomic charge

$$O^+ + H \leftrightarrows H^+ + O,$$

in which an atomic oxygen ion, O^+, and a neutral atom of hydrogen, H, react to produce a neutral atom of oxygen, O, and a hydrogen ion, H^+, and vice versa. This equilibrium results from the fact that the two atoms have nearly equal ionization energies, and, if the concentrations are great enough, they will pass readily from one state to the other in accordance with the proportions of the two constituents, H and O, present in the neutral

state. But ionization equilibrium conditions between atomic oxygen and hydrogen cannot be maintained at the highest altitudes. The mechanism of exchange of charge depends on the frequency of the collisions of the ions with neutral atoms, whose concentration decreases with increasing altitude. Consequently, when the exchanges are infrequent the phenomenon of diffusion of the ions in the gravity field appears. The hydrogen ion, H^+, which is only $\frac{1}{16}$ as heavy as the oxygen ion, O^+, will have a tendency to distribute itself along a line of force of the magnetic field in accord with its particular weight and to escape reaction with the neutral oxygen atom. As a result, atomic-hydrogen ions become much more numerous than atomic-oxygen ions.

THE EXOSPHERE

When it is stated that the temperature is about 273° K at sea level or at 50 kilometres altitude or that it is about 1,500° K at 500 kilometres, the same physical basis of determination is used, and it is incorrect to suppose that a change of concept is involved. The thermometer currently used in meteorology is not sufficiently sensitive to measure the exact temperature of the air at high altitudes. Even at sea level, many precautions must be taken in order to obtain an exact measurement of the air temperature.

Division of molecular energies

The temperature of gases. The kinetic theory of gases states that the molecules are characterized by very rapid motions, with a distribution of velocities such that a mean velocity corresponding to a specific temperature can be determined. When the collisions are very numerous, there is an equal division of the energy among the various kinds of molecules, which defines a unique temperature. To state that the temperature is of the order of 273°K at a certain altitude means that the kinetic energies of the molecules are identical, because there are enough collisions to assure equal division of the energy. Consequently, perfect thermometers (which could attain equilibrium immediately) would give the same indication. No such thermometers can be made, however, and, accordingly, indirect procedures are used, based on the determination of pressure and on the law of perfect gases.

Reactions between particles. When the heterosphere is analyzed in terms of its physical constitution, in which electrons, ions, and neutral atoms are simultaneously involved, it is necessary to go into the details of various types of reactions and to determine their exact nature. First of all, it must be remembered that the interaction between two electrons is much more marked than that occurring between an ion and an electron. Also, the interaction between an electron and a neutral atom is much weaker than that between an ion and an electron. All these interactions can take various forms, ranging from elastic collisions to exchanges of charge or of energy. It follows that the electrons can have their own temperature when there are so many electrons relative to neutral particles that the interaction between electrons predominates over other interactions. Moreover, when the time required for exchanges of energy in interactions between ions and neutral particles is relatively long, it is possible for the ions and the neutral atoms to be maintained at different temperatures.

Collision frequency and temperature differences

Distribution of energy exchange. The problem of the temperature of a neutral gas, ions, or electrons can be described as follows: a photoelectron ejected, under the action of solar radiation, from a neutral atom of oxygen possesses a certain kinetic energy distinctly greater than that of the electrons already present at the same altitude. In the lower thermosphere (below 120 kilometres), the photoelectron will collide with the relatively numerous neutral molecules, rapidly lose its excess energy to them, and reach the temperature of the neutral gas. In the middle thermosphere (above 150 kilometres), the photoelectron will come into contact primarily with the electrons already present, and its excess energy will serve initially to increase their total energy; the result will be an increase of the temperature of the electrons as a group relative to the temperature of the ions and to that of the neutral atoms. In the upper strata, high-energy photoelectrons will come initially into contact with the other electrons that undergo collisions with the ions. Again, the photoelectrons will cause a rise in the electron temperature and indirectly in that of the ions, but, just as in the neutral atmosphere, it must be taken into account that heat conduction exists within an electronic gas exhibiting a temperature gradient. Consequently, there will be a tendency to isothermy at the highest altitudes by reason of the rapid heat transport. Thus, it is evident how greatly the physical state of the neutral atoms, the ions, and the electrons in the upper atmosphere differs from that created in the laboratory.

Atom velocities at the critical level

The gradation of particle collisions. Beginning at the critical level, temperature loses its ordinary meaning, because the particles follow their free-space trajectory and practically do not undergo collisions. Because of this it must be kept in mind that, although for a given temperature the atoms have the same energy, they do not have the same velocity; the latter depends on their mass. The light atoms move with greater velocities than the heavy ones; for a given energy, hydrogen has double the velocity of helium, which in turn has twice the velocity of atomic oxygen. Consequently, in the exosphere a light atom such as hydrogen can escape from the terrestrial atmosphere because of its velocity, four times that of atomic oxygen, which remains permanently subject to the gravity field.

Assuming that all the elementary physical processes can operate without restriction in the heterosphere, it is then clear that the changes of structure and composition can be explained within the usual framework of a hydrostatic equilibrium. The collisions are always sufficiently numerous to provide a complete distribution of particle velocities as defined in the kinetic theory of gases. This is what is called a Maxwell distribution, leading to a mean kinetic energy of the particles and a clear definition of the temperature. With increasing altitude above the thermopause, however, the mean free path of the atoms gradually increases. Beyond a certain altitude this mean free path is so great that collisions between the atoms are too few to maintain the uniformity of the atmospheric gas. The particles—atoms, ions, electrons—can traverse long distances under the influence only of the field of gravity in the case of the neutral particles, and only of gravity and the magnetic and electric fields in the case of the charged particles.

Thus, the heterosphere, in which the collisions must be taken into account, is separated from the exosphere, where the collisions can be neglected, by a critical zone in which transition occurs from the point, at its base, where the number of collisions is significant to the zone above it, where the number of collisions is negligible.

Determination of the critical zone. The most satisfactory means of determining the critical zone is to search for a level at which the aeronomic conditions can be defined for each element, neutral or ionized. It is possible, for example, to study the mean free path of each element. To specify the probability that, for example, in one out of two instances a particle arriving at a particular level will be able to escape without undergoing a collision is also to specify the mean free path characterizing a critical level corresponding to the base of the exosphere.

Neutral and ionic exosphere differences

Computation shows that, for the neutral atoms, the critical level for temperatures ranging from 750° to 2,000° K is located at altitudes ranging from 400 to 800 kilometres (250 to 500 miles), respectively. In other words, the neutral exosphere, considered as the region in which the collisions between neutral atoms are so infrequent as to permit the escape of atoms whose velocity is sufficient (11 kilometres per second), begins at 400 kilometres with minimum solar activity and can reach 800 kilometres with maximum solar activity. The ionic exosphere begins at an altitude 2,000 or 3,000 kilometres (1,000 or 2,000 miles) higher than that of the neutral exosphere, because collisions between charged particles are always more numerous. The ionic exosphere differs in another respect from the neutral ionosphere: although the gravity field influences the trajectory of the charged particles as it does in the neutral exosphere, the charged particles also are

controlled by the geomagnetic field and by the electrostatic field between heavy ions and electrons.

Particle trajectories in the exosphere. In any case, a given volume of the exosphere can contain a certain number of particles having a priori the following trajectories (see Figure 8): (1) particles coming from the critical level and returning there; (2) particles coming neither from the critical level nor from interplanetary space; (3) particles coming from the critical level and proceeding to interplanetary space or vice versa; (4) particles coming from and returning to interplanetary space. These four groups represent all the possibilities arising from a Maxwell distribution of velocities; it can accordingly be said that at any altitude the density, ρ, of a given element normalized to unity is composed of four densities:

$$\rho = \rho_{\mathrm{I}} + \rho_{\mathrm{II}} + \rho_{\mathrm{III}} + \rho_{\mathrm{IV}} = 1.$$

Particles with ballistic trajectories. The first group (see Figure 8) consists of particles with ballistic trajectories: those that leave the critical level at a certain angle with a certain velocity and return to it after having attained their limiting altitudes in the exosphere. The kinetic energy at departure from the critical level is less than that required for escape from terrestrial attraction. This is the case of the neutral particle the trajectory of which is regulated by the gravity field; but a charged particle is constrained to follow the lines of force of the geomagnetic field. A completely closed line of force does not permit the escape of an ion or an electron, whatever its velocity.

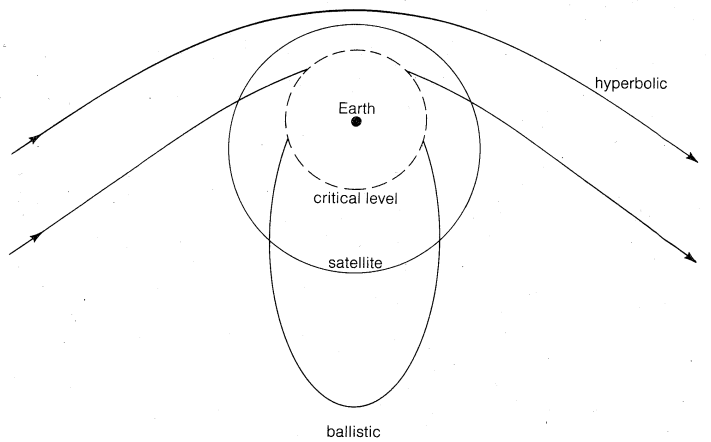


Figure 8: Trajectories of particles in the exosphere. In the vicinity of the critical level, 400 to 800 kilometres above the Earth, ballistic trajectories are most numerous, whereas at much greater distances, hyperbolic trajectories are evident.

Trapped particles. The second group of particles (Figure 8), not making contact with the critical level or with the exterior, corresponds to trapped particles. These are satellite neutral atoms or ions, subject to backward and forward movements along a line of force of the magnetic field. It can be shown that the conditions to be satisfied in order to maintain thermal particles in elliptic orbits are difficult to achieve. In general, the presence of satellite or entrapped particles in the exospheric distribution can be disregarded; the density ρ_{II} of exospheric particles following elliptical orbits in the presence of particles of density ρ_{I} following ballistic orbits is not significant.

Significant and insignificant exosphere constituents

Particles with escape velocities. The third group of particles must be subdivided according to whether they leave the critical level and escape to interplanetary space or arrive at the critical level from interplanetary space. In general, it is not necessary to consider particles with adequate velocities arriving from interplanetary space; the significant component of the third group consists of atoms at the critical level capable of attaining velocities above the escape velocity of 11 kilometres per second. This is the case with the hydrogen atom and sometimes the helium atom when the temperature of the heterosphere is sufficiently high. In the case of the ionic exosphere, escape is possible only where the particles follow open lines of force of the geomagnetic field into free space. In general, the escape of charged particles from the ionic exosphere is neglected, because a line of force of the geomagnetic field joins one hemisphere to the other except near the poles. Only through the channel of the polar cap, where the lines of force of the geomagnetic field are open toward the nocturnal magnetosphere, do charged particles have the possibility of escaping. Indeed, such is found to be the case for ions of hydrogen and helium, which escape easily from the polar ionic exosphere.

Extraterrestrial particles. Finally, it is possible to conceive of particles moving at high velocities that arrive from the exterior of the exosphere and return without touching the critical level. These would comprise the fourth group. Such particles are naturally neglected in the exospheric population, as they never attain sufficient concentration.

To sum up, the population of the neutral or the ionic exosphere is composed almost exclusively of particles coming from the critical level. In fact, the ballistic trajectories represent the practical conditions of vertical distribution for such elements as atomic oxygen. The concentration of neutral helium differs according to whether the temperature is high or low. In the case of neutral hydrogen, the trajectories representing permanent escape of this atom from the terrestrial atmosphere into interplanetary space are significant compared with the ballistic trajectories. As for the ions, it may equally be said that it is the elements that come to the critical level that constitute the population of the ionic exosphere. The oxygen ion, which is too heavy, does not escape, whereas the ions of hydrogen and helium escape into the polar exosphere. Thus, the atmosphere of the Earth, terminating in its neutral or ionic exosphere, is populated at its extreme limits by neutral and ionized atoms of hydrogen.

BIBLIOGRAPHY. During the last 20 years, available information on the atmosphere has been subject to a constant evolution. T.F. MALONE (ed.), *The Compendium of Meteorology* (1951), discusses all aspects prior to space observations. G.P. KUIPER (ed.), *The Earth As a Planet* (1954), shows the relations between the Earth's interior and the atmosphere. M. NICOLET, "The Properties and Constitution of the Upper Atmosphere," in *Physics of the Upper Atmosphere*, ed. by J.A. RATCLIFFE (1960), is an exposition of knowledge after the launching of the first artificial satellites. H.E. NEWELL, "The Upper Atmosphere Studied by Rockets and Satellites," *ibid.*, is a summary of the results obtained from rocket observations. H. FRIEDMAN, "The Sun's Ionizing Radiation," *ibid.*, indicates the first results obtained on the X-ray and ultraviolet observations of the Sun. D.R. BATES, "The Airglow," *ibid.*, presents a modern view on the emissions of the upper atmosphere at the beginning of space research. F.S. JOHNSON (ed.), *Satellite Environment Handbook* (1961), is a useful summary on the various phenomena of the upper atmosphere after the beginning of space observations. See also J.W. CHAMBERLAIN, *Physics of the Aurora and Airglow* (1961), a comprehensive text with extensive bibliography on the upper-atmosphere radiations; M. NICOLET, "The Structure of the Upper Atmosphere," in *Research in Geophysics*, ed. by H. ODISHAW (1964), an exposition of the general knowledge of the upper atmosphere after the International Geophysical Year (1957–58); R.A. CRAIG, *The Upper Atmosphere* (1965), meteorological view on the atmospheric problems; and *Physics of the Earth's Upper Atmosphere*, ed. by C.O. HINES *et al.* (1965), a modern review dealing with the ionospheric physics. *Solar-Terrestrial Physics: Terrestrial Aspects*, ed. by A.C. STICKLAND (1969), is a series that describes various aspects of the physics of the upper atmosphere after the Year of the Quiet Sun (1964–65).

(M.N.)

Atmosphere, Development of

The development of the Earth's atmosphere is a subject on which there is little direct evidence. Conjecture is limited only by geological consequences attributable to atmospheric evolution and by available theory. Among the important questions to which answers are sought are the origin of the gaseous components, the nature of the initial and early atmosphere, and the kinds of compositional changes that subsequently occurred in response to additions of some gases and losses of others. The answers to such questions must be sought within diverse areas of knowledge, because atmospheric evolution is related to

the general evolution of the planet Earth and, hence, to the development and interactions through time of the Earth's hydrosphere (waters), biosphere (life), and outer shell of rocks.

It would be helpful if a general theory of development of planetary atmospheres was available. The only planetary system known, however, is the solar system, and the question at once arises whether this system is a representative sample of the universe at large. The Sun appears to be a very ordinary star, and it seems likely that many millions of the stars in the visible universe are similar to it. Of still greater importance, it is likely that these visible stars have undergone or are undergoing similar evolutionary development.

If, then, our solar system does not depart drastically in attributes from the mean of all solar systems in the universe, it is probable that the existence, composition, and history of planetary atmospheres are related to the size, distance from parent star (or sun), and subsequent physical, chemical, and biological evolution of the planet concerned. Related variables include the temperature, electromagnetic and particle radiation, and previous history of the central sun; the presence or absence and the strength of a planetary magnetic field that would deflect the solar wind; the relative proportions of the elements; and the volatilities that different gases require to attain escape velocities from a planet of a given mass, size, and temperature. The origin of the chemical elements must also be considered. Their very regular periodic relationships signify that they have been put together from universally similar elementary particles. Hydrogen and helium are the basic building blocks involved in the processes and events usual in the course of stellar evolution. The abundances of the elements are (roughly) inversely proportional to their atomic weights, and the heavier elements are found only in stars that have undergone one or more novations or supernovations (explosions during which the outer shells of stars are blown off). The Earth's Sun, therefore, is a second- or third-generation star; and the Earth's composition of the heavy elements and radioactive isotopes and, thus, its internal differentiation, thermal history, and atmospheric evolution are different than would have been the case otherwise.

Distribution of elements

After the Sun's last novation (or supernovation), a solar nebula or plasma of luminous, incandescently hot, thinly dispersed matter presumably extended beyond the present orbit of Neptune and perhaps as far out as Pluto. If the Sun had simply condensed from such a state, the distribution of angular momentum in the solar system would be very different from what it is, however. Angular momentum is the product of the moment of inertia of a body (the distribution of mass about its rotational axis) and its speed of rotation, or angular velocity. If angular momentum is to be conserved then one of these quantities must increase as the other decreases. The Sun has more than 99.9 percent of the matter in the solar system, but less than 1 percent of the total angular momentum of the system. The great decrease in the moment of inertia was not accompanied by the much higher speed of rotation needed to compensate for this and to conserve angular momentum within the Sun. Instead, as several astronomers have suggested, there was apparently an outward transfer of angular momentum and of matter, in the form of hot ionized gases moving along, but not across, magnetic lines of force, such that the angular momentum of the entire solar system became concentrated in the outer planets, especially in Jupiter.

Planetary atmospheres

Outward transfer of matter was restrained gravitationally, however, so that the lighter components travelled preferentially to greater distances from the central Sun, leaving the heavier ones behind. Thus the planets closest to the Sun have relatively larger fractions of the heavier elements, and those more distant are composed mainly of hydrogen and helium, all of which bears on the question of what manner of atmosphere the Earth had when it first acquired its present mass and before it underwent subsequent evolution.

One approach to this question is to look at the atmospheres of the outer or Jovian planets (Jupiter, Saturn, Uranus, Neptune, and Pluto), which, having far greater mass and gravitational attraction than the terrestrial planets (Mercury, Venus, Earth, and Mars), would accumulate with and retain primitive planetary atmospheres. Such an approach can be taken to imply that the primary terrestrial atmosphere probably contained a lot of methane and ammonia, as well as hydrogen and helium. A difficulty with this method, as implied above, is that the atmospheres of the outer planets may be products of their location within the evolving solar nebula; hence, they would not be representative of initial planetary atmospheres.

A second approach, adopted here, is to seek a model that is consistent with what is known about the present chemistry, probable origin, and evolution of the solid Earth and its life forms. Such a search leads to the concept of a primary Earth without an atmosphere or with a very thin one. The present atmosphere, then, has probably evolved from primary components occluded within the Earth as a consequence of outgassing (emanation of gases from the Earth's interior, principally by volcanism), weathering, irradiation, biological processes, and selective escape and geochemical alternation and removal of components. Hydrogen and helium, however, are the only elements that now escape the Earth's gravity field.

This article treats the evidence sustaining this model of the Earth's original atmosphere and its subsequent development. A section on the effect of man's activities also is included. For relevant information on the evolution of the Earth's hydrosphere, see OCEANS, DEVELOPMENT OF; and for treatment of the development of the Earth and its life forms through time, see PRECAMBRIAN TIME; FOSSIL RECORD; and EARTH, GEOLOGICAL HISTORY OF. See ATMOSPHERE for a detailed account of modern properties and composition.

ORIGINAL ATMOSPHERE OF THE EARTH

Noble gases and theory of planetary origins. A number of investigators have commented on the great deficiencies of the noble gases, such as neon, argon, krypton, and xenon, in the terrestrial atmosphere as compared with their cosmic abundances. These gases are generally inert and relatively heavy. Helium is the only one having sufficiently high volatility and sufficiently low escape velocity (besides hydrogen) to be lost from the terrestrial atmosphere at existing temperatures. This can only mean that the Earth's present atmosphere has evolved from one deficient in the noble gases, either because the Earth's surface generally was exposed to very high temperatures (5,000° to 8,000° K) at some time in the geologic past or because the Earth accumulated without a primary atmosphere. (The Kelvin [K] temperature scale is an absolute Celsius scale; 0° K equals −273.15° C or −459.7° F, which is absolute zero.)

It is difficult to choose between such alternatives on the basis of direct evidence. Because life could not have survived the temperatures needed for the escape of the noble gases, no such thermal episode can have occurred during the continuous record of biological evolution, which began probably 3,200,000,000 years or more ago (3.2 aeons). There is a record of a possible major thermal event of global dimensions perhaps 3.6 aeons ago, however, that could account for such a loss of the primary atmosphere, and the only record so far available of Earth history prior to that time is the lead- and xenon-isotope data that imply birth of the planets from the Sun over a short interval about 4.6 to 4.8 aeons ago.

The nature of the initial atmosphere, if any, can only be deduced from the theory of planetary origins. Assuming that the Earth originated as a result either of the condensation of hot gases or the gravitational aggregation of cold particles, what kind of volatile envelope would it have had at the conclusion of this process? However it began, the final aggregation of the Earth had to be the result of the gravitational aggregation of solid particulate matter. When large enough, such a nucleus would be able to retain, by gravitational attraction, any gaseous components that had not previously radiated

away and whose escape velocities were low enough at temperatures then existing to prevent their loss. The ratios of present planetary masses, compared with their atmospheric compositions, strongly imply that the terrestrial planets are too small for the latter stage to have occurred. The great bulk of volatile gases in the solar system at that time was either within a then retracted solar disk or grouped around the outer, Jovian planets. The atmospheres of the terrestrial planets, including Earth, can have had only a trivial primary component, at most.

Even assuming the condensation of the atmosphere and hydrosphere from a completely molten Earth (a once popular model), without subsequent additions only about 10 percent of the water present as vapour and liquid could be accounted for. This would pose drastic problems of scale in explaining the present quantities and ratios of the atmospheric gases (see ATMOSPHERE).

Juvenile gases

Thus, the most probable explanation is that the Earth's atmosphere and the hydrosphere that condensed from it originally arose from within the Earth and thus most probably consisted of those gases that can reasonably be regarded as "juvenile." A gas is considered to be juvenile if it is found in gaseous inclusions in plutonic rocks (deep-seated crystalline rocks) or if it is a component of lavas or thermal waters. Because waters can recycle a significant fraction of such volcanic gases in the course of the hydrologic cycle, however, more work is needed to determine precisely the kinds and proportions of truly juvenile gases. Gases now thought to be juvenile and thus likely components of primitive atmosphere include, above all, water vapour. There is a difference of opinion as to whether the other main volatile elements were primarily in the form of methane (CH_4) and ammonia (NH_3), subsequently oxidized, with escape of hydrogen (H_2) to carbon monoxide (CO), carbon dioxide (CO_2), and nitrogen (N_2) or whether, in fact, the latter gases are juvenile, along with lesser quantities of H_2, hydrogen chloride (HCl), sulfur (S), Fluorine (F_2), Argon (Ar), helium (He), hydrogen sulfide (H_2S), ammonia (NH_3), and methane (CH_4).

Anoxic state of original atmosphere. There is no juvenile or primary source of gaseous oxygen (O_2) in the Earth, although combined oxygen is the most abundant element in the lithosphere and accounts for over 20 percent of the atmosphere. The free O_2 in Earth's atmosphere, therefore, must have a secondary source. Moreover, because of its striking chemical activity, it should be possible from geochemical evidence to deduce something both about when free O_2 began to accumulate in the terrestrial atmosphere and about prior and subsequent atmospheric evolution.

Whatever may have been the source and nature of the atmosphere before the beginning of the record of Earth history about 3.6 aeons ago, there is ample independent evidence that the atmosphere during the interval from 3.6 to about 1.9 aeons ago was essentially devoid of O_2; that is, it was anoxic, if not reducing (a reducing environment is one in which substances will not be oxidized by loss of electrons—*e.g.*, iron will not rust). This evidence is of four main kinds: (1) Prebiological evolution of the primary organic molecules and the origin of life could not occur in the presence of free O_2. (2) It is difficult to account for the transportation of the iron that formed the banded iron formations (BIF) unique to this part of geologic time except in the ferrous state, which could not persist in the presence of free O_2. (3) Fresh, worn, and rounded grains of pyrite (iron disulfide) and uraninite (a uranium-bearing mineral), obviously transported and shaped by traction in shallow waters, are abundant and widely distributed in some rocks older than about two eons. Such readily oxidizable minerals do not today appear as detrital (particulate) sediments except under local glacial regimes. (4) The fact that sedimentary carbon older than about two aeons is in the form of elemental carbon and kerogen and not ordinarily as carbonate rocks implies a scarcity of free O_2, as does the absence of other oxidized sediments (other than BIF) in the older rocks.

Evidence of the absence of O_2

Inasmuch as such evidence affords almost the only controls on speculation as to the nature of the original atmosphere other than that of a purely theoretical nature, it is worth examining in some detail.

1. Why, for instance, is it said that life and its precursor molecules could not have originated in the presence of free O_2? The evidence is partly experimental. Assuming an evolutionary origin, the large organic molecules of which proteins and the nucleic acids are constructed must exist before they can be assembled into an organism. Experiments utilizing various energizing mechanisms on different atmospheric models consistently fail to produce such molecules when O_2 is present. They do produce them, however, in plausible primitive-atmosphere models where free O_2 is absent and under a variety of temperatures. Ultraviolet (UV) irradiation of an atmosphere of likely juvenile gases characteristically yields hydrogen cyanide (HCN), formaldehyde (HCHO), and sometimes ammonia (NH_3), from which amino acids, polypeptides, sugars, lipids, and nucleotide bases can be synthesized, provided there is no free O_2 to degrade primary components or products. Even if such a mixture could originate in the presence of O_2, it could not survive but would oxidize back to H_2O, CO_2, and N_2 or NO_2 (nitrogen dioxide). And if it could persist long enough for living cells to arise from it, they could not survive in the absence of suitable oxygen-mediating enzymes, which represent a significant biochemical advance over the simple, anaerobic first organisms. The fundamental life processes are reducing processes, and free O_2 generally is lethal to all forms of life. The existence of living things, therefore, is in itself evidence that the original atmosphere was deficient in free O_2.

Banded iron formations

2. The banded iron formations (BIF) have long been a geochemical mystery. Occurrences of these distinctive, rhythmically banded, iron-rich rocks are accepted by all as chemical deposits precipitated from open water bodies as a result of some recurrent equilibration process. But what process will explain simultaneously how the vast quantities of iron involved could have been transported, presumably in the ferrous state, given the presence of free O_2 or, alternatively, how the iron-bearing minerals hematite (Fe_2O_3) and magnetite (Fe_3O_4), which are represented in the oxide facies of the iron formation, could have originated in the absence of free O_2? Some of the hematite, to be sure, is a recognizable alteration product, and it is barely possible that most of the O_2 is secondary; but whether the oxidation be secondary or primary, the presence of contemporaneous atmospheric O_2 would have inhibited the mobilization and transport of the source iron by causing its retention within the weathering profile (or near the source area) in the ferric state. Thus the existence of the BIF and its limitation in geologic time is best explained under a contemporaneous atmosphere without free O_2 but with a limited source of local O_2 that could combine with the ferrous oxide (FeO) in solution to bring about its precipitation (as hematite or magnetite). A biological source for such O_2, involving local anaerobic photosynthesis, with the ferrous ion as an O_2 acceptor, is a possibility to be considered below.

Detrital pyrite grains

3. The occurrence of readily oxidizable minerals as detrital grains in sedimentary rocks older than about two eons but not in younger sediments (except locally in glacially transported deposits) is the third item of concrete geological evidence opposed to appreciable oxygen in the original atmosphere. That such minerals do occur abundantly as detritally rounded grains in conglomerates and sandstones associated with middle pre-Paleozoic (older than 2,000,000,000 years) sequences of the Witwatersrand in South Africa, the Blind River area of Ontario, and other places is now well established. What is not established is whether they necessarily imply absence or scarcity of atmospheric oxygen. On the present surface of the Earth such minerals survive unaltered in detrital form only associated with glacial deposits in high latitude or alpine regions. A similar relation is possible for the pre-Paleozoic occurrences. Their former extent and abundance, compared with their absence from younger occurrences, however, favours a more general

explanation and is most consistent with the evidence cited above as indicative of an anoxic atmosphere for the first half of pre-Paleozoic time.

4. Carbonate rocks older than two aeons are uncommon, sedimentary carbon of this age being mainly in the form of graphite and kerogen. This implies an insufficiency of free O_2 to balance against the ultimate loss occasioned by conversion of CO_2 to CO_3^{2-} (carbonate ion).

There are, of course, uncertainties. In addition to the fact that detrital pyrite and uraninite do, in special circumstances, survive unaltered under the present highly oxidizing atmosphere, there is the painful incompleteness of the pre-Paleozoic stratigraphic record, with opportunity for the loss or nonrecording of intervals equivalent to eras of geological time. The possible range of geological, geochemical, and experimental error in the chronological resolution and correlation of pre-Paleozoic events is also considerable—in many instances as much as 10 percent or more of the total time involved.

A degree of confidence in the conclusion that the original atmosphere was anoxic and that the change to an oxygenous atmosphere began about 1.8 to 2 aeons ago is not provided by the force of any particular line of evidence, but by the convergence of many different lines of evidence toward this same conclusion.

EVOLUTION OF THE PRESENT TERRESTRIAL ATMOSPHERE

An initial atmosphere of any nature is bound to undergo change. Factors involved in atmospheric change include atmospheric growth through addition from juvenile sources and variation in composition as a result of removal of components or change in state due to physical, chemical, and biological processes. Foremost among additive mechanisms is outgassing of volatiles from the Earth's interior, accompanying plutonism and volcanism (including hot springs, and geysers and fumaroles [*q.v.*]). Of lesser effect is the production of helium and argon resulting from nuclear decay of naturally radioactive isotopes and implantation of hydrogen and other gases by the solar wind.

Changes in composition of the atmosphere have resulted from the photodissociation of water vapour, followed by escape of hydrogen from the Earth's gravity field, and from green-plant photosynthesis, followed by sedimentary removal of carbon. Other changes have been brought about by oxidation of formerly reduced gases in the atmosphere, ions in the hydrosphere, and minerals at the surface of the Earth. And still others have resulted from the removal of both oxidized and reduced substances, as well as of N_2 and CO_2 as a result of soil formation and chemical and biogenic rock-forming processes. Such mechanisms are considered below.

Escape of volatiles from the Earth's interior

Additions to the atmosphere. Escape of occluded volatiles from the Earth's interior is the primary mechanism of atmospheric origin and growth. Such escape, of course, cannot take place until the rocks in which the gases were originally sealed are raised to melting temperatures and moved to low-pressure environments at or near the surface of the Earth. Such melting could have occurred during or following the Earth's accumulation, as a result of the conversion of gravitational energy to heat or later in response to the release of energy during the nuclear decay of potassium-40, or still later as a result of tidal forces set up by near approach of the Moon to the Earth. A very early melting event or one that occurred as recently as about 3.6 aeons ago could account for a large initial outgassing that would start the atmosphere and hydrosphere of the Earth at substantial fractions of their present size. William W. Rubey, a U.S. scientist who is one of the leading investigators of atmospheric evolution, has shown that no more than about 10 percent of the water in the hydrosphere can remain in the gaseous state in equilibrium with an initially molten Earth, the rest being contained in solution in the melt. For this and other reasons of geochemical balance elaborated by Rubey, it is necessary to assume that the hydrosphere and, thus, the atmosphere grew to their present size as a result of continuous or recurrent additions over geologic time and, in fact, are still growing.

To obtain a reasonable geochemical balance the atmosphere and hydrosphere must be considered together, as different phases of the Earth's volatile envelope. Rubey showed that, omitting O_2, the volatiles that now make up the Earth's hydrosphere and atmosphere (H_2O, 92.8 percent; C as CO_2, 5.1 percent; Cl_2, 1.7 percent; S, 0.13 percent; N_2, 0.24 percent; H_2, 0.07 percent; and traces of Ar and F_2) are found in crudely similar proportions in lavas, hot springs, fumaroles, and geysers and that current rates of emission, continued over geologic time, are more than sufficient to account for all of the present atmosphere and hydrosphere. It would be a digression here to consider the many subtleties involving recycling of metric waters and gases, chemical reactions with previously deposited rocks, and variations in rate that may be involved.

A second source of addition to the atmosphere, trivial in quantity but important for other reasons, is the production of helium as a by-product of the nuclear decay of uranium and thorium to lead and of argon as a result of the similar decay of potassium-40 to argon-40 (and calcium-40). This suffices to account for most of the roughly 1 percent Ar and 0.0005 percent He in the atmosphere and for the fact that these are relatively more abundant than noble gases of exclusively cosmic source.

Thus, the atmosphere has grown by accretion over geologic time; but its composition has changed concurrently as a result of processes mentioned, and the juvenile gases on the primitive Earth may have appeared in different guise. Much or all of the carbon dioxide, for instance, may have started out as carbon monoxide or even methane; hydrogen may have been more abundant and perhaps mainly in the form of methane and ammonia. The nitrogen also may conceivably have been initially in the form of ammonia. The compositions of the initial gases vary in accord with the chemical processes an investigator visualizes as taking place within the primitive Earth, and how strongly reducing the initial atmosphere is estimated to have been. If it contained large amounts of H_2, for instance, this would react with any CO or CO_2 to produce CH_4 and other C–H compounds (plus H_2O), while N_2 and H_2 would react to form NH_3; if the oceans were relatively alkaline (ph about 8.2) NH_3 would appear there as NH_4^+ (ammonium ion).

Changes in composition. What are the processes that brought about changes in the composition of the atmosphere, including changes in the relative proportion of gases that were not physically split or chemically degraded? The most striking aspect of the Earth's present atmosphere is its highly oxidized state. Excluding essentially inert nitrogen and argon, it consists primarily of oxygen with a little carbon dioxide and water, the latter being the oxidized counterparts of carbon monoxide and hydrogen. Likewise, any primary methane or ammonia has been oxidized to carbon dioxide (CO_2), nitrogen dioxide (NO_2), nitrogen, and water. Above all, therefore, a source of oxygen unknown as a juvenile gas, must be sought to account for this remarkable situation.

Sources of free O_2

Two sources of free O_2 are known, both involving photolytic dissociation—the action of light in splitting the H_2O molecule. A strictly photolytic dissociation of H_2O takes place under the impact of light photons (bundles, or pulses, of electromagnetic radiation) of wavelength less than 2390 Å (one angstrom unit [Å] equals 10^{-8} centimetre) with maximum absorption at around 1650 Å and effective absorption between 1750 and 2030 Å: this dissociation can be expressed

$$H_2O + UV \rightarrow OH + H,$$

in which H_2O is water, UV is the incident ultraviolet radiation and OH and H are the hydroxyl and hydrogen released. Green-plant photosynthesis involves carbon dioxide assimilation under the energy derived from electrons acquired by splitting the water molecule. The removal of carbon and its compounds from the system can then lead to the accumulation of oxygen.

Photolytic dissociation of water vapour in the outer atmosphere, permitting escape of hydrogen atoms from the Earth's gravitational field, has been considered to be

an inadequate source of the Earth's atmospheric and combined O_2, because the process becomes self-limiting when the quantity of photons absorbed by O_2 and O_3 (ozone) exceeds that absorbed by H_2O. It has been stressed, however, that O_2 is a much more important absorber than O_3 and that photochemical limitations cannot rule out the possibility that a substantial number of hydrogen atoms are liberated by photolysis and do survive long enough to avoid recombination to H_2O; thus they would escape the Earth's gravity field. Given a relatively constant volume of atmospheric H_2O during geological time and a roughly constant density of solar radiation, it can be concluded that O_2 derived from photolytic dissociation of H_2O alone (that is, without green-plant photosynthesis) could have reached about one-fourth the present atmospheric level early in Earth history and would have remained above such a level for most of geological time.

This contradicts the conclusion, expressed above, that free O_2 did not appear in appreciable quantities in the Earth's atmosphere prior to about 1.8 to 2 aeons ago. What could explain the discrepancy? The most likely explanation is that removal of O_2 by recombination with atomic hydrogen and by combination with other reduced atmospheric components, solid substances, and ions in solution, equalled or exceeded its rate of photolytic generation through much or all of geological time. In considering this factor, it has been argued that most atmospheric oxidation involves the formation of hematite (Fe_2O_3) from FeO. As Rubey showed in 1951, however, even larger pools of O_2 are tied up with organic matter in sedimentary rocks, in sedimentary sulfates, with the sulfate ion in seawater, and with the carbon dioxide required by the amount of carbonate ion in carbonate rocks. It even has been estimated that the rate of release of juvenile CO alone is more than sufficient to use up all photolytic O_2 during oxidation to CO_2, although this is probably not true. Finally, the rates of escape of hydrogen from the Earth are not really known, and it is unlikely that the solar UV flux was as high during the first aeons after the aggregation of the planets as it is now. In view of all of these difficulties and the fact that Rubey's data strongly imply a geochemical balance between carbon and oxygen, it seems reasonable to believe the evidence in favour of an initial anoxic atmosphere.

Something in addition to photolytic O_2 was evidently needed not only to provide a source for the present, richly oxygenous atmosphere but even to account for the lesser levels of oxygen concentration required to explain the atmospheric oxidation of the oldest recorded terrestrial red beds (ferric-iron containing terrestrial sediments) about 1.8 to 1.9 aeons ago.

Production of O_2 by green-plant photosynthesis is the obvious mechanism but, like the O_2 resulting from UV dissociation of H_2O, it requires sequestering of products that otherwise would be reoxidized to their original states. That is why the suggestion that the reproductive rates of phytoplankton are sufficient to account for all the Earth's oxygen in a very short time, starting from a small population, is not to be taken seriously. Unless carbon is removed in some form (*e.g.*, as reduced carbon or hydrocarbons) photosynthetic O_2 cannot accumulate. And, because O_2 is itself lost as a result of a variety of processes, implied above, sequestering of accessory products must exceed loss of O_2 in order for the latter to accumulate. The counterparts of our heritage of photosynthetic O_2 are to be sought in the graphitic schists, carbonaceous shales, limestones, and coal beds of Phanerozoic and older geological time as well as in the carbon that has been recirculated along the Benioff zones (seismic and volcanic zones associated with drifting continental margins) of the past. Together with O_2 from outer atmosphere photolysis, this was sufficient to outweigh the many reactions that work against the accumulation of oxygen in the atmosphere. Just how the various interacting factors may have functioned to bring about that result is suggested in the accompanying table.

Hypothetical course of atmospheric evolution. When the aggregation of the planets from eddies within the former solar disk was completed about 4.6 to 4.8 aeons ago, the mass of the Earth was too small to retain a surrounding cloud of volatiles of primary cosmic origin. Instead, volatiles occluded within the aggregated rocks, and minerals were thermally released to form the proto-atmosphere as a result of postaggregational core-forming processes and radioactive heating. There are theoretical reasons for concluding that such an atmosphere in its earliest stages may have included substantial quantities of CH_4 and perhaps NH_3, as well as the components now identified as the dominant juvenile gases, but there is as yet no geological record of such an atmosphere.

Oldest sedimentary rocks

The first concrete evidence of a terrestrial atmosphere is provided by the oldest sedimentary rocks, those more than 3.2 aeons old in the Swaziland System of eastern South Africa and northern Swaziland. These rocks, which may be as old as 3.4 or more aeons, could not have originated except in the presence of an atmosphere to produce the needed weathering and erosion of pre-existing rocks and a hydrosphere in which they could acquire their structures, which indicate water transportation and sedimentation. The Swaziland succession of rocks is a fairly normal succession of geosynclinal sediments (those deposited in subsiding trenches), including island-arc types of volcanic rocks in the lower part, terminating above in a coarse nonmarine sequence (but without oxidized sediments), and containing abundant carbonaceous shale and black chert and very little limestone in the middle and lower parts. It implies formation under a hydrologic regime not greatly different from that now prevalent except with regard to the absence of free O_2 and SO_4^2. Had there been much NH_3, given atmospheric and hydrospheric CO_2 no less than now, carbonate ion should have been abundant and limestone common instead of rare among rocks of this age. Had there been enough H_2 to account for much of the carbon in the atmosphere as CH_4, one would expect the nitrogen to have been present as NH_3. The oldest sediments thus suggest a contemporaneous hydrosphere having a pH (measure of acidity–alkalinity) similar to or slightly less than that of the present and an atmosphere composed of likely juvenile gases, primarily H_2O, CO, and CO_2 with lesser N_2 and perhaps small quantities of H_2, CH_4, and NH_3.

Earliest life

From such an atmosphere, energized by UV irradiation in the 2400 to 2900 Å wavelength range, chemical evolution could have brought about the synthesis of large organic molecules, which accumulated and underwent hydrolysis, polymerization, and other changes in the waters beneath, eventually becoming organized into living cells. The record of metamorphic ages in the oldest rocks from several continents combines with the evidence of lead isotopes to suggest that such an atmosphere may have originated (perhaps after loss at high temperatures of an antecedent atmosphere of unknown composition) as a result of outgassing accompanying a global thermal event about 3.6 aeons ago. If that is the case, life probably arose very soon after a substantial atmosphere and hydrosphere had accumulated, because the old sedimentary rocks of the Swaziland System contain what are reasonably (but not conclusively) interpreted as fossil microorganisms, as well as much carbon of probable biological origin.

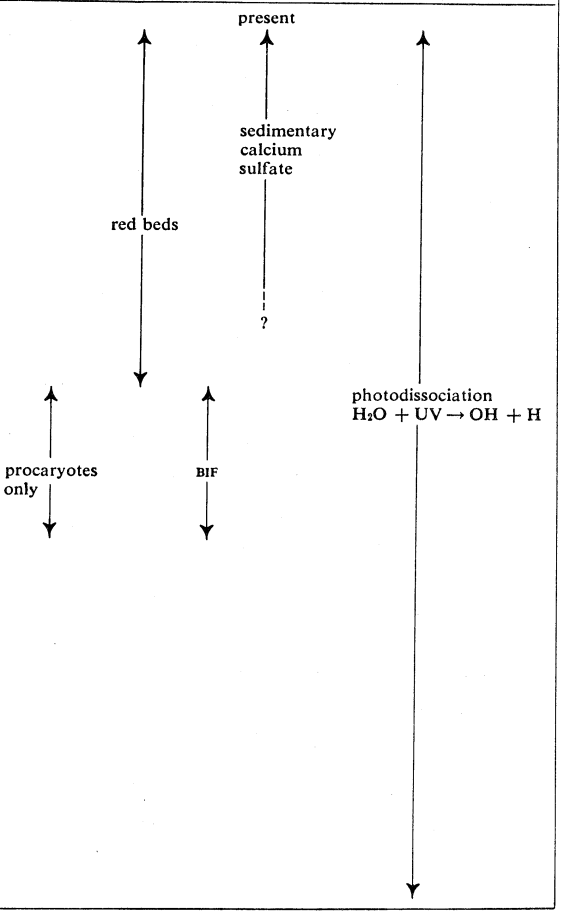
For reasons given earlier, it is not likely that the first cells were able to tolerate free O_2. Because of the biochemically advanced state of the processes involved in creating food from a mineral or gaseous substratum, it is also likely that they were dependent on a pre-existing nutrient source of abiotically evolved organic molecules. For these and other reasons, it is probable that the first cells were anaerobic, heterotrophic, and procaryotic (without nuclear membrane, organelles, or mitotic cell division).

Green-plant photosynthesis

The evolution of oxygen-releasing photosynthesis is probably signalled in the geologic record by the first appearance of BIF (banded iron formations), implying a local source of O_2 in the then generally anoxic hydrosphere. Such a signal is given by the BIF of the Swaziland System and by possibly near-contemporaneous deposits such as the Soudan BIF of northern Minnesota. Although

Content and Progress of Earth's Atmospheric Evolution

approximate years before present $\times 10^{-9}$ (not to scale)	
0.02	grasses appear, mammals diversify, atmosphere approaching present composition
0.07	more volcanism, more coal; carbon dioxide then oxygen increases
0.10	mammals, flowering plants, social insects appear
0.23	great volcanism (outgassing of carbon monoxide, carbon dioxide, water)
0.30	great coal formation, elaboration of land plants and insects
0.38	insects appear
0.40	land plants appear
0.64	oxygen reaches 3 percent of present atmospheric level (PAL); metazoa appear; glaciation; oxidative enrichment of banded iron formation (BIF); oxygen increasing; carbon dioxide decreasing; ozone accumulating as ultraviolet radiation shield; eucaryotes appear
1.8–2.0	begin evasion of oxygen from hydrosphere to atmosphere; advanced oxygen-mediating enzymes; biological production of oxygen in hydrosphere (plus photolytic oxygen) in fluctuating balance with ferrous oxide and other oxygen-sequestering systems: $4FeO + O_2 \rightleftarrows 2Fe_2O_3$; start of oxygen generating photosynthesis in hydrosphere in dependence on oxygen sink of ferrous ion
> 3.2	oldest terrestrial sediments; clear evidence of extant hydrosphere and atmosphere; origin of anaerobic, heterotrophic, procaryotic cells; abiogenic evolution of organic molecules under ultraviolet irradiation; major outgassing gives rise to primitive atmosphere of water, carbon monoxide, carbon dioxide, hydrogen, hydrogen chloride, nitrogen, etc; no free oxygen
3.5–3.6	oldest earth rocks; global melting event; moon approaches earth (?); no records available; (possible time of core formation and origin of magnetic field)
4.6–4.8	age of meteorites and terrestrial lead; origin of solar system; no records available; (possible time of core formation and origin of magnetic field)



it is an open question in biology whether the initiation of green-plant photosynthesis preceded, followed, or was accompanied by the evolution of oxygen-mediating enzymes, it is at least possible that FeO or the simple ferrous ion served as an immediate oxygen acceptor, and that the corresponding reduced biologic compounds were sequestered in other parts of the iron basins, perhaps in areas where the sulfide facies of the BIF were being formed. Some such mechanism could provide the episodic local sources of O_2 needed to account for the precipitation of the oxide facies of the BIF, while simultaneously accounting for its limitation in time and the ability of the first green-plant photosynthesizers to survive until suitable oxygen-mediating enzymes could arise. There is presumptive, although by no means compelling, evidence that photosynthetic procaryotes did exist in association with the oldest BIF and conclusive evidence that blue-green algae similar to modern forms were extant around 1.8 to two aeons ago. Thus the suggested mechanism is consistent with available evidence.

But the reduction of peroxides and of oxygen is universal among aerobic organisms, including modern blue-green algae. And the entry of oxygen-mediating enzymes capable of accomplishing these tasks was an event of such significance in biochemical and atmospheric evolution that it should be recognizable in the sedimentary records.

Oxidation of ferrous iron and appearance of red beds

Once advanced oxygen-mediating enzymes appeared, organisms possessing them would be freed from dependence on local oxygen acceptors such as ferrous iron. The photosynthetic population would then be able to exist independently and, given removal of carbon, would presumably bring about a depletion of iron in solution in the marine hydrosphere, followed by the general evasion of O_2 from hydrosphere to atmosphere. This would lead to a buildup of O_2 in the atmosphere and, consequently, oxidation of ferrous iron at the Earth's surface and its retention in the weathering profile. As a result, ferrous iron would no longer be transported to the sea in quantity in surface waters. No more BIF could form (except locally near volcanic sources of iron). Subsequent iron formation instead would be in local reducing environments (bog iron ores) and as earthy replacement deposits where iron transported in oxygen-depleted groundwaters moved into environments suitable for chemical replacement of calcium carbonates by ferrocarbonates and ferric oxides (minette ores). At the same time, the retention of iron in the weathering profile would lead to the coating of terrestrial detrital grains by ferric oxides and their deposition on land surfaces as red beds, consisting of such grains surrounded by a matrix rich in ferric oxides. Inasmuch as the oldest volumetrically significant red beds and the youngest volumetrically significant BIF both appear to be about 1.8 to two aeons old, it seems likely that this may have been the interval in geologic time during which free O_2 began to accumulate in the atmosphere, perhaps as a consequence of the evolution of the advanced oxygen-mediating enzymes among micro-organisms that were like modern blue-green algae except in their previous lack of such enzymes.

Thereafter there appears to have been a fairly progressive, though, no doubt, fluctuating accumulation of atmospheric O_2, accompanied by residual enrichment of N_2 and presumably by decrease of CO_2 as both it and O_2 were removed to form carbonate rocks and hydrocarbons. The eucaryotic cell, implying an appreciable level

of free O_2, appeared in the geologic record at least 1.2 to 1.4 aeons ago; and there was enough O_2 by the beginning of Phanerozoic time (about 0.68 aeon ago) to account for the massive formation of sedimentary sulfates and a widespread episode of oxidative enrichment of the BIF. A possibly related feature is the widespread glaciation attributed to younger pre-Paleozoic time. Such glaciation, if it proves to be real, could have been triggered by concomitant decrease in heat-retaining CO_2 and could have facilitated the concurrent oxidation of the BIF by bringing about a general lowering of continental water tables.

Attainment of present atmospheric levels of O_2

At the onset of Phanerozoic time, perhaps several percent of the present atmospheric level (PAL) of O_2 was already in existence. This follows from the fact that differentiated multicellular animal life (Metazoa) dependent on relatively high levels of free O_2 was then extant.

For the later records of Phanerozoic history, estimates of approximate O_2 level in the atmosphere may be wrung from such crumbs of evidence as the likely O_2 requirements of advanced life forms, episodes of volcanism, and probably from studies of sulfur isotopes, although it is not completely clear yet what variations in the latter signify. Among other things, they seem to imply little O_2 in the atmosphere before 0.8 aeon ago.

It has been suggested that the onset of terrestrial vegetation in Late Silurian or Early Devonian periods represented the attainment of ten percent of the present atmospheric level of O_2 and that the active Carboniferous insects might indicate something near present levels. Such estimates, bolstered by impressive mathematics, are still merely that; and there are other events that may equally well correlate with similar levels of O_2. The emergence or ascendancy of the social insects and the mammals or the formation of vast coal deposits during Carboniferous and Late Cretaceous periods are examples.

At the same time, great episodes of volcanism such as those of later Permian and Triassic and of Late Cretaceous periods may well have introduced sufficient CO into the atmosphere to lower the quantities of O_2 as a consequence of the oxidation of CO to CO_2. Or there may have been some causal linkage between increased CO and CO_2 from volcanism, enhanced plant growth, coal formation, and eventually increased O_2. The continuity of Metazoan evolution is sufficient proof that the atmosphere from about 0.68 aeon onward was never wholly depleted of O_2, yet there could have been wide excursions in abundance. Such excursions would certainly be reflected in and might be deciphered from the vagaries of biospheric, lithospheric, and chemospheric evolution. Although the possibility of such effects needs study, invoking sulfur and oxygen isotope ratios as indicators, it is too soon to go beyond the suggestion that such investigation offers hope of a more refined atmospheric history than is available at present, including both composition and temperature regime.

By processes such as those outlined above, dim though the details may be, the evolution of the Earth's present atmosphere can be accounted for from the outgassing and subsequent compositional changes of volatiles originally occluded within a waterless and airless primitive Earth. Although confidence in the broad validity of the model outlined is warranted by the fact that many lines of evidence are consistent with it, the model is still only a very rough approximation of reality, having many corollaries that remain to be investigated and many details to be added.

EFFECTS OF MAN'S ACTIVITIES

Nothing remains unchanged for very long, and the Earth's atmosphere also will continue to evolve with time. Volatile gases are constantly being added (though probably in decreasing quantity) by volcanic and hot-spring activity, and they come under the same chemical and biological influences that have acted on such additions from the beginning.

But industrial man has now become an agent of geological magnitude, acting to bring about changes in the composition of the atmosphere that may be to his own detriment. Conspicuous among the processes instituted by man is the burning of fossil fuels for industrial and other purposes, beginning early in the 18th century and attaining a particularly high rate of combustion since about 1938. Such combustion adds a variety of noxious gases such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and hydrogen chloride (HCl), as well as large quantities of CO_2 and particulate matter that may affect the heat balance of the Earth and the atmospheric ratio of CO_2 and O_2. The processes involved are so complexly interwoven that it is hard to be sure just what is happening other than that the atmosphere is becoming increasingly polluted (see URBAN CLIMATES). A diminution of, if not an end to, such pollution can be foreseen with the expected conversion to nuclear sources as the main energy supply. But there will be trade-offs. Although nuclear waste disposal itself is a manageable item, given sufficient funds and effort, some of the expected benefits of nuclear energy will pose problems of a different sort. Electrolysis of water to produce hydrogen for use as a reductant in extractive metallurgy could lead to oxygen increases that, on a large scale, might be as much of a problem in their own way as carbon dioxide increases. Or continuing or even increasing utilization of biocides to feed expected population increases may kill off enough of the world's phytoplankton to reduce oxygen levels. Without a great deal of new research it is impossible to predict the effects to expect. An examination of atmospheric history, however, clearly shows that there is a high level of interaction among the various threads of atmospheric, biospheric, hydrospheric, and lithospheric evolution. It would be well, therefore, to consider carefully the likely consequences of all actions for the future of the environment, including its atmosphere, lest the balance inadvertently be tipped against mankind.

BIBLIOGRAPHY. L.V. BERKNER and L.C. MARSHALL, "Limitation on Oxygen Concentration in a Primitive Planetary Atmosphere," *J. Atmos. Sci.*, 23:133–143 (1966), discusses factors that affect the concentration of oxygen and ozone in the atmosphere, with suggested correlations between atmospheric and biospheric evolution. P.J. BRANCAZIO and A.G.W. CAMERON (eds.), *The Origin and Evolution of Atmospheres and Oceans* (1964), a collection of papers dealing with the sources and history of volatiles in planets of the solar system, reprints W.W. RUBEY's classic paper on the origin of the Earth's volatiles by exhalation of juvenile gases from its interior. R.T. BRINKMANN, "Dissociation of Water Vapor and Evolution of Oxygen in the Terrestrial Atmosphere," *J. Geophys. Res.*, 74: 5355–68 (1969), challenges previous views holding photolytic dissociation of H_2O to be incapable of accounting for a large part of the free oxygen in the Earth's atmosphere. PRESTON CLOUD, "Pre-Metazoan Evolution and the Origins of the Metazoa," in E.T. DRAKE (ed.), *Evolution and Environment*, pp. 1–72 (1968), summarizes biochemical, geochemical, geological, and paleontological evidence and presents a model of atmospheric and biospheric evolution antecedent to that above. G.P. KUIPER (ed.), *The Earth As a Planet* (1954), is an important collection of facts and ideas about the nature and history of the Earth as it seemed in 1954. G.P. KUIPER, "Origin, Age, and Possible Ultimate Fate of the Earth," in D.R. BATES (ed.), *The Earth and Its Atmosphere*, pp. 12–30 (1958), is a good summary of facts and views about the origin and early atmospheric history of the Earth. W.M. LATIMER, "Astrochemical Problems in the Formation of the Earth," *Science*, 112:101–104 (1950), places thermodynamic limits on permissible conjecture. E.J. OPIK, "The Moon's Surface," *A. Rev. Astr. Astrophys.*, 7:473–526 (1969), discusses Earth–Moon relations and implications of the modern theory of planetary origins for the evolution of the Earth's atmosphere. H.C. UREY, "The Atmospheres of the Planets," *Handb. Phys.*, 52:363–418 (1959), examines the data and theory for the evolution of planetary atmospheres and discusses a self-limiting model for oxygen genesis from photolytic dissociation of water vapour. F.L. WHIPPLE, "The History of the Solar System," *Proc. Natn. Acad. Sci. U.S.A.*, 52:565–594 (1964), shows how the planets may have originated as a result of condensation of the solar nebula with magnetohydrodynamic outward transfer of angular momentum and segregation of planetary matter according to mass and volatility. I.P. WILLIAMS and A.W. CREMIN, "A Survey of Theories Relating to the Origin of the Solar System," *Q. Jl. R. Astr. Soc.*, 9:40–62 (1968), is a critique of theories proposed to date.

(P.C.)

Atmospheric Sciences

Scope and organization

The atmospheric sciences are concerned with understanding the composition, structure, and behaviour of the atmosphere. Their role can be described in terms of the natural phenomena to be understood: development of cumulus clouds, raindrops and ice crystals, middle-latitude storms, hurricanes, tornadoes, clear-air turbulence, fog, fair weather, monsoons, drought, the mechanisms of atmospheric tides, the trade winds, the intertropical convergence zone, changes of climate, formation of smog, dispersion of pollutants, radio "whistlers," the blueness of the sky, ultraviolet shielding, airglow, auroras, ionospheric storms, or the evolution of the atmosphere. This partial list suggests the scope of the atmospheric sciences; they provide understanding of a very wide range of natural phenomena, many of direct, intimate impact on life, some of more subtle significance.

The atmospheric sciences can also be described in terms of their relation to the basic scientific disciplines of physics and chemistry; the categories appropriate to this classification would be fluid dynamics, photochemistry, radiation, electromagnetism, optics, and thermodynamics. They may also be classified according to regions of the atmosphere or in terms of their application to weather modification, weather prediction, space exploration, communications, management of air quality, agriculture, and transportation. The atmospheric sciences are young and rapidly developing, so it is not surprising that they are approached in a variety of ways.

An important distinction exists between the basic disciplines of physics and chemistry and the atmospheric sciences. Physics and chemistry are concerned with understanding the fundamental properties of matter. Consequently, attention is focussed on matter in the small, and controlled laboratory experiments are crucial. The atmospheric sciences utilize the principles of physics and chemistry and are concerned with properties of matter in the large. They are addressed to understanding phenomena that are characteristic of extensive systems, phenomena that for the most part cannot be studied in the laboratory.

This article treats the disciplines of meteorology, climatology, and aeronomy and includes coverage of the study of planetary atmospheres and the applications and technology of the atmospheric sciences. The section on *Meteorology* is addressed to the physical state and the dominant phenomena associated with the portion of the atmosphere that is electrically neutral (that is, not ionized) and for which the molecules are in low energy states. Under *Meteorology* the following topics are discussed:

1. *Radiation:* transfer of energy by electromagnetic waves from the Sun to the Earth, from the Earth's surface to the atmosphere, within the atmosphere, and from the surface and atmosphere to the cosmos.
2. *Thermodynamics:* transfer of energy within the Earth–atmosphere system by the processes of heating, compression and expansion, condensation, evaporation, freezing, and melting.
3. *Cloud physics:* processes that lead to the formation, growth, and precipitation of droplets and ice crystals.
4. *Turbulence:* the unorganized chaotic motions, usually of small scale, that are responsible for transfer of heat, gases, particles, or other atmospheric constituents or properties, and for the dissipation of energy.
5. *Chemistry:* the chemical changes occurring in gases and particulate matter, especially those changes associated with air pollution near the ground and with the composition of the atmosphere up to a height of about 100 kilometres (60 miles).
6. *Analysis:* development of concepts and generalizations from observations, using graphical and statistical techniques primarily.
7. *Dynamics:* mathematical theory of atmospheric structure and flow based on the fundamental principles of physics.

In addition to these subjects, air–sea interaction constitutes an important special combination of dynamics and turbulence, which overlaps meteorology and oceanography. The fields of glaciology and hydrology constitute an analogous overlapping of meteorology and the geological sciences.

The section on *Climatology* is addressed to the study of stable or slowly changing statistical states of the atmosphere—past, present, and future. Under this section are included the classification of climate based on the equations of water balance or energy balance; the history of climate based on geological, botanical, and historical records; and the theory of climate, which explains (quite inexactly and tentatively at present) climate in terms of fundamental principles.

The discipline of *Aeronomy* is concerned with the physical state and dominant processes of the ionized and photodissociated upper atmosphere. The following topics are included:

1. Photochemistry: chemical changes resulting from absorption of electromagnetic radiation above a height of about 80 kilometres.
2. Airglow: processes of the upper atmosphere that give rise to the emission of light during both day and night.
3. Ionospheric physics: structure and behaviour of layers of electrons and positive ions in the upper atmosphere and their effects on the Earth's magnetic field and on radio waves.
4. Magnetosphere: structure and behaviour of the Earth's magnetic field and its interactions with charged particles.
5. Aurora: effects of energetic charged particles in the regions of the geomagnetic poles.

The short section on *Planetary atmospheres* is addressed to an understanding of the composition and structure of the atmosphere of the planets. It treats the observations made by space probes and by Earth-based spectroscopic observations.

The following applications of the atmospheric sciences to the needs of society are discussed in the next section:

1. Prediction of dangerous weather events and other changes in the atmosphere, which is the central objective of the atmospheric sciences.
2. Modification of weather and climate in order to reduce hazards and achieve economic benefits and to anticipate possible irreversible effects on climate.
3. Management of air quality to minimize dangerous effects on man, on the ecological balance of the Earth, and on property.
4. Informational services needed for the safety and economy of air and sea transportation, for public safety on lakes and on sea, for agriculture, lumbering operations, space operations, communications, and for military operations.

Finally, a short section is devoted to the technology of the atmospheric sciences: satellites, computers, instrumentation, and the data collection, transmission, and processing systems that make possible research and operational programs in the atmospheric sciences.

The total body of information associated with the foregoing topics is enormous, and detailed accounts are provided principally in the following related articles: CLIMATE; MICROCLIMATES; URBAN CLIMATES; CLIMATIC CHANGE; WINDS AND STORMS; CYCLONES AND ANTICYCLONES; THUNDERSTORMS; LIGHTNING; HURRICANES AND TYPHOONS; TORNADOES, WHIRLWINDS, AND WATERSPOUTS; MONSOONS; JET STREAMS; AURORAS; VAN ALLEN RADIATION BELTS; IONOSPHERE; ATMOSPHERE; ATMOSPHERE, DEVELOPMENT OF; CLOUDS; PRECIPITATION; SNOW AND SNOWFLAKES; WEATHER FORECASTING; WEATHER MODIFICATION; and WIND ACTION. See also HYDROLOGIC SCIENCES, which parallels this article, treating the nature, scope, and methods of hydrology, oceanography, limnology, and glaciology; EARTH SCIENCES, which treats the historical development of atmospheric sciences, and the interrelations between atmospheric, hydrologic, geologic, and other sciences; and METEOROLOGICAL MEASUREMENT, which covers some instrumentation.

METEOROLOGY

Understanding of the structure and behaviour of the atmosphere may be gained by viewing the Earth and its atmosphere from a satellite that circles the Earth while maintaining itself at a distance of 36,000 kilometres (22,000 miles) above the Equator. Such views have been published frequently in popular magazines and books, and they show that roughly half of the Earth's surface is covered with clouds, organized into a variety of forms on many different scales. Large storm systems thousands of

Scale of atmospheric motions

kilometres in horizontal extent are associated with much smaller cloud elements organized in complex ways; spiral bands, parallel bands, isolated clouds, and other forms. By viewing a sequence of pictures, the motion of the clouds and the airflow can be detected and measured.

A satellite picture showing the horizon just after sunset demonstrates, by the scattering of light from air molecules and dust particles, that the atmosphere is strongly concentrated close to the Earth's surface. Ninety-nine percent of the mass of the atmosphere is confined to a layer 30 kilometres (19 miles) in thickness, and clouds are confined for the most part to the lowest ten kilometres (six miles).

Some of the large-scale features of atmospheric motion are persistent and recurrent; it is possible to identify these features on any series of upper air charts or from any series of satellite pictures. This global scale of motion, or planetary circulation, is characterized in middle and high altitudes by slowly changing wavelike zonal flow from west to east; in low latitudes the flow is weaker and, near the Earth's surface, is directed from east to west.

Atmospheric phenomena occur on a very wide range of time and space scales in addition to the global or planetary scale. Turbulent eddies may be a few metres or less in size, thunderstorms have horizontal dimensions of a few kilometres, fronts may be thousands of kilometres in length and a hundred kilometres in width, an area of urban smog may be tens of kilometres in extent and last for days, drought may cover an area thousands of kilometres in width and may last for months or years. All of these phenomena and others are of scientific interest and practical importance.

The fundamental problem of meteorology is to account for the most important aspects of this range of phenomena in terms of physical principles, mathematically expressed. Here it is possible only to suggest the elements of the discipline of meteorology. Mathematical formulations are not included, and the more sophisticated concepts are beyond the reach of this article. Fuller exposition is provided in the references listed at the end of the article.

Radiation. Electromagnetic radiation emitted by the Sun provides nearly all the energy that, after transformations at the Earth's surface and in the atmosphere, is responsible for the generation of storms and air circulation on all scales. Changes in the intensity of solar energy undoubtedly occur, but they are smaller than the uncertainty of present measurement.

Absorption and reflection of solar radiation

In passing downward through the atmosphere some of the solar energy is absorbed by water vapour in the troposphere. Averaged over the globe, about 18 percent of the solar radiation is absorbed in the atmosphere, and about 35 percent is reflected to space by clouds, the Earth, and the atmosphere. The remaining 47 percent is absorbed at the Earth's surface.

Clouds, water surfaces, and solid Earth surfaces emit electromagnetic energy as infrared radiation, the intensity of which increases with temperature. Of the radiation emitted upward from the Earth's surface under clear skies, roughly 60 to 70 percent is absorbed by the carbon dioxide and water vapour in the troposphere. Thus, energy is trapped by the atmosphere, with the result that the Earth's average surface temperature is about 35° C (63° F) higher than it would be without an atmosphere. Cloud layers enhance this blanketing effect. Radiation absorbed and emitted by the Earth and atmosphere is in near balance when averaged over the whole Earth; otherwise large changes in climate would occur. In the tropics the radiation absorbed exceeds that emitted, whereas at higher latitudes radiation emitted exceeds that absorbed. These differences are balanced by the heat transported poleward by large-scale air and water currents.

The net radiational warming or cooling of the atmosphere can be studied and measured by sensors carried aloft in balloons or from Earth satellites. These measurements are inherently difficult to make accurately. Net radiational warming or cooling can also be calculated from knowledge of the scattering and absorption properties of the constituents of the atmosphere: gases, cloud particles, and aerosols. Research programs are directed at improving the accuracy of measured and calculated atmospheric radiation and at incorporating radiation processes in dynamical models of the atmosphere.

Thermodynamics. The electrically neutral atmosphere consists of a mixture of so-called permanent gases, whose concentrations are essentially constant, as well as several gases whose concentrations are variable, and various solid and liquid particles. The permanent gases are: nitrogen, 78 percent by volume; oxygen, 21 percent; argon, 1 percent; and other gases in much smaller concentrations. Because of turbulence they are mixed uniformly throughout the first 100 kilometres, or 60 miles (of altitude) of the atmosphere. The most important variable gases are: water vapour, 0 to 7 percent; carbon dioxide, 0.01 to 0.17 percent; ozone, 0 to 0.1 percent; sulfur dioxide, 0 to 0.0001 percent; nitrogen dioxide, 0 to 0.00002 percent.

The permanent gases can be treated as ideal gases; that is, an equation of state can express accurately the relation between gas pressure, density, and temperature over the whole range encountered in the atmosphere. The variable gases (especially water vapour) are treated separately.

Study of air and water vapour

For a static atmosphere the pressure (force per unit area) is equal to the weight of a column of air of unit cross-section extending to the top of the atmosphere. Because air is compressible, atmospheric pressure and density decrease with height exponentially, assuming uniform composition and temperature. One consequence is that pressure and density decrease much more rapidly with height near the Earth's surface than higher in the atmosphere. Therefore, the atmosphere extends hundreds of kilometres above the Earth's surface even though a column one square centimetre in cross-section contains only about one kilogram (two pounds) of air.

The first law of thermodynamics may be used to calculate that if a parcel of dry air is lifted adiabatically (without adding to or taking away heat), the temperature must decrease with height at the rate of 9.8° C per kilometre (28.3° F per mile), called the adiabatic lapse rate; that is, air lifted adiabatically cools 9.8° C for every kilometre it is lifted. Descending air warms at the same rate. In a region where temperature decreases with height more rapidly than the adiabatic rate, small vertical displacements will grow and active vertical convection occurs. The air is said to be statically unstable. In a region where temperature decreases with height at less than the adiabatic rate, vertical displacements are suppressed, and the air is said to be statically stable.

Temperature normally decreases with height within the lowest ten to 15 kilometres (six to nine miles) of the atmosphere (the troposphere) as a result of the dominant effect of vertical convection and mixing. Within the overlying stratosphere, as a result of absorption of radiation by the ozone layer, temperature increases gradually with height to a maximum at about 50 kilometres. In the mesosphere, between 50 and 80 kilometres (30 to 50 miles), temperature decreases on the average at about 3° C per kilometre, but over limited intervals of time and height at nearly the adiabatic rate, indicating that vertical mixing and convection are dominant processes. Above 80 kilometres (the thermosphere) temperature increases rapidly with height as a result of absorption of solar radiation by oxygen and nitrogen and possibly other processes. Above 150 kilometres (90 miles) temperatures exceed 1,000° K (absolute zero, −273° C, or −460° F, is taken as the zero point on the Kelvin temperature scale).

Effects of adiabatic cooling

Water vapour condenses and evaporates at atmospheric temperatures, so that water-vapour concentration varies in time and space. As a consequence of the adiabatic cooling that accompanies lifting, clouds normally form at some distance above the Earth's surface. Condensation of cloud droplets releases latent heat, which tends to warm the ascending air, thus creating a saturation–adiabatic lapse rate that is roughly half the adiabatic lapse rate. One consequence of this is that rising saturated air is more buoyant than unsaturated air, a phenomenon

that accounts for the development of cumulus clouds and the characteristic roughness of airplane flight in clouds, a subject of considerable interest and study.

Cloud physics. In order for cloud droplets to form in saturated (or even in supersaturated) air, there must be present solid or liquid particles on which condensation can occur. Natural air contains large numbers of very small particles on which condensation may occur when air is lifted. The resulting small droplets form clouds.

The initial formation of small ice crystals also occurs on minute particles called ice nuclei, which are much fewer in number than condensation nuclei. Ice nuclei are effective at a variety of temperatures depending upon the material of the nucleus. Most natural ice nuclei are effective only at temperatures below −10° C (14° F), so that cloud droplets normally remain in the liquid state at temperatures well below 0° C (32° F). Such supercooled clouds may be changed to ice-crystal clouds if ice crystals fall into them. They may also be converted to ice-crystal clouds by adding effective artificial nuclei. Silver iodide is effective at −5° C (23° F), the highest nucleating temperature of any known material other than ice crystals.

The vapour pressure of a liquid droplet exceeds the vapour pressure of ice at the same temperature. Consequently, in a mixture of ice crystals and liquid droplets, the ice crystals grow by evaporation of the droplets, diffusion of the vapour, and sublimation on the crystal surfaces. The difference in vapour pressure, hence the rate of growth, is greatest at about −12.5° C (9.5° F) where it amounts to 12.5 percent supersaturation. Understanding of the processes of ice nucleation and of the processes by which ice crystals multiply in clouds remains obscure, however. Laboratory and field research is directed at gaining a fuller understanding of the microphysics of clouds and of their interactions with the cloud environment.

Separation of electrical charge is intimately associated with the collision of ice crystals and supercooled droplets and also with collision of droplets; the mechanism or mechanisms have not yet been adequately explained.

When charge separation occurs in convective clouds, positive charges are carried on the small crystals to the upper part of the cloud, while negative charges are carried downward with the falling particles. In this way an active convective cell may generate electrical charge amounting to 1,000 coulombs in 20 minutes or so. A large electrical potential difference may develop between the cloud and ground leading to a lightning stroke, which carries negative charge to the Earth. At the top of the cloud positive charge is conducted upward to the conducting ionosphere. By this mechanism, thunderstorms maintain a potential difference between the Earth and ionosphere of several hundred thousand volts. The resulting normal vertical electrical field amounts to about 100 volts per metre near the ground; the associated electrical current of about 2,000 amperes distributed over the Earth's surface completes the electrical circuit for which the world's thunderstorms provide the generators.

Turbulence. In a fundamental sense the atmosphere is a turbulent fluid; there are few if any simple regularities. There are organized motion systems that can be described and understood in more or less complete detail, however; and there are other highly disorganized motions for which statistical analysis seems to be the only feasible approach to generalization. The latter are referred to as turbulence.

Atmospheric properties—for example, humidity, temperature, momentum, or aerosol concentrations—are transferred by turbulent air motions. Near the Earth's surface vertical flux of humidity and other conservative properties occurs entirely through turbulence. The latent heat supplied to the atmosphere by evaporation at the surface and vertical transfer by turbulence constitutes most of the energy that drives the atmosphere.

Inversions and clear-air turbulence

Above the surface layer, ten metres (30 feet) or so in thickness, vertical flux usually decreases with height. The extent of this flux is often limited by the height of an inversion layer in which temperature increases upward. An inversion often is present at a height of one or two kilometres. Inversions are statically stable in the sense that vertical displacements result in a damped periodic oscillation with periods of several minutes. When the surface layer is strongly heated, vertical convection currents develop that transport heat, water vapour, momentum, smoke, and other properties up into the middle troposphere or even higher.

A second important effect of turbulence is its role in the dissipation of energy. Thus, turbulence is essential both to the process that drives the atmosphere and to the final dissipation of the atmosphere's kinetic energy into heat.

Turbulence and dissipation occur not only near the Earth's surface but also in the atmosphere well above the planetary boundary layer. Turbulence is commonly associated with clouds, for reasons indicated earlier, and also occurs as clear-air turbulence in regions of strong shear (velocity gradients). Such turbulence may be generated by localized absorption of gravity wave energy that may have been introduced some distance away. It is estimated that the energy dissipated in clear-air turbulence is comparable to that dissipated in the boundary layer.

Chemistry. Chemical changes occurring in the atmosphere are often subtle and hard to measure, but they have important consequences. Chemical processes are important in the development of condensation nuclei and ice nuclei, in the changes in atmospheric composition that may influence climate, in the life histories of pollutant gases and particulates, and in their effects on visibility and on humans, animals, plants, and solid materials. Atmospheric aerosols are produced by many processes, in particular: condensation and sublimation of materials of low vapour pressures, reactions between gases of low concentration in the presence of sunlight or water, and dispersal of surface material, sea spray, mineral dust, or smoke. Chemical reaction rates vary with incident shortwave radiation, temperature, and humidity, as well as with concentration and the physical state of particles, so that changes in chemical state are occurring constantly, especially in urban areas. The sources of particles and of gases that form particles include: volcanoes, sea spray, agriculture, burning of fossil fuels, and evaporation of organic materials from forests.

Two processes in aerosol particle distribution

Two physical processes are important in determining the distribution of aerosol particles after they are injected into the atmosphere. Coagulation of particles smaller than 0.01 micron (one micron equals 0.001 millimetre) occurs rapidly as a result of Brownian motion (agitation of particles by molecular collision) so that smaller particles have a short lifetime. Particles of radius greater than 20 microns have an appreciable fall velocity relative to still air, so they are precipitated quickly. Under equilibrium conditions, the number density of aerosol particles is a maximum between 0.01 and 0.1 micron. Particles larger than one micron are most effective as condensation nuclei.

A worldwide layer of large aerosol particles exists in the stratosphere at a height of roughly 20 kilometres (12 miles). Sulfur is a dominant component of these particles, and it is inferred that they are formed by oxidation of sulfur dioxide, which diffuses upward from the Earth's surface.

Particulates are removed from the atmosphere by dry fallout and by attachment to cloud droplets and falling precipitation. Electrical effects undoubtedly play important parts in both processes. Gases are removed by rain and by chemical reactions with particles and with surfaces. Downwind from industrial areas, distinctly acidic rainfall is common.

Dissociation of molecular oxygen (O_2) to monatomic oxygen by absorption of ultraviolet radiation from the Sun has important effects on the composition of the atmosphere above a height of about 20 kilometres. In the region between about 20 and 50 kilometres (12 and 30 miles) the monatomic oxygen reacts with O_2 to form ozone (O_3). The resulting worldwide layer of ozone, although its relative concentration is less than 1/10,000, is sufficient to absorb ultraviolet radiation and thereby

serve as a vital protective shield for life on Earth. The equilibrium concentration of ozone depends strongly on air density, water vapour, and other variable gases, as well as on the intensities of various lines in the solar spectrum. At heights above about 80 kilometres (50 miles) where air density is about 2×10^{-8} gram per cubic centimetre, dissociated oxygen reacts slowly with O_3 and above 110 kilometres (70 miles) nearly all the oxygen is in monatomic form.

Analysis. Observations of atmospheric phenomena result in enormous numbers of data, which vary in three-space dimensions and in time. These data are reduced to tractable order by a number of analytical techniques. Graphical techniques have been important throughout development of the atmospheric sciences. In recent times, statistical techniques and objective numerical techniques have been developed, but the complexity of atmospheric behaviour is such that graphical techniques in the hands of well-trained, capable analysts remain essential aids to understanding and primary research tools.

Graphical techniques as primary research tools

Radiation charts have been devised to determine the net long-wave radiation absorbed or emitted in an atmosphere with arbitrary distribution of water vapour. From adiabatic charts both the thermodynamic energy that might be released in a specific situation and the vertical distribution of static stability can be determined.

In order to represent atmospheric structures and motions graphically, observations are plotted on vertical and horizontal cross-sections. Analysis of these charts requires that the analyst have in mind a generalized concept of the structures and fields under study. The analyst must fit constraints imposed by the dynamic equations and must make the analysis consistent with the sequence of analyses of which it is part. In many cases the objective is to depict a particular scale of phenomenon while filtering out the effects of other scales of motion. Experience, skill, and physical insight are all essential to good analysis. Many special techniques of graphical analysis are useful for determining the distributions of various scalar quantities: mean temperature of a layer, horizontal velocity divergence, vertical component of velocity, vertical component of vorticity, and others.

Objective numerical analysis can be carried out by fitting an arbitrary mathematical function to a particular set of data or by objectively smoothing the observations. These methods are particularly efficient and valuable for use with computers. In research on the structure and behaviour of complex phenomena, however, subjective analysis by an individual is indispensable.

Analysis of time series observations made at a fixed point—or in some instances from a moving airplane, rocket, or balloon—is used widely in identifying features or phenomena that are heavily obscured by fluctuations of other origins and in determining phase relationships. Spectrum analysis of fluctuating data provides a powerful tool for investigating properties of turbulence. The power spectrum represents the variance or energy of the signal in relation to its frequency; the cospectrum of two time series represents the correlation of the two series with respect to frequency; and the quadrature spectrum determines the phase relation of the two series with respect to frequency. Aspects of greatest interest are the frequencies and magnitudes of the peaks of the spectrums, location of gaps, total "energy" or area under the curve, and the shape of the spectrum.

Dynamics. To understand and predict the structure and behaviour of the atmosphere, the fundamental physical principles of conservation of mass, momentum, and energy are expressed by appropriate mathematical equations. In the most general forms these include the effects of the physical processes discussed earlier in this article: radiation, thermodynamics, cloud physics, turbulence, and chemistry. The equations are partial differential equations in space and time, which are far too complex to be mathematically tractable. Atmospheric dynamics is concerned with development and analysis of mathematical models that are simple enough to be integrated but that contain the essential physics of the phenomena of interest. Thus, there are many kinds of models: of the general global circulation, of hurricanes, of cumulus clouds, of mountain gravity waves, and so on; for each model special simplified equations are developed from which predictions of the model behaviour can be made. Many interesting and useful results may be obtained neglecting turbulence and surface friction.

Mathematical models and atmospheric-waves

Atmospheric disturbances are of several types, which differ in their driving mechanism and in the fluid-mechanical response of the atmosphere. In order to understand these disturbances or the wave forms into which they can be resolved in a dynamical sense and to predict their behaviour, separate simplifications of the general equations are invoked. Upon linearizing the equations—that is, retaining only terms of the first power in the dependent variables—the relation between speed and wavelength, the phase relations, and criteria for instability can be determined for the following major types of atmospheric waves.

1. Acoustic. Waves that are formed by sudden shocks or impulses and travel by virtue of the compressibility of the atmosphere; sound waves are of insignificant meteorological interest and can be simply filtered from the dynamic equations.

2. Gravity. Waves that result from localized vertical displacements in the statically stable atmosphere; short gravity waves are unaffected by rotation of the Earth, but waves of several thousand kilometres length are increased in speed by the Earth's rotation, and their particle motions are inclined toward the horizontal plane. Such waves are referred to as inertia-gravity waves. The energy of gravity waves may propagate vertically as well as horizontally. In this way, gravity waves originating near the Earth's surface may transmit energy into the upper atmosphere. At such heights, where the density is small, the wave amplitudes and associated velocities are correspondingly large. It has been proposed but not yet thoroughly established that absorption of gravity-wave energy in the upper atmosphere constitutes an important source of thermal energy at these heights.

3. Barotropic. Gravity waves can be filtered from the dynamic equations by a mathematical approximation. Then, assuming that the mean air flow is zonal and unchanging with height, the existence of waves that move toward the west relative to the zonal flow can be demonstrated. These waves are manifestations of the conservation of absolute vorticity of vertical columns moving in a barotropic (surfaces of equal density parallel to surfaces of equal pressure) atmosphere.

4. Baroclinic analysis of atmospheric models in which there is a north–south gradient of mean density shows that small amplitude disturbances may be unstable and that greatest instability occurs for wavelengths in the range of 3,000 to 5,000 kilometres (2,000 to 3,000 miles), approximately. Consequently, waves of this length should grow most rapidly and ultimately dominate the global circulation in middle and high latitudes. The process can be understood as transformation of the potential energy associated with the normal meridional gradient of density to kinetic energy of middle-latitude storms. In this process the cold air sinks in flowing from polar to tropical regions, while the warm air rises in flowing from tropical to polar regions; this results in widespread cloudiness and precipitation in the poleward flowing air. These air currents transport most of the energy contributed by the net radiation absorbed by the tropical oceans.

Warm air flow from tropical to polar regions

5. Tropical. Close to the Equator both gravity and barotropic effects must be retained in the linear models. There arise waves of a mixed barotropic–gravity type that are strongly dependent on latitude and tend to be trapped in a region within 20° or so of the Equator. These tropical waves are probably driven by the energy released by condensation in the tropical troposphere.

6. Atmospheric tides. Internal gravity waves of planetary scale may result from the radiational heating and cooling of the atmosphere. Theory predicts semidiurnal and diurnal waves, which correspond closely to the observed atmospheric tides. At heights above about 30 kilometres tidal motions probably dominate the flow.

In the planetary boundary layer, the one- or two-kilometre layer just above the Earth's surface, surface friction and heat transfer by turbulence are important. Interaction of the boundary layer with the free atmosphere leads to reduction of the vorticity of the free atmosphere and dissipation of its energy. Secondary circulations develop within the boundary layer on a scale of several kilometres. These circulation systems may be recognized when they produce long parallel bands of clouds or other organized cloud forms near the top of the planetary boundary layer.

In order to make specific predictions of changes, the model equations (usually nonlinear) must be integrated over finite intervals of time, and this requires numerical methods using electronic computers of very large capacity and speed. Special numerical algorithms are utilized to avoid computational instability. Numerical models of the general circulation duplicate many features of the real atmosphere, and they permit quantitative analysis of the effects of surface influences, condensation, radiation, or other aspects of the model. They have revealed that growing baroclinic disturbances transfer momentum to the zonal current, thus generating the middle-latitude "jet stream" and maintaining it against frictional dissipation. These models contain the essential dynamical elements of a continuously operating general circulation and, therefore, permit an attack on long-range prediction and the theory of climate.

CLIMATOLOGY

Climatic classification. The climate of a region is defined by weather statistics over extended periods, usually 30 years or more. Description of the climate includes statistical analyses of a variety of weather elements: rainfall and snowfall, temperature, humidity, solar radiation, atmospheric radiation, wind velocity, aerosol content, soil heat transfer, evaporation, atmospheric heat transfer, and cloudiness. Among the statistical quantities considered are: annual, monthly, and daily mean values; mean annual and diurnal ranges; and the variances of individual values from the means. For special purposes more complicated statistics are compiled; for example, correlations of wind velocity with temperature or static stability with wind velocity and with aerosol content.

The energy balance and water balance

Climate changes so slowly that balance between incoming and outgoing energy can be assumed for suitable periods. Also, water balance can be assumed for corresponding periods. The two equations representing these balances provide a quantitative basis for classification of climates. The equation relating the incoming and outgoing energy for a portion of the Earth's surface is

$$S(1-a) + R_A + H_A = R_G + H_G + H_E,$$

in which S is incident solar flux, a is albedo of surface, R_A is long-wave radiation from the atmosphere, H_A is heat conduction from the atmosphere, R_G is long-wave radiation emitted by the surface, H_G is heat conduction into the Earth, and H_E is latent heat of evaporation or condensation on the surface. The corresponding equation of water balance, relating precipitation and condensation to losses by evaporation and drainage, is

$$P + C = E + D,$$

in which P is water mass added by precipitation, C is water mass added by condensation, E is water lost by evaporation, and D is water lost by drainage. The individual terms in these two equations vary with latitude, elevation above sea level, soil characteristics, and distance from a large body of water and also vary in relation to major features of the atmospheric and oceanic circulation systems. All these factors affect the climate.

Among the many indices of climate that can be defined on the basis of the water-balance equation are the runoff ratio, $(D-E)/P$; the evaporation ratio, E/P; and the potential evapotranspiration ratio, E_0/E, in which E_0 represents the evaporation that would occur from a wet surface. Some typical values of these indices are shown in the Table.

Climatic Indices

region	$(D-E)/P$	E/P	E_0/E
Tundra	>0.7	<0.3	<1.1
Forest	0.3–0.7	0.3–0.7	1.1–1.43
Steppe	0.1–0.3	0.7–0.9	1.43–2.22
Semidesert	0.03–0.1	0.9–0.97	2.22–3.09
Desert	<0.03	>0.97	>3.09

Source: W.D. Sellers, *Physical Climatology* (1965).

Climate classifications may be based on the energy balance. For example, the Bowen ratio, defined as H_A/H_E, varies from roughly unity over land, on the average, to 0.1 over the oceans and increases strongly with latitude.

Climatic change. The record provided by geologic and paleobotanical evidence proves that climates have changed substantially. For most of the past 500,000,000 years higher temperatures than at present prevailed, at least over the Earth north of 40° N. Major glaciations did occur, however; residual evidence of the most recent ice advance is found in the glaciated regions of Antarctica, Greenland, and high-latitude mountainous areas. For the past 10,000 years the record of climatic change is reasonably detailed, with evidence from such varied sources as lake levels, tree rings, pollen found in peat bogs, and anthropological and historical records. Among the climatic features that have probably influenced recent human history in important ways are the Little Climatic Optimum in northern Europe (1150–1300), the "Little Ice Age" (1500–1700), and recent worldwide warming (1900–50).

Evidence for the detailed climatic record of the past 10,000 years

During periods of glaciation, global evaporation and precipitation were undoubtedly reduced somewhat from present values, but in regions on the Equator side of the ice sheets, considerably increased precipitation must have occurred as the middle-latitude storm track was shifted nearer the Equator. In this way substantial shifts of forests and deserts must have occurred.

Small sustained change in any one or more of the terms in the water-balance or energy-balance equations might account for change in climate. Among the mechanisms that may be of importance are the following: (1) interactive coupling of the ocean and atmosphere (two fluids with quite different response characteristics), along with the possible roles of glacial and sea ice in such coupling; (2) changes in solar radiation received by the Earth because of astronomical periodicities in the Earth's orbit about the Sun or to astrophysical changes in solar radiation; (3) effects of changed turbidity caused by volcanic eruptions and man-made pollution on the cloud cover and on absorption and scattering properties of the cloudless atmosphere; (4) effects of changed carbon dioxide content of the atmosphere from weathering of rocks, volcanic eruptions, decomposition of organic matter, combustion of fossil fuels, and the storage of carbon compounds in the ocean and in land plants; (5) changes in the albedo (reflectivity) of the Earth's surface accompanying development of agriculture or other changes in land-surface characteristics; and (6) changes in the distribution and elevation of land surfaces associated with continental drift and mountain building and erosion.

In addition to the changes that affect climate through the energy supplied to the atmospheric system, other statistical changes may be inherent in an atmosphere with fixed boundary conditions and fixed energy input. It is not known which of these mechanisms may account for past changes of climate; several or all of them may have operated simultaneously.

Numerical models operating for simulated times of one year to ten or 100 years have duplicated climatological statistics closely. Such results give some confidence that similar experiments in which the energy input or other features of the models were altered to simulate possible realistic processes might lead to an understanding of the mechanisms of climate change. The effect on temperature of doubling the carbon dioxide concentration in a model that considers some but not all the important atmospheric interactions has been computed to

be 2° C (4° F). Coupled ocean–atmosphere dynamical systems are beginning to be studied effectively. Many such experiments using more sophisticated models will be necessary to explain past climate changes with confidence and to predict future changes in climate.

AERONOMY

Prior to the 1960s knowledge of the atmosphere above about 30 kilometres (20 miles) was based largely on inferences from ground-based observations and was fragmentary and uncertain. The advent of Earth satellites and space probes made possible a decade of intensive exploratory research on the environment of the Earth in its relation to the Sun and interplanetary space. Description of this environment is now sufficiently complete and accurate that the mechanisms that control the interactions of the lower and the upper atmosphere and of the atmosphere with the Sun and with interplanetary space can now be examined effectively. To do this requires that electromagnetic forces be included in the fluid dynamical equations and that detailed observations of the state of the upper atmosphere and of its field of motion be made. Using the technique of incoherent scatter of radar waves, it is possible to measure many properties of atmospheric ions and electrons and to infer others in the region from 50 kilometres to 10,000 kilometres (30 to 6,000 miles). Included among the inferred properties are the temperature, density, and velocity of the neutral atmosphere.

In the region of the atmosphere above 50 to 80 kilometres (30 to 50 miles), corresponding to densities of 10^{16} to 10^{14} molecules per cubic centimetre, dissociation and ionization of air molecules by absorption of solar shortwave radiation are dominant physical processes. They are largely responsible for determining the composition of the upper atmosphere, its temperature distribution, and the distribution of charged particles. Interaction of ions and electrons driven by atmospheric tidal motions with the geomagnetic field gives rise to electrical currents in the upper atmosphere. Visible radiation is produced by recombination of ions with electrons and the recombination of neutral atoms. Interaction of ions and electrons streaming outward from the Sun with the Earth's magnetic field creates a boundary between the solar wind and the geomagnetic field at a distance of 14 or more Earth radii and strongly influences the behaviour of charged particles. These are a few of the dominant features of the upper atmosphere which aeronomy seeks to explain. The major subdivisions of aeronomy are those treating photochemistry, airglow, ionospheric physics, the magnetosphere, and the auroras.

The major subdivisions of aeronomy

Photochemistry. The most important photochemical processes in the upper atmosphere are those in which oxygen plays a part. Formation and maintenance of the ozone layer between heights of about 20 and 50 kilometres (ten and 30 miles) has already been discussed. At heights above 110 kilometres (70 miles) virtually all the oxygen is in monatomic form (that is, oxygen is present as O rather than the molecular O_2). In the exosphere, or region from which neutral molecules may escape from the Earth's gravitational field, dissociation of water vapour may provide a source for the hydrogen, which is a major component of this region. Dissociation of nitrogen, nitric oxide, and carbon dioxide are of lesser importance.

In the height range from about 20 to 80 kilometres (ten to 50 miles) dissociation is in approximate balance with recombination; the balance is determined by reaction rates that depend upon intensity of absorbed radiation, the concentrations of components, air density, and temperature. At heights of 100 kilometres (60 miles) or more, dissociated molecules remain dissociated for a day or more and during this long interval may be carried by air movements upward or downward to regions of greatly different photochemical equilibrium. Recombination releases energy in the form of visible light (chemiluminescence) or heat.

Airglow. Visible light is emitted at night in the upper atmosphere by chemiluminescence and by ion–electron recombination reactions. The night glow contributes radiation that is about equal to that from starlight. In the daytime, resonance scattering from sodium, atomic oxygen, nitric oxide, and nitrogen together with chemiluminescence contributes to the airglow. Charged-particle excitation of neutral atoms and molecules also contributes to airglow both day and night. Observations of the airglow have been made from the Earth's surface and also from rockets and satellites. Most of the airglow emanates from the region from 80 to 120 kilometres (50 to 75 miles) above the Earth. Airglow is, of course, difficult to observe during the day.

Ionospheric physics. At wavelengths of 0.1216 micron (Lyman α) or less, absorption of radiation may result in production of an ion–electron pair. The rate of production is proportional to radiation intensity and air density. Radiation intensity decreases at lower levels because of absorption, whereas air density increases downward. Therefore, the rate of ionization is small at very great heights in the atmosphere; it reaches a maximum at a characteristic height and decreases lower down where radiation intensity is low. In this way a layer of maximum concentration of ions or electrons is produced; in fact, separate layers can be produced corresponding to different constituents of the atmosphere. The layers overlap, however, so that there is only one distinct maximum (called the F layer), which occurs at a height of about 300 kilometres (200 miles). Inflections in the concentration curve occur at a height of roughly 150 kilometres or 90 miles (E layer) and 60 to 80 kilometres or 40 to 50 miles (D layer). The processes of photo-ionization and recombination are complex and are different for each of the layers. Ratios of ions or electrons to neutral particles during the day are roughly 10^{-4} for the F layer, 10^{-8} for the E layer, and 10^{-11} for the D layer. The D and E layers virtually disappear at night. The concentration of charged particles is sufficient to make the ionosphere an electrically conducting sphere. Conductivity is strongly influenced by the Earth's magnetic field, however, as well as by air density, ion or electron concentration, and mass of the charged particles. Conductivity consequently is directionally dependent; it is largest in the direction of the Earth's magnetic field and, in the normal directions, depends in complicated ways on frequency of collision, frequency of rotation of charges about magnetic lines of force, and the mobilities of the ions and electrons. Conductivity is largest in the E layer.

Atmospheric tides and electrical conductivity

The periodic motion of the atmospheric tides drives an electrical current in the ionosphere. For the diurnal tide, the total current amounts to about 62,000 amperes, most of it flowing on the sunlit side of the Earth where charge concentration is highest. Calculation of the electrical current for the E layer can be carried out using the equations of electrodynamics with appropriate values for the distribution of the conductivity tensor. The current for the equinoxes for each hemisphere can be visualized as a large vortex centred in middle latitudes with strong current from west to east along the Equator. The induced magnetic field is measurable at the surface of the Earth, providing a means for measuring the current. Current flow in the E layer must induce current flow in the F layer, but details are not known. When solar flares occur, greatly enhanced X-ray radiation results in increased ionization and unusually high concentrations in the D region. Sudden increases in concentration are also produced by meteor trails, which produce metallic ions, and by corpuscular radiation from the Sun.

Electrically conducting media tend to reflect electromagnetic waves. For a vertically directed radio signal, reflection occurs for particle concentration that is proportional to mass of the charged particle. Thus, electrons, by virtue of their small mass, are far more effective in reflecting radio waves than are ions. By varying the radio frequency and timing the transmit time for the reflected signal, the electron density can be scanned up to the height of maximum concentration (F layer). A similar observation made from an Earth-orbiting satellite can scan the electron density down to the F layer. In this way, changes in the ionosphere are observed at many

locations around the world. Above the F layer, concentration of charged particles decreases with height as air density decreases.

Magnetospheric studies. In the vast region far beyond the ionosphere, nearly all particles are electrically charged, and their paths are strongly bound to the geomagnetic field. The equations of electrodynamics require that moving charges spiral around the geomagnetic lines of force, following these lines from the Northern Hemisphere to the Southern as they loop far out into space. In an infinite and empty space the geomagnetic field would be visualized as doughnut-shaped with the Earth at the centre, the strength of the field falling off with the third power of distance. This was, in fact, the concept of the geomagnetic field that prevailed up to the era of space probes and satellites. This simple, symmetrical picture is now known to be changed drastically as a result of the solar wind, the flow of electrons, protons, and other ions outward from the Sun. Under more or less steady conditions, the solar wind flowing at 300 to 700 kilometres (200 to 400 miles) per second interacts with the geomagnetic field. On the side facing the solar wind the field is compressed and a frontal surface develops that is analogous to the bow wave produced by a ship anchored in a flowing stream. Behind this frontal surface is a transition region several Earth radii in thickness in which the solar-wind particles experience increase in energy and in which rapid magnetic fluctuations occur. Inside the transition zone, the boundary of the geomagnetic field, the magnetopause, is reached. The magnetopause forms an appropriate boundary to the atmosphere and an interface between the domains of solar and interplanetary physics and the atmospheric sciences. On the downstream side of the Earth the geomagnetic field is extended like a comet's tail in the solar wind, perhaps as far as 1,000 Earth radii. Thus, the magnetosphere is strongly influenced by the solar wind, and fluctuations in the solar wind produce distortions and pulsations of the geomagnetic field, which in turn affect the motions of charged particles throughout the magnetosphere.

The plasma sheet and magnetic shells of the Earth

Observations show that ions and electrons of a wide range of energies exist in the magnetosphere, but the source of the particles and the mechanism of their acceleration are not fully understood. High-energy particles exist in the plasma sheet, a region extending outward through the centre of the tail of the magnetosphere. In the plasma sheet, ion-pair density is about 0.5 pair per cubic centimetre; electron energy density is about 10^3 electron volts per cubic centimetre; and proton energy density 5×10^3 electron volts per cubic centimetre. Well to the north or south of the plasma sheet, ion-pair density is only about 0.01 pair per cubic centimetre, and energy density is 30 electron volts per cubic centimetre. Under quiet conditions, the plasma sheet extends inward to a distance of about ten Earth radii from the Earth, where it spreads out to join the Earth's magnetic field lines, which finally converge at low elevations in the polar regions. Under disturbed conditions the plasma sheet may extend in as close as five Earth radii.

Within the region extending out to about ten Earth radii, charges are constrained to the lines of force and, therefore, are trapped in the magnetosphere. They are influenced by magnetic effects induced by changes in solar wind and by hydromagnetic and electrostatic waves. Magnetospheric substorms having characteristic periods of about 2½ hours have been observed as gigantic pulsations affecting particles and fields everywhere from the Earth's surface through the magnetosphere far out in the tail. Coupling of the magnetosphere and ionosphere undoubtedly occurs, but the mechanisms are not clear.

The magnetic shells extending outward from the Earth's surface into the magnetosphere act as barriers to particulate radiation from space. The shell that is located at a height of four Earth radii above the geomagnetic equator may be considered as bounding an interior region to which solar particles do not penetrate under undisturbed conditions. Within this region ions and electrons are produced by photo-ionization; outside it they are produced by particulate radiation and by photo-ionization.

Auroral studies. The mechanism of the aurora, the most dramatic and one of the earliest sources of information about upper-atmosphere processes, remains unknown in important respects. Fundamental difficulties exist in understanding how the various auroral phenomena are produced. Present knowledge indicates that auroras are produced when high-energy electrons and protons excite atmospheric atoms and molecules, which then emit characteristic visible radiation. Most of the auroral light is emitted by oxygen and nitrogen. The electrons and protons, constrained by the Earth's magnetic field to spiral around lines of force, are guided down into the denser layers of the atmosphere in the vicinity of the geomagnetic poles. Normally they are reflected back and forth between the hemispheres along lines of force. This process occurs continuously; when the charged particles penetrate in the polar regions to within about 80 to several hundred kilometres above the Earth, sufficient collisions occur to produce visible effects. The normal auroral zone is an oval belt surrounding the geomagnetic poles at an angular radius of about 18°. The magnetic shell at four Earth radii is located at roughly the lower latitude limit of frequent auroral observation, and this also is the region of the often observed stable red arc. Auroras occur at all times of the day, with maximum frequency of occurrence during the first half of the night. This asymmetry of timing probably is related to the asymmetry of the magnetosphere. The plasma sheet provides a channel through which charges can enter the auroral zones, but the required acceleration has not yet been explained. Brilliant displays, which may be visible over large regions, may reflect major distortions of the magnetosphere, or they may result from acceleration of charges by mechanisms that are still obscure.

Occurrence of the auroras producing visible effects

Auroras occur in several forms, descriptively characterized as draperies, arcs, rays, and corona. Perhaps the most spectacular is the appearance of several parallel sheets or draperies extending east to west from horizon to horizon, changing over an interval of minutes in form, intensity, and colour. The draperies are formed of narrow rays parallel to the geomagnetic lines of force, typically visible between 90 and 250 kilometres (60 and 150 miles) above the Earth. In the auroral zone the rays have an angle of 10° to 15° to the vertical.

PLANETARY ATMOSPHERES

Comparison of the atmospheres of the planets of the solar system provides a unique guide to understanding the present state and the evolution of the solar system and of the Earth's atmosphere. Although the Earth's atmosphere is much more accessible to observation, certain atmospheric processes occurring on other planets may be simpler than those occurring on Earth. Direct measurements of the atmospheres of Venus and Mars have been carried out by Soviet space probes of the Venera and Mars series, while the Mariner vehicles launched by the United States have employed remote sensing techniques to obtain similar knowledge of these two planets and Mercury. The Pioneer probes of the United States have transmitted information about the atmosphere of Jupiter. Pioneer 11 revealed a polar icecap on Callisto, one of the satellites of Jupiter, as it passed by in late 1974; its path was expected to carry it close to Saturn in September 1979. Mercury appears to have no atmosphere. Jupiter, Saturn, Uranus, and Neptune have atmospheres of hydrogen, helium, methane, and ammonia, but the atmospheres of those planets are not clearly distinguished from the material of the underlying planets.

Carbon dioxide makes up more than 90 percent of the atmospheres of Mars and Venus; traces of other gases have been detected. Water vapour appears to amount to only about 0.01 percent on Mars and even less on Venus. Pressure of the Venus atmosphere at the surface is about 100 Earth atmospheres, whereas pressure of the Mars atmosphere is only 0.5 percent of that of the Earth atmosphere.

Magnetic fields around Mars and Venus are very weak, so that there are no magnetospheres shielding the planets from the solar wind. Ionospheres formed by ionization

of helium, however, provide shields much closer to the planet's surfaces than is the case on Earth. Each planet exhibits maximum ionization at heights of from 100 to 150 kilometres (60 to 90 miles).

Ultraviolet radiation, which is absorbed in the ozone layer of the Earth's atmosphere, penetrates to the Mars surface; consequently, terrestrial forms of life could not be expected to develop or to survive in that environment. The Venus clouds, which cover the planet at a height of 60 to 70 kilometres (40 to 45 miles), shield the surface from ultraviolet radiation. The composition of those clouds remains uncertain; ice and mercurous chloride have been proposed, but there are difficulties with either hypothesis.

The vertical temperature distributions for Venus and Mars are qualitatively similar to the distribution in the Earth's atmosphere, except that the Earth's warm ozone layer is absent from the atmospheres of Mars and Venus. In each case the thermosphere is evident above about 100 kilometres (60 miles) for Mars and 130 kilometres (80 miles) for Venus. The tropospheres are also evident, but there are no layers corresponding to the Earth's stratosphere or mesosphere.

Radiating temperatures and atmospheric circulation

The effective radiating temperature (based on the Stefan-Boltzmann relationship between emitted radiation and temperature) for Mars is 217° K (the zero point on the Kelvin temperature scale is absolute zero, or − 273° C, − 460° F), for Venus it is 243° K, and for Earth it is 250° K. The corresponding values for Mercury and Jupiter are, respectively, 443° K and 89° K. The diurnal range of surface temperature on the Martian equator is from 300° K in the daytime to 170° K at night, an enormous range in comparison to the diurnal range on Earth. The surface temperature on Venus cannot be determined from measurements of infrared radiation because of the obscuring clouds, but microwave observations indicate a surface temperature in the neighborhood of 700° K, and this is consistent with extrapolation of the Venera observations in the upper troposphere down to the surface.

The atmospheric circulations of Mars and Venus differ markedly because of the fact that Venus rotates only once in 243 days whereas Mars rotates at close to the rotation rate for the Earth. The Mariner 9 observations (1971 and 1972) have provided a wealth of information about the Mars atmosphere. Understanding of the circulation of Mars and Venus, however, is still largely derived from calculations using dynamic models rather than from direct observations; consequently, the following account must remain uncertain until direct observations are available. The dominant motion on Venus is thought to be that of a single cell with rising fluid at the subsolar point and sinking fluid at the antisolar point. This circulation creates the near adiabatic lapse rate in the Venus troposphere, and it provides an efficient mechanism for maintaining a horizontally uniform distribution of temperature. In the upper atmosphere, transport of ions and electrons by the circulation also accounts for the existence of the nighttime ionosphere. Circulation on Mars is much more like that on Earth because of similarities in the rotation rates and in the meridional gradients of solar heating. Because oceans are absent on Mars, the temperature gradients are larger, and the zonal wind velocity and the shear associated with baroclinic disturbances should be larger than on Earth. The large diurnal temperature change should result in an enhanced thermal tide. The extreme relief of the Martian topography must also affect circulation significantly.

APPLICATIONS

Much research effort in the atmospheric sciences is closely linked to applications of immediate importance. Research into the dynamics of storms is directed toward improving the ability to predict these events and thus to avoid damage and loss of life. Research in atmospheric chemistry is motivated by the need to understand and to specify the effectiveness of the atmosphere in dispersing and precipitating air pollution in a safe manner. And research in cloud physics is concerned with understanding processes in clouds so that they might be modified to produce more rainfall or to reduce the severity of storms. These examples illustrate that atmospheric research is often both basic, in the sense that it concerns the fundamental understanding of atmospheric processes, and applied, in the sense that it is directly useful. The distinction is more a reflection of individual interest and motivation than of the scientific nature of the problem or the mode of research.

Predictions and forecasts. The central function of the atmospheric sciences is to predict changes in the atmosphere within useful limits of accuracy. It would be most desirable to predict changes of the full range of scales which include tornadoes, middle latitude storms, droughts, and climate changes. Efforts to predict atmospheric changes, however, are for the most part concerned with forecast periods from about 12 hours to two or three days. The lower limit of 12 hours is a consequence of the fact that turbulence, vertical convection, and other microscale and mesoscale phenomena are strongly nonlinear and their physics so complex that simplifying approximations are invalid. The upper prediction limit is a consequence of lack of data covering large areas of the Earth and failure to represent the physics of atmospheric processes with sufficient accuracy. For periods of 12 hours to two or three days predictions are made routinely using the equations of large-scale atmospheric motions. These equations are transformed to finite-difference form and applied to a three-dimensional space grid. The model in current use by the National Meteorological Center (NMC) of the National Oceanic and Atmospheric Administration (NOAA) in the United States requires that initial values of the dependent variables be specified at grid points for six layers in the free atmosphere between the top of the atmosphere and the top of the boundary layer at a pressure of 950 millibars. An additional surface layer is introduced between 950 and 1,000 millibars (approximate sea level) in which the effects of turbulence are included. The grid is centred at the North Pole and is artificially bounded near the Equator, with the grid points approximately 380 kilometres apart. Other prediction models are in use in Australia, Canada, China, East Germany, France, Italy, Japan, Norway, Sweden, United Kingdom, Soviet Union, West Germany, and other countries.

Three-dimensional prediction grid

The initial data determine the values of all terms in the equations at each grid point, except the local derivatives with respect to time; therefore, changes in the dependent variables for small intervals of time (ten minutes in the current model) can be computed at each grid point. This provides a new set of data that serves to compute a new set of changes. By repeating this procedure, forecasts for the entire globe could be made in principle for as far in the future as desired. In fact, the forecast period is limited by (1) the existence of unavoidable errors in the initial data, (2) errors introduced by the finite-difference representation of continuous functions, and (3) features smaller than the grid intervals that cannot be represented by the model, as well as by lack of global data. These effects grow during the forecast period, and experiment has shown that they become as large as differences between randomly selected states of the system after about two weeks of forecast time. Thus, the limit for deterministic prediction is in the neighbourhood of two weeks; this result provides strong incentive and justification to the nations of the world to obtain complete global data. In response to a resolution of the United Nations General Assembly, the International Council of Scientific Unions and the World Meteorological Organization have undertaken to develop a system of global observations and to carry out a series of experiments to extend the range of forecasting as nearly as practicable to the theoretical limit.

Forecast models have become steadily more complex and more accurate as understanding of atmospheric dynamics has increased and computer speed and capacity has grown. The 36-hour NMC "skill score," which is a rough measure of the error of the wind forecast, has decreased over the past 15 years as each major model im-

provement has been introduced. The skill score for the current 500-millibar model is 30 percent below the skill score achieved prior to introduction of numerical prediction models in 1958. The model in use since 1966 represents, at least crudely, boundary layer and cloud processes; these steps have resulted in extension of routine forecasts since 1966 from 48 to 72 hours.

Decreasing error in wind forecasting

Operational weather forecasts have improved significantly and steadily over the past half century or more. The improvements are attributable to the effects of better basic data, improved communications, and more competent manpower, as well as to development of numerical models. Improvement in forecasts of one type of severe storm, the hurricane, has resulted in greatly reducing the loss of life in the United States from these storms since the early part of the 20th century. At the same time, however, property loss has increased because population and industries have grown along exposed coasts.

Forecasts for specific local areas utilize results of model calculations as a basis and add other considerations that must be treated more subjectively, terrain and surface effects or subgrid scale variability, for example. Roughly half of the variability of surface weather is contained in such subgrid scale features as thunderstorms, fronts, and sea breezes. This variability is superimposed on the variability of the numerical model prediction and therefore results in degrading the model forecasts for specific small areas and time periods. It is difficult to make valid comparisons of specific, local forecasts on a long-time basis. The limited data available indicate small but significant improvement in temperature and precipitation forecasts since numerical prediction has been used operationally.

For time periods longer than about two weeks, the mechanisms that produce changes in weather and climate are not understood well enough to make useful predictions, and, as noted earlier, numerical experiments indicate that the approach through deterministic prediction of specific weather states cannot be successful. There are many hypotheses of the causes of long-term variation of weather and climate, including: (1) some long-term variation in weather must be inherent in an atmosphere with fixed boundary conditions and composition, but how much is not known; (2) the atmosphere and the ocean have greatly different response times; as a result, when they are treated as a coupled fluid mechanical and thermal system, long-term fluctuations may occur; (3) it is plausible that changes in surface reflectivity accompanying the expansion of agriculture or other surface changes may have influenced climate to an important degree; and (4) past or future changes in atmospheric carbon dioxide concentration or in particulate concentration may change the radiation balance of the atmosphere and bring about changes in climate.

Statistics from models of the global atmosphere

After running models of the global atmosphere for simulated periods of several years, climatological statistics can be compiled from the sum of all the individual weather states. Although the separate daily predictions are not accurate, the statistics closely simulate present climate. This conformity gives a degree of confidence in the models and suggests that they might be used to identify the principal causes of climatic change. Experiments using arbitrary changes in carbon dioxide concentration have been conducted; these constitute the first early steps in predicting possible future climate. Other experiments can be carried out to test other hypotheses, but the physical subtleties are very great and the computer requirements severe.

Weather modification. The understanding of atmospheric processes, which is essential for successful prediction, also affords opportunities to modify weather deliberately for useful purposes and may make it possible to anticipate inadvertent, deleterious effects before it is too late to avoid them. Operational programs of weather modification have been carried out, but most of them have been exploratory in character. Weather modification so far has been in large part a research field, but it appears likely to become increasingly important as an applied field. It is pertinent to note also that as technical and scientific problems are solved, the associated public policy problems are looming larger. The most important question is not "Can precipitation be increased?" but "What programs of weather modification are in the public interest?"

Efforts to modify weather deliberately have been directed largely at cloud seeding to increase precipitation, mostly by introducing ice nuclei into supercooled clouds. Many programs have been poorly designed or executed, and only rarely have their effects been adequately observed or reported. Effects, even when observed carefully, are usually well within the natural variability of clouds. It is not surprising, therefore, that the results of cloud seeding have been regarded as controversial and uncertain. Sufficient reliable statistics have been accumulated to indicate that the results of individual cloud-seeding experiments vary depending upon weather situations and mode of seeding. Randomized experiments indicate that in some situations increases of precipitation of 10 to 30 percent have been achieved, whereas, in other situations, corresponding decreases have occurred. There is only limited understanding of the factors that distinguish positive from negative effects.

Role of latent heat in cloud seeding

The latent heat released when large numbers of ice crystals are formed and grow in a supercooled cloud may result in enhancing the updraft in a cumulus cloud. This may increase the height of the cloud, increase the total moist air flowing into the cloud, and therefore increase precipitation considerably. Numerical models have been developed to predict cloud growth and to aid in design of field experiments. In a series of experiments in Florida, it was found that three times as much rain fell to the ground from seeded single clouds as from unseeded clouds.

The dynamic effects of the release of latent heat by adding ice nuclei to supercooled clouds may also be useful in modifying the pressure field and in reducing the wind speed in hurricanes. This possibility has been examined by model calculations, indicating that significant results might result from seeding. A field test of Hurricane Debbie in August 1969 was followed by reductions of 15 to 30 percent in maximum wind speeds after silver-iodide seeding. This result suggests that it may be possible to reduce the damage inflicted by hurricanes and possibly to influence the path they follow, but further investigations are needed to establish this possibility and to determine how best to achieve the desired results.

Other examples of deliberate weather modification include hail suppression, which is carried on operationally at one site in the Soviet Union, and cold fog dissipation, which is an operational practice at several United States airports.

Inadvertent effects on weather and climate may in the long run be more important than deliberate effects. Research effort is just beginning to be directed toward long-range problems. Climate simulation models, which include a few of the interactive effects of the atmosphere and the underlying Earth and ocean, have indicated that the expected increase of 18 percent in atmospheric carbon dioxide concentration by the year 2000 might result in a surface temperature increase of ½ ° C. Because this is comparable to the natural variability of temperature over a period of 30 years, it appears that the climatic effects of increased carbon dioxide are not likely to be critical in the next few decades, but that over longer periods they may become very important.

Another example of a potentially important inadvertent effect is provided by major sources of air pollution. The acidity of rainfall in western Europe has increased markedly in recent decades because of the rising volume of effluents from industrial regions, largely sulfur dioxide. Adverse effects on the forests of Sweden have already been reported; and, as industrialization proceeds in many areas of the world, this example must be expected to be multiplied many times over.

Air pollution

Management of air quality. The problems posed by air pollution, already serious, are likely to become more so in the future; they can be managed only by the joint

efforts of chemists, biologists, engineers, medical doctors, and atmospheric scientists. The work of these professionals must be carried on within constraints of law, economics, traditional practice, and public acceptability. About 98 percent of world energy is produced by combustion of fossil fuels: coal, petroleum, and natural gas. Nuclear and hydroelectric sources account for only 2 percent of energy production. This ratio will remain the same until after 1980, even though world energy production is expected to nearly double by that time. Most of the waste products are injected into the atmosphere for ultimate disposal, and no alternative to this mode is feasible in the foreseeable future. In addition to the products of fossil fuels, radioactive gases and particles also may be introduced into the atmosphere. There are also natural pollutants that are of concern: aeroallergens, spores, fungi, agricultural and desert dust, and volcanic materials.

The specific fate of pollutants and their effects on man, animals, plants, and materials are strongly dependent on atmospheric processes; consequently, management of air quality must be based on the ability to predict the state of the atmosphere. The atmospheric scientist's responsibility, therefore, is to specify the capacity of the atmosphere to store, disperse, modify, and deposit pollutant materials.

Materials are emitted into the atmosphere as particles of various sizes and as gases. Particle size and chemical composition at the source can be controlled technologically to a limited extent, but the mass of effluent usually cannot be modified. Once the particulates and gases have been emitted into the atmosphere, they are transported horizontally by wind and mixed vertically by turbulence; chemical and photochemical changes occur; new particles are formed from gaseous pollutants; and rain or snow may wash out some of the particles and gases. These processes are seldom known in detail. The common pollutants sulfur dioxide and carbon monoxide evidently are transformed chemically within the atmosphere, but their histories are not understood, and efforts to identify how they are removed from the atmosphere have so far been unsuccessful. Understanding of the effects of solar radiation, humidity, and turbulent mixing in the chemical reactions that lead to urban smog are in a primitive state. So research on the chemistry of air pollution under various natural conditions is needed to make possible accurate specification of pollutant concentrations under the range of conditions for which they are needed.

The complexity of the problem requires the use of numerical models; the needed computers must be capable of storing large numbers of data and making many intricate calculations quickly. Beginnings have been made in constructing models that predict concentrations in certain urban areas, but they do not incorporate chemical changes or other sophisticated aspects of turbulence, cloud physics, radiation, and atmospheric dynamics.

In the United States urban pollution monitoring systems constitute an important part of the Air Pollution Control Meteorological Service under the Environmental Protection Agency. Together with data from the Basic Meteorological Service of the National Oceanic and Atmospheric Administration (NOAA) and, in some cases, from state agencies, these stations provide the data that determine safe limits for emission of air pollution by industry, homes, or motor vehicles. Dangerous levels of air pollution are especially likely to be encountered in valleys or in bowl-shaped topographic regions under conditions of low wind speed and strong surface temperature inversions. In such areas, where pollution is confined to the surface layer below the inversion, if emission continues, pollution concentrations may rise to lethal levels.

Information services. Atmospheric data serve a variety of additional specific uses. Sometimes the data alone are provided, or, in other cases, interpretation also is provided. In most of the developed countries, such services are available from governmental agencies or from private professionals employed for their specialized services; organizational details are unique to each country.

Specialized weather reports

Continuous, comprehensive weather information and forecasts are essential to airlines, private pilots, and terminal operators to ensure safe flying and to maintain schedules. This information is provided in the United States by the Aviation Meteorological Service of NOAA. Coastal and open-ocean shipping and small boat operators rely on weather reports and storm warnings issued by the Marine Meteorological Service. Specialized storm surge and tsunami (seismically generated sea waves) warnings are issued to the public in vulnerable low-lying coastal areas.

Decisions on what species of crops to plant and of when and where to plant are contingent on climatological information which is provided by the Agriculture Meteorological Service. The National Climate Center provides data useful for planning in many fields in addition to agriculture: architecture, highway construction and maintenance, truck operations, food distribution, and many others.

Radio communication over long distances on Earth is made possible by reflection of the radio signals from the underside of the ionosphere. Ionosphere monitoring stations are operated by many nations, and exchange of information occurs through World Data Centers in the United States, Soviet Union, and other countries. Scientific groups under the International Council of Scientific Unions (ICSU) sponsor joint research programs concerned with the ionosphere. In the United States ionospheric information is provided by the Aeronomy and Space Data Center of NOAA in order to make possible efficient, reliable radio communication. Safety in space operations depends on reliable radio communication and also in avoiding highly dangerous electromagnetic and particulate radiation, which may accompany solar flares or other eruptions on the surface of the Sun. These necessary data are provided by the Space Operations Meteorological Service. The Committee on Space Research of ICSU maintains an international roster of objects launched into space.

TECHNOLOGY OF THE ATMOSPHERIC SCIENCES

The atmospheric sciences are heavily dependent on sophisticated technology. Most of the scientific advances of the past three decades are directly related to developments in technology, and many of the critical problems facing atmospheric scientists will require further technological advances. Satellites have made possible for the first time observation of the entire Earth. The first experimental meteorological satellite was launched in 1960; it was capable of taking pictures of limited areas of clouds with low resolution. Present-day satellites provide far more data of better quality and greater usefulness. Stationary satellites situated above the Equator provide color pictures of cloud distribution at time intervals of 20 minutes each covering a circular area 5,500 kilometres (3,400 miles) in radius. Polar orbiting satellites scan the entire Earth each day, providing daytime and nighttime cloud images with high resolution. Even more important, infrared and microwave spectrometers flown on satellites can provide the vertical distributions of temperature and humidity for the entire globe.

Satellites and remote sensing

The vertical distributions of density and electrical charge can be measured by optical and radio occultation measurements. Other satellite instrumentation measures the incident and reflected solar and terrestrial radiation, the characteristics of the ionosphere, the airglow, the geomagnetic field, flux of ions and electrons, and other properties. Related instrumentation on space probes is used to observe planetary atmospheres and interplanetary space.

Other remote-sensing instruments have been developed for use on aircraft and on the ground. A review of remote-sensing technology for atmospheric measurements was published in 1969 by the National Academy of Sciences, Washington, D.C. Doppler radar is useful in examining the turbulent motions occurring in precipitation clouds. In clear air, small-scale temperature and humidity structure can be detected and tracked using highly sensitive microwave radar. With optical radar (lidar) equip-

ment the distribution of water vapour and aerosol can be measured; and acoustic echo-sounding may be applied to remote measurement of wind and turbulence by using natural small-scale temperature fluctuations as tracers. Spectrophotometric techniques can probably be used to detect and measure air pollution. Upper-atmosphere properties are measured by a variety of ground-based remote-sensing instruments; for example, radio waves are used to measure ionospheric properties and magnetospheric temperatures, and, by employing a correlation technique, wind velocity and turbulence in specified remote volumes can be measured. With incoherent scatter radar many properties of electrons, ions, and electrically neutral gas in the ionosphere and magnetosphere can be directly determined.

The great numbers of data provided by modern technology could not be put to profitable use if it were not for the development of high-speed, high-capacity computers. They are essential for processing raw data transmitted from satellite sensors, for organizing and analyzing data, and for carrying out numerical predictions. For operational use the computer time used in making a prediction must be considerably less than the forecast interval; this requirement limits the physical complexity of the operational models in current use. More sophisticated operational and research models require computers of the highest capacity and speed contemplated at this time.

Atmospheric data is exchanged internationally through the Global Telecommunications System, which utilizes satellite communications, radio, and land lines. World centres in Washington, Moscow, and Melbourne collect and store world data and provide two-way access to Regional Telecommunication Hubs in Tokyo, Brasília, Bracknell (London), Paris, Offenbach (Frankfurt am Main), Prague, Cairo, Nairobi, and New Delhi. The United States Automatic Picture Transmission (APT) satellites provide direct readout of cloud images to surface receiving stations all over the world. The World Meteorological Organization with headquarters in Geneva, Switzerland, provides an effective channel for exchange of data and multigovernmental agreements on matters concerning the atmosphere.

BIBLIOGRAPHY. General references on the varied topics that comprise the atmospheric sciences include the following works: M.I. BUDYKO, *The Heat Balance of the Earth's Surface* (1958; orig. pub. in Russian, 1956), a summary and interpretation of the great quantity of climatological data needed to describe the heat and vapour transfer distributions at the earth's surface; J.W. CHAMBERLAIN, *Physics of the Aurora and Airglow* (1961), theoretical background, observations, and interpretation as known in 1961, with emphasis on the processes of light emission in the upper atmosphere; COMMITTEE ON ATMOSPHERIC SCIENCES, *Atmospheric Exploration by Remote Probes*, 2 vol. (1969), an appraisal of the potential of remote probing techniques for atmospheric observations; R.A. CRAIG, *The Upper Atmosphere: Meteorology and Physics* (1965), an introduction to the study of the upper atmosphere with emphasis on the structure, composition, circulation, and interactions; R.G. FLEAGLE and J.A. BUSINGER, *An Introduction to Atmospheric Physics* (1963), an introduction at an intermediate level to physical processes that determine atmospheric structure, energy transfer, and transmission of electromagnetic and sound waves; N.H. FLETCHER, *The Physics of Rainclouds* (1962), a thorough summary of the microphysical processes of clouds; J.R. HOLTON, *An Introduction to Dynamic Meteorology* (1973), an introduction at an intermediate level of large-scale atmospheric dynamics; C.E. JUNGE, *Air Chemistry and Radioactivity* (1963), an organized summary of the observations with a critical discussion of areas of uncertainty; K.Y. KONDRATYEV, *Radiation in the Atmosphere* (1969), a comprehensive account of physical processes associated with radiation and their importance in determining large-scale atmospheric structure; J.L. LUMLEY and H.A. PANOFSKY, *The Structure of Atmospheric Turbulence* (1964), a summary of the mechanics and statistics of turbulence and their relation to transport and diffusion problems; J.M. MITCHELL, JR. (ed.), *Causes of Climatic Change* (1968), a group of papers that reviews understanding of past changes in climate; E. PALMEN and C.W. NEWTON, *Atmospheric Circulation Systems* (1969), a comprehensive account of circulation systems based on an analysis of upper air and surface observations; H.A. PANOFSKY and G.W. BRIER, *Some Applications of Statistics to Meteorology* (1958), a concise account of some of the important applications of statistics to atmospheric analysis; W.D. SELLERS, *Physical Climatology* (1965), an organized account of the physical basis of climate; H. RISHBETH and O.K. GARRIOTT, *Introduction to Ionospheric Physics* (1969), an organized and clear account of the physical processes occurring in the ionosphere; SPACE SCIENCE BOARD—H. FRIEDMAN and F.S. JOHNSON (eds.), *Physics of the Earth in Space: A Program of Research, 1968–1975* (1968), a brief summary of insights and understanding resulting from the space exploration of the 1960s; R.C. WHITTEN and I.G. POPPOFF, *Fundamentals of Aeronomy* (1971), theoretical background, observations, and interpretation of aeronomy with emphasis on photochemical processes and physics of the ionosphere.

(R.G.F.)

Atomic Structure

The atom is defined as the smallest unit of a chemical element that retains its elemental identity. It consists, according to present knowledge, of a tiny but massive central core, called the nucleus, surrounded by a number of electrons. The arrangement and behaviour of these electrons determine the interaction of the atoms with one another and thus govern not only chemical processes but also most of the physical properties of bulk matter. Such arrangement and behaviour are what is meant by atomic structure, and they form the subject of this article.

The article follows this outline:

General considerations
- Nature of the atom
- Atomic magnitudes
- Atomic structure and interactions

Development of atomic concepts
- Early philosophical speculations
- Origins of the atomic theory in chemistry
- Spectroscopy and spectral regularities
- Light and related radiations
- Discovery of the electron
- Discovery of X-rays and radioactivity
- Discovery of the existence of isotopes

Models of atomic structure
- Early atomic models
- The Rutherford model
- The Bohr–Sommerfeld atom
- Contemporary atomic models

The origin of atomic spectra

Atomic structure and molecules
- The nature of forces between atoms
- Formation of a molecular bond
- The origin of molecular spectra

Atomic structure and bulk matter
- The nature of bonding in solids
- Energy bands in the solid state
- Optical behaviour
- Magnetism

GENERAL CONSIDERATIONS

Nature of the atom. There exist a limited number of substances of a special class, called elements, that cannot be reduced or separated into chemically different components. A few examples would include the metals iron, aluminum, copper, and uranium; the gases oxygen, nitrogen, helium, and hydrogen; and the nonmetallic solids carbon, silicon, and sulfur. When elements are chemically bound together, they form compounds. Water, ammonia, and quartz are compounds of the elements; steel, brass, whiskey, and air are mixtures of elements and compounds. About 90 different elements occur in nature, and about 20 more have been produced artificially. The ultimate particles of these elements are the atoms, and thus there are as many basic kinds of atoms as there are elements.

Atoms are not indivisible but have a structure consisting of two kinds of particles, one with a positive electrical charge and the other with a negative. Some of the negatively charged particles are loosely bound in the atom and are easy to dislodge. Heating a wire, for example, releases negative particles. Regardless of what substance they come from or how they are set free, all are identical and have been given the name electrons.

An atom (and thus all matter) is mostly empty space. At the centre is a positively charged core, the nucleus,

which accounts for 99.95 percent of the mass of the atom, although it occupies only about 10^{-15} of the volume. The electrons are scattered through the remaining volume.

Atomic magnitudes. Because a molecule is a close association of atoms, the size of a molecule containing only a few atoms is not very much greater than that of an individual atom. The determination of molecular sizes is a useful step toward finding the sizes of individual atoms, and historically this was the way the problem was approached. The first quantitative estimates were obtained by the English scientist Thomas Young in the early 19th century and presented by him in the *Encyclopædia Britannica Supplement* of 1816. Young based his conclusions on a study of surface tension and tensile strength of liquids. Although his conception of the relation between these two effects was substantially correct, his ignorance of the intimate structure of liquids and vapours led him to false conclusions. The first truly significant values of molecular size came from the study of the diffusion, or mixing, of gases.

Early estimates of atomic size

One of the simplest and most direct deductions from the kinetic theory was that the molecules of gases must have high speeds, 1,000 feet per second more or less. Seemingly opposing this deduction, however, was the fact that if a flask of, say, gaseous carbon dioxide is opened to the air, no great amount of mixing of the two gases occurs over a period of hours. These results could be reconciled by ascribing to the molecules a diameter of such a size that any one molecule (at least in gas at normal atmospheric pressure) would travel only a very small distance before undergoing a collision with another molecule. In this way the movement or diffusion of the gas as a whole in a particular direction would be greatly discouraged. From the diffusive properties it is possible to estimate the average distance between successive molecular collisions. But this quantity, called the mean free path and denoted by the Greek letter λ, is determined from kinetic theory by the molecular diameter (d) and the number of molecules in a unit volume (n) according to the formula $\lambda = 1/(\sqrt{2}\pi nd^2)$. If, now, a volume V of gas or vapour, containing Vn molecules, condenses into liquid form so that the molecules are assumed to be densely packed, the volume v of condensed material cannot be less than $\pi Vnd^3/6$, because $\pi d^3/6$ is the volume of a spherical molecule of diameter d. A more realistic estimate would take simply Vnd^3, because the packing of equal spheres does not fill space completely. Multiplying v and λ together gives $v\lambda = Vd/\sqrt{2}\pi$, or $d = \sqrt{2}\pi\lambda(v/V)$. Experimental results show that for many gases v is equal to $0.005V$ approximately (*i.e.*, a given volume V of a gas at atmospheric pressure condenses to about $0.005V$ when converted to the liquid form of the substance) and that $\lambda = 2 \times 10^{-6}$ centimetre approximately. This would mean that the diameter, d, of a molecule is roughly equal to 5×10^{-8} centimetre.

Because of its extraordinary simplicity, a method of estimating molecular dimensions, developed in 1890 by physicists W.C. Röntgen of Germany and Lord Rayleigh of Great Britain, deserves mention. It consists of depositing a tiny drop of oil on a water surface; the assumption is that the oil can spread until it forms a layer only one molecule thick. The limits of the oil patch can be seen by first sprinkling the water surface with a fine powder, such as chalk dust or lycopodium powder. If the volume of the oil drop is V and the area of the resulting patch (*i.e.*, the film thickness) is A, then the molecular dimension is volume divided by the area, V/A.

Other methods give much the same results. Clearly, a convenient unit to use for lengths on the atomic scale is 10^{-8} centimetre—called the angstrom and abbreviated Å. It is a rather remarkable fact that all atoms have about the same diameter, roughly one angstrom.

Atomic weights

The relative weights of different kinds of atoms, which lead to the concept of atomic weights, can be determined easily, provided chemical formulas are known. Thus, water has the formula H_2O, which means that two atoms of hydrogen (H) combine with one atom of oxygen (O) to form a molecule of water. When the water is decomposed, as by electrolysis (the passage of an electric current through an electrolyte, in this case water), the weight of the oxygen obtained is eight times that of the hydrogen. This difference means, there being twice as many atoms of hydrogen as of oxygen, that an atom of oxygen weighs 16 times as much as an atom of hydrogen. Similar observations for other compounds give the relative masses of most of the elements. (In addition, the requirement that all these experiments give self-consistent results provides a check on the assumed formulas.) It is then only necessary to choose one kind of atom as a standard. The choice currently accepted as standard by physicists is that of the most abundant isotope (atoms of an element having different mass are isotopes) of carbon, which is taken to have a mass value of exactly 12; some typical masses on this scale are given in the Table (see also ISOTOPES; ATOMIC WEIGHT).

The quantity of an element having a mass in grams equal to its atomic weight is called a gram atom. The number of atoms in a gram atom is the same for all elements and is called Avogadro's number. This number can be determined in several ways and has the value 6.025×10^{23} per gram atom. With this number, it is possible to calculate the mass represented by one unit on the scale described above; it turns out to be 1.6604×10^{-24} gram. This figure is also, it will be noted, approximately the mass of a hydrogen atom; the heaviest atom yet produced, early in the 1970s, has a mass of 4.2×10^{-22} gram, still an unimaginably small number. The mass of the electron can be determined from the nature of its response to electric and magnetic forces and has a value a little more than 1/2,000 of the mass of the hydrogen atom; its precise value is 9.106×10^{-28} gram.

The electron's charge

The electric charge on an electron has been measured directly and can also be deduced, along with other constants, from several kinds of experiments. Its value is 1.6×10^{-19} coulomb. This value means that in a current of one ampere—roughly what a 100-watt light bulb uses in the ordinary 110-volt household circuit—about 6×10^{18} electrons pass through the wire every second. It turns out that the charge (e) carried by an electron is a natural unit of electric charge in atomic and nuclear physics. Every particle yet discovered, if it has a charge different from zero, has a charge that is a whole multiple of this unit, although sometimes positive rather than negative. (A class of subatomic particles called quarks, with charges $+\frac{2}{3}e$ and $-\frac{1}{3}e$, has been proposed on a theoretical basis, but their existence is still doubtful.) In these units the charge on the nucleus of an atom is equal to the atomic number, a number indicating the position of the element in the periodic table. Because the normal atom is electrically neutral, this positive charge must be just balanced by the negative charges of the electrons, so that the number of electrons in the normal atom is also given by the atomic number of the element. (If one or more electrons are added to or removed from the system, it acquires a net electric charge and is called an ion. Its elemental identity does not change, however.)

Energy is as important in the atomic realm as it is in everyday life. On the human scale, energy is often measured in kilowatt-hours. A kilowatt-hour is the energy consumed by a 100-watt light bulb in ten hours. It is also

Atomic Numbers (Z), Atomic Mass Numbers (A), and Masses (Atomic Weights) of Some Elements

	Z	A	mass
Neutron	0	1	1.00866522(6)
Hydrogen	1	1	1.00782522(4)
		2	2.01410222(7)
		3	3.01604972(16)
Carbon	6	12	12.00000000
		13	13.00335508(23)
		14	14.00324202(30)
Oxygen	8	16	15.99491502(20)
		17	16.9991333(10)
		18	17.9915996(29)
Iron	26	54	53.9396120(29)
		56	55.9349339(26)
		57	56.9353907(26)
		58	57.9332745(27)
Tin	50	112	111.904834(8)
		114	113.902776(7)
		115	114.903353(7)
		116	115.9017483(34)
		117	116.9029606(21)
		118	117.9016126(21)
		119	118.9033159(21)
		120	119.9022073(22)
		122	121.9034511(33)
		124	123.905283(4)
Bismuth	83	209	208.980401(7)
Uranium	92	234	234.040975(11)
		235	235.043944(11)
		238	238.050816(11)
Nobelium	102	257	257.096885(34)

the energy needed to lift a standard automobile, with a mass of about 1,500 kilograms (3,307 pounds), to a vertical height of about 200 metres—100 feet higher than 555-foot-high Washington Monument. On the atomic scale, energies are more conveniently expressed in electron volts. There are 2.26×10^{25} electron volts in one kilowatt-hour. To remove an electron from an atom requires an energy of the order of a few electron volts, the exact value depending on which atom is being dealt with. This value is also the typical amount of energy required or released in chemical reactions such as burning. Thus, because a gram of an element contains about 10^{23} atoms, the burning of that gram of matter yields roughly 10^{23} electron volts of energy, or roughly 1/200 of a kilowatt-hour.

Atomic structure and interactions. Because the atom is so much smaller than anything familiar to ordinary experience, it is not too surprising that physics on the atomic scale is different from the physics derived from experiments with larger objects. In the latter case, the laws of physics are an approximation that is more than adequate for objects on the everyday scale but that fails when atomic dimensions are involved; not only the laws of classical physics, however, but also some daily-life concepts are no longer valid on the atomic scale. This failure of classical physics has a peculiar consequence: if all the properties of atoms and their components as deduced from modern experimental investigations are considered, it is not possible to combine them into a pictorial representation of the atom. Nevertheless, pictures supposedly representing atoms are often seen. Probably the most familiar is one representing the atom as a sort of planetary system, with the nucleus at the centre and paths representing orbits of the surrounding electrons, as in the upper part of Figure 1. Another shows the electrons not in orbits but "smeared out" as clouds of various regular shapes that surround and hide the nucleus, as shown in the lower part of Figure 1. In fact, however, each of these pictures is only an approximation, even a caricature, because each picture overemphasizes some characteristics of the atom while suppressing others. The planetary picture stresses the properties of a single, isolated atom, especially its electrical and magnetic properties, but gives practically no clue to its interaction with other atoms. The cloud picture, on the other hand, can be used to advantage in helping to understand how atoms interact and form molecules but says little about other properties. Despite the fact that the structure of the atom is so hard to render pictorially, it is the basis of all differences as well as similarities in the properties of elements and ultimately of the behaviour of molecules and bulk matter.

Representations of atomic structure

By courtesy of H.E. White

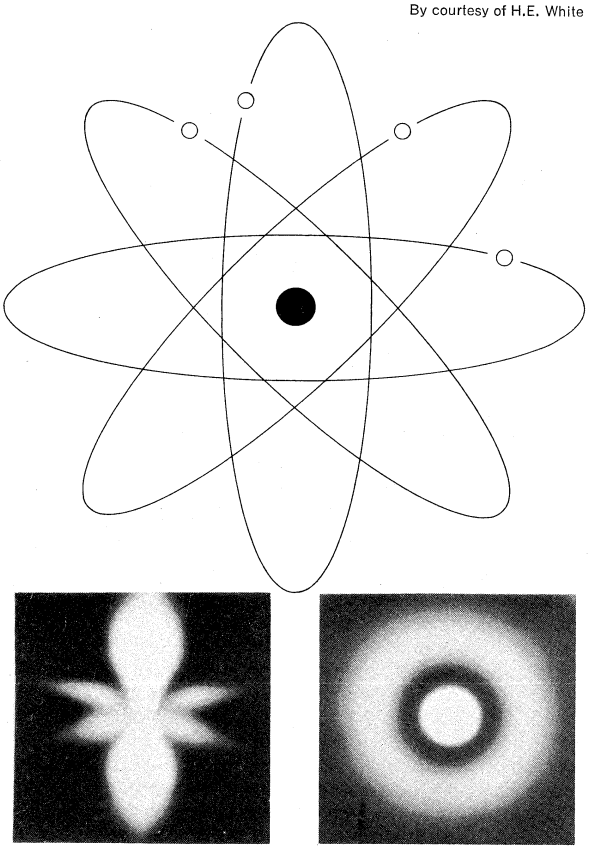


Figure 1: *Typical representations of atomic structure.*
(Top) The "planetary" or "orbital" model of atomic structure. (Bottom) Electron "clouds" for two representative states; note the highly directional features on left, and their absence on right.

In the neutral atom, the electrons are held to the nucleus by the electrostatic forces of attraction resulting from the opposite charges they carry. Particles with similar charges—*e.g.*, the electrons—repel one another with the same electrostatic force. There are other kinds of forces acting on electrons; among other things, every electron behaves as if it were a tiny bar magnet. As far as electrical charge is concerned, a neutral atom should not have any effect on a passing electron, but, as a result of electrons behaving as magnets, some neutral atoms exhibit a positive affinity for the addition of an extra electron or even two. The measure of this tendency is the binding energy of the electron to be added. If it is larger than the binding energy of the electron when it is in a neighbouring atom, then it is likely that the electron will be transferred, thereby producing two ions, the former host atom being now positive and the new host being negative. These oppositely charged ions will then be attracted to one another, resulting in the formation of a molecule of a compound.

More generally, the electrons most loosely bound in an atom are the ones most likely to have their properties greatly modified by the proximity of other atoms. In some situations, the outer electrons of each of a group of atoms will be so strongly mutually influenced that they can no longer be characterized as "belonging to" one atom rather than another; as a result, they tend to draw the atoms together into a single entity, a molecule of some kind.

Thus, these loosely bound electrons have the essential role in what is called chemical behaviour; they also provide the key to understanding many of the properties of condensed matter, including the conduction of heat and electricity. The electron clouds have directional properties that are reflected in the geometrical characteristics of molecules and of crystals. Magnetic properties of matter depend on the overall electronic structure of atoms and molecules, as does much of its optical behaviour.

DEVELOPMENT OF ATOMIC CONCEPTS

Early philosophical speculations. The concept of an atom with structure is quite recent. The Greek word *atomos*, from which the English word atom is derived, means "undivided"; hence, structureless. An understanding of the atom, however, has grown through incorporation of developments in virtually every field of physics and many areas of chemistry. Thus, a tracing of the historical development of the concepts involved amounts almost to an outline of the history of the physical sciences.

The atomic hypothesis appears to have originated with a Greek philosopher of the 5th century BC, Leucippus. Although nothing is known of his life, and only a single line of his own work survives, his ideas were carried forward by his followers, notably the 5th-century-BC Greek philosopher Democritus of Abdera; in addition to atomism, they include a rigid mechanistic determinism and at least the foreshadowings of the modern principles of conservation of matter and energy. The theories expounded by Democritus were in turn taken over by the 4th- and 3rd-century-BC philosopher Epicurus and his school, who incorporated them into a materialistic philosophy of life; their fullest surviving description is that given by the Roman poet Lucretius in the 1st century BC, in his book *De rerum natura* ("On the Nature of Things"). The es-

sential features are these: (1) All matter is composed of ultimate, indivisible entities, the atoms; and all changes in matter are merely changes in the groupings of atoms. (2) There are many kinds of atoms, of different sizes and shapes, and the properties of bulk matter reflect in some way the properties of the atoms of which it is composed. Thus, for example, a white colour might be a manifestation of smooth atoms on the surface; a sour taste, a manifestation of needle-shaped atoms; and so on. (3) The only forces are those of collision between atoms, and the only causation is that produced by the atomic forces. (4) Apart from atoms, the only reality is infinite empty space, through which the atoms move. (5) An atom in its eternal "falling" through space may occasionally undergo a minute deviation of its path, which then enables it to collide with other atoms, leading ultimately through a succession of collisions to the formation of a whorl that develops into a material object. Of these principles, the first, third, and fourth still appear, with some modification or conceptual extension, as basic tenets of atomism today.

During the medieval period, atomism shared in the general decline of scientific thought in the Western world, as well as experiencing specific opposition to its highly deterministic approach. It seems to have attracted negligible attention in the Muslim world, where, however, significant advances were made in optics and chemistry, as well as substantial contributions to the advances in mathematical notation that were essential to the development of science in the quantitative form now known.

Beginning of modern science

The 17th century witnessed a veritable revolution in science, extending roughly from Galileo Galilei, at the end of the 16th century, to Sir Isaac Newton, one century later. It was characterized by a resurgence of careful, quantitative observation and, more especially, controlled experimentation and by a rise in synthesis and generalization. Out of it there emerged two results of importance to the atomic concept. One was an essentially completely developed science of mechanics, and particularly an emphasis on the role of the mass of a body as the property most important in determining the behaviour of the body in response to an external influence. The other was a dispute over whether light consists of a wave or of a stream of corpuscles. This dispute, won temporarily by the corpuscularists on the weight of Newton's authority, was resolved experimentally in favour of the wave theory a century later, only to be partially revived after still another century. A side development of the work in optics was Newton's discovery of the dispersion of white light into a spectrum by a refracting body such as a prism, which made possible the science of spectroscopy, which in turn has advanced knowledge of atomic structure, as will be seen later.

All the important scientific figures of the 17th century seem to have been more or less staunch atomists, including not only Galileo and Newton but also the English scientists Robert Boyle and Robert Hooke and the Dutch physicist Christiaan Huygens. Hooke even propounded a theory whereby the properties of matter, particularly gases, were to be understood in terms of the motion and collision of atoms, a concept later known as kinetic theory. His idea was considerably ahead of its time; unfortunately, he did not have enough mathematical ability to do it justice. Under the circumstances, it is somewhat surprising that the first work in which atoms were regarded not merely as tiny hard objects but rather as centres of force did not appear until after the middle of the 18th century. This work was published in Venice but was written largely in Vienna, by Ruggero Boscovich, a remarkable scientist-poet-diplomat-priest, born in what is now Yugoslavia and educated in Rome.

Origin of the atomic theory in chemistry. The rise of modern chemistry, too, came later than that of physics. Although Boyle had given effectively the modern definition of an element in 1661, significant progress appeared only when chemical studies became more quantitative, partly under the influence of Newton's emphasis on the importance of mass. Meanwhile, techniques had been developed for the isolation of gases, making possible a wider range of studies. Thus, it became possible to deduce the law of definite proportions, stating that, in a given compound, no matter where or how formed, the components always occur in the same proportions by weight. This law in turn led to tabulations of "equivalent weights," or simply "equivalents," of elements, defined as the weight that would combine with or replace in combination a unit weight of some reference element; *e.g.*, hydrogen.

Laws of proportions

Dalton's postulates. In this setting appeared the first attempt to relate atoms to chemistry: *A New System of Chemical Philosophy*, published in 1808 by an English chemist, John Dalton. Dalton had been led to believe in the atomic hypothesis because of the failure of mixtures of gases of different densities to stratify. He also knew of the law of definite proportions, and he himself had deduced the law of simple multiple proportions, which states that, when two elements combine in more than one proportion by weight, forming more than one compound, the weights of one that combine with a fixed weight of the other are in simple ratios to one another. Thus, for example, nitrogen and oxygen form five different compounds; the weights of oxygen combining with unit weight of nitrogen in each compound are in the ratio 1:2:3:4:5. Dalton was thus led to postulate that (1) all atoms of a given element are identical, and they are different from atoms of any other element; (2) chemical reactions are merely a rearrangement of atoms; and (3) the formation of a compound from its elements takes place by the formation of "compound atoms" from a fixed small number of atoms of the compound elements. These postulates alone were of no particular significance; they were not even essentially new. Dalton wanted to establish the relative weights of atoms, and to do so required at least one additional postulate. A natural preference for simplicity led him to a wrong guess—that (4) if two elements *A* and *B* form only one compound, it is formed by combination of one atom of *A* with one atom of *B;* in particular, the compound atom of water was assumed to consist of one atom of hydrogen and one atom of oxygen.

The law of combining volumes of gases. With this additional assumption, specific results were possible, and tables of relative atomic weights began to be published. The fallacy of the rule of greatest simplicity, however, led to some confusion; related to this was the tacit assumption that the "ultimate particles" of any element are single atoms. The resolution of the difficulty had as its basis a law enunciated in 1808 by a French scientist, Joseph-Louis Gay-Lussac: when gases combine chemically, the ratios of the combining volumes are ratios of simple integers. Thus, hydrogen and oxygen combine to form water in the ratio of two volumes of hydrogen to one of oxygen; carbon monoxide and oxygen combine to form carbon dioxide in the ratio of two volumes of carbon monoxide to one volume of oxygen, forming just two volumes of carbon dioxide; and so on.

Avogadro's hypothesis. These results led an Italian chemist, Amedeo Avogadro, to propose, in 1811, a two-part hypothesis: first, that the ultimate particles (in Dalton's terminology) of even elemental gases are not necessarily atoms but may be groups of atoms joined to form molecules; and, second, that equal volumes of gases contain equal numbers of molecules. Despite the fact that this proposal would have resulted in an entirely consistent scheme, it attracted little notice, because it flew in the face of the prejudices of outstanding chemists of the day. Rather, it was largely neglected until 1858, when another Italian chemist, Stanislao Cannizzaro, presented it, with further supporting evidence, to a chemists' congress at Karlsruhe, Germany. This presentation brought fairly widespread acceptance, but there was still considerable resistance until the early years of the 20th century.

Electrochemistry. In a slightly different vein, the rapid advances in electricity, especially the development of the voltaic cell in 1800, by an Italian physicist, Alessandro Volta, made possible a study of what is now known as electrochemistry, carried out in England by the scientist Michael Faraday. His work in this field is best remem-

The first study of electrochemistry

bered because of his deduction of the quantitative laws of electrolysis; but it is to be noted that these studies led him to recognize that the forces holding atoms together in chemical combination are electrical in origin and extremely strong.

The periodic law. One other development in chemistry was important not because it contributed to knowledge about atoms but because it provided a fundamental datum that a satisfactory atomic theory would have to explain. The existence of groups of elements with similar properties, such as lithium, sodium, potassium, rubidium, and cesium or iron, cobalt, and nickel, had long been known. Once reasonably accurate atomic weights began to be obtained, there appeared certain regularities in the atomic weights of the members of such groups. Often a group consisted of three elements; in many of those cases, either the atomic weights were close together, 56 for iron, 59 for both cobalt and nickel, with no known element having an intermediate value, or else they were widely but approximately equally spaced, 39 for potassium, 85.4 for rubidium, and 133 for cesium. In 1869 the Russian chemist Dmitry Mendeleyev found that, if the elements were listed in the order of increasing atomic weight, these groupings appeared as part of a much more striking and extensive regularity, involving the approximately periodic recurrence of similar sequences of elements. Thus, if hydrogen is skipped, it being somewhat anomalous, and if helium, which along with the other inert gases was unknown at the time, is also ignored, then the sequence lithium, beryllium, boron, carbon, nitrogen, oxygen, fluorine is obtained (see PERIODIC LAW). This sequence runs from a quite reactive element, lithium, which forms positive ions, through the relatively inert and ambivalent element carbon, to a highly active element, fluorine, which forms negative ions. The next seven elements are sodium, magnesium, aluminum, silicon, phosphorus, sulfur, and chlorine. Not only does this sequence display the same variation of properties as the first, but each member of it is chemically quite similar to the corresponding member of the preceding. Continuing in this way, Mendeleyev was led to form a tabular array similar in principle to the modern periodic table, as shown in the articles CHEMICAL ELEMENTS and PERIODIC LAW. It was, of course, not as complete as these, because many elements besides the inert gases were not yet known; in addition, there were some problems arising from the fact that some atomic weights were incorrect at the time (notably uranium, for which Mendeleyev used the value 118, and indium, for which he used 75). Actually, however, its very incompleteness turned out to be an advantage. Mendeleyev was sufficiently convinced of the essential correctness of his scheme that he predicted the existence and the approximate properties of several elements, notably those now known as gallium, scandium, and germanium. These were discovered in 1875, 1879, and 1886, respectively, leaving no doubt as to the validity of Mendeleyev's idea.

Mendeleyev's predictions verified

Spectroscopy and spectral regularities. As was mentioned earlier, Newton discovered that white light could be dispersed into a band of colours by a prism and that the colours were inherent in the light itself. Further work revealed that light from any incandescent solid or liquid gave a continuous band of colours. A luminous vapour, on the other hand, gave a set of isolated colours separated by darkness. These colours became known as spectrum lines, each colour of the spectrum being displayed as the image of a narrow slit. A crucial step was taken in 1859, when the German physicists Gustav R. Kirchhoff and R.W. Bunsen showed that the pattern of lines given off by a given chemical substance is characteristic of the substance.

A search then began for some sort of regularities in spectra, under the assumption that the spectra must originate in some sort of mechanical vibrations within the atom. This assumption led to the expectation of harmonic relations, which were not found.

What ultimately proved to be a major breakthrough was a relationship enunciated in 1885 by Johann Balmer, a Swiss schoolmaster who seems to have been led to his discovery by a love of numerology. By this time, the wave nature of light had been established, and spectroscopy had become the measurement of wavelengths. Balmer found that the wavelengths (symbolized by lambda, λ) of the four visible lines in the spectrum of hydrogen are given to high accuracy by the formula

$$\lambda = b\left(\frac{m^2}{m^2 - n^2}\right), \qquad (1)$$

in which b is a constant having the value 3645.6 angstroms, and m and n are integers; specifically, $n = 2$ and $m = 3,4,5,6$, according to the line involved. Before publication of his findings, he learned of the existence of several more lines, extending into the ultraviolet range of the spectrum; all had wavelengths in striking agreement with the values predicted by his formula with successively higher values of integer m.

In his paper announcing this formula, Balmer suggested that it might be a special case of a more general relationship applicable to other elements, but he carried the work no further. It was left to Johannes Rydberg, a Swedish physicist, to determine, in 1890, that for a large number of series of spectral lines for various elements the wave number (number of waves per unit length, or the reciprocal of the wavelength, symbolized by the Greek letter nu with a superior tilde, $\tilde{\nu}$) of a member of the series is given by an equation of the form

$$\tilde{\nu} = \tilde{\nu}_\infty - \left[\frac{R}{(m + \mu)^2}\right], \qquad (2)$$

in which the subscript ∞ (representing infinity) indicates a particular value of $\tilde{\nu}$, a value that varies from one series to another; m is an integer labelling the position of the line within the series; the Greek letter mu (μ) denotes another constant that depends on the series; and R is a constant that is nearly the same, $1.097 \times 10^5\ \mathrm{cm}^{-1}$, for all series and all elements; *i.e.*, the Rydberg constant.

It is important to note that Balmer's formula, equation (1), is a special case of equation (2). Written in terms of wave numbers instead of frequencies, equation (1) becomes

$$\tilde{\nu} = \frac{1}{\lambda} = \frac{n^2}{b}\left(\frac{1}{n^2} - \frac{1}{m^2}\right) = R\left(\frac{1}{n^2} - \frac{1}{m^2}\right). \qquad (3)$$

This equation is of the form of equation (2) with $\tilde{\nu}_\infty = R/n^2$ (n^2 was fixed at 4 in Balmer's formula) and $\mu = 0$. Other series in hydrogen were not known at the time of Rydberg's work; but in 1908 one was discovered in the infrared for which n in the last form of equation (3) is 3, and m is successively 4,5,6,. . . . Others have been found since.

These relationships and others recognized later finally led to a broad general principle, enunciated in 1908, that states: for any atom there exists a set of characteristic sequences of terms, numbers of the form $T_n = R/(n + \mu)^2$, with n taking on successive integer values and with μ varying from sequence to sequence; the wave number of any line in the spectrum of the atom is expressible as the difference between terms from two such sequences,

$$\tilde{\nu} = R\left[\frac{1}{(n + \mu_1)^2} - \frac{1}{(m + \mu_2)^2}\right] \qquad (4)$$

and a series of lines all having similar characteristics results from the combination of a fixed term of one sequence with all the terms in succession of another sequence; *i.e.*, keeping n fixed in equation (4) and varying m.

Light and related radiations. The 19th century saw an enormous growth in the understanding of light, only to close with a development that would lead to a thoroughgoing modification of the entire structure of that understanding. The century opened with the discovery that there are invisible radiations that have every characteristic of light except visibility: infrared radiation was discovered in 1800 and ultraviolet in 1801. Also in 1801, the English physicist Thomas Young pointed out that a wave theory of light implied the wave phenomenon called interference; and he proceeded to produce examples of it.

Proofs of the wave theory of light

The corpuscularists were not convinced. The dispute was not finally settled until 1850, when it was shown by direct measurement that light travels more slowly in water than in air, a result consistent with the wave description of refraction, whereas the corpuscular description requires the reverse. The phenomenon of polarization was discovered in 1809. This phenomenon implies that light waves are transverse; *i.e.*, whatever the vibration consists of, it takes place in a plane perpendicular to the direction of propagation of the wave.

In 1865 a Scottish physicist, James Clerk Maxwell, produced "A Dynamical Theory of the Electromagnetic Field," which encompassed all the previous laws of electricity and magnetism in unified form. Maxwell concluded from his theory that light is indeed an electromagnetic wave; and his conclusion gained support when in 1887 a German physicist, Heinrich Hertz, was able to generate waves, now known to have been actually radio waves, by purely electromagnetic means.

As the century ended, two pairs of workers at the Imperial Physical and Technological Institution in Berlin undertook experimental measurements of the radiation spectrum of a blackbody (defined as an ideal body that absorbs all radiation incident on it). They made use of the equivalence, established in 1884 by Ludwig Boltzmann, an Austrian physicist, between radiation from a blackbody and radiation emanating through a small hole in a wall of an oven held at uniform temperature. Their experimental findings were analyzed by a German physicist, Max Planck. Planck found that, in order to produce a consistent interpretation of the results, he had to assume that radiant energy is "quantized"; that is, is emitted or absorbed only in discrete amounts, proportional to the frequency that is symbolized by the Greek letter nu, ν. The factor of proportionality he denoted by the symbol h, so that quantized energy (E) is written $E = nh\nu$, in which n is an integer. The factor h had to be a universal constant. Planck calculated its value, and the presently accepted value, deduced from a variety of experiments analyzed simultaneously, is 6.624×10^{-34} joule second. The dimensions joule second—*i.e.*, energy times time, or momentum times distance—are those of a quantity in classical analytical dynamics called the action; thus h is often referred to as the quantum of action, and it plays an essential part in the description of atomic structure.

Planck's constant

Discovery of the electron. In 1858 a form of radiation was discovered emanating from the cathode, or negative plate, of a tube in which an electrical discharge was being passed through a gas; the radiation caused fluorescence when it struck the glass wall of the tube, and the position of the fluorescent spot was affected by a magnetic field in the vicinity. The radiation, called cathode rays, attracted much study. In 1879 an English physicist, Sir William Crookes, found strong indications that they consisted of negatively charged particles. This evidence was confirmed in 1897 by another English physicist, Sir J.J. Thomson, who also measured the ratio of the mass (m) to the charge (e) of the particles, m/e, obtaining a value of about 10^{-7} gram per electromagnetic unit of charge. This value was smaller by a factor of about 1,000 than the smallest value previously known for such a ratio, that of the hydrogen ion in electrolysis; moreover, it was independent of the nature of the gas in which the discharge was produced. Thomson proceeded to study the negative electrification discharged from metals by ultraviolet light and that produced by an incandescent filament in hydrogen gas. He concluded that these all represented a single kind of particle, which he called corpuscles but which soon became known as electrons. He also suggested that electrons were constituents of every kind of atom, which meant, in effect, that the atom could be subdivided.

Multiple independent discoveries of the photoelectric effect

In 1888 several workers independently discovered the photoelectric effect: light falling on a metal surface ejects negatively charged particles from the surface, provided the light includes wavelengths shorter than a critical value dependent on the metal used. As noted above, Thomson identified the particles as electrons. The existence of a threshold wavelength, together with the fact that the energy of the electrons is directly proportional to the difference between the stimulating frequency (ν) and the frequency (ν_0) corresponding to the threshold wavelength, could not be understood on the basis of Maxwell's theory of light. In 1905 Albert Einstein showed that these features followed from Planck's hypothesis of quanta, if it is assumed that each quantum of energy $h\nu$ interacts with just one electron and that an amount of energy $h\nu_0$ is needed just to free the electron from the metal; the remainder, $h\nu - h\nu_0$, is available to the electron as energy of motion. The conclusive experimental confirmation of this theory by a U.S. physicist, Robert A. Millikan, in 1914 established the fact that radiation exhibits some properties normally associated with particles.

Discovery of X-rays and radioactivity. In 1895 the German physicist Wilhelm Röntgen discovered X-rays. Their nature remained controversial for at least 20 years, though some wavelike behaviour was established as early as 1904. In that same year an English physicist, Charles Glover Barkla, showed that each element could be made to emit one or more characteristic groups of X-rays. These groups were labelled K, L, and M, in order of decreasing penetrating power, each group shifting to a more penetrating quality as the atomic weight increased. Barkla also established that the number of electrons in an atom is roughly half the atomic weight. Finally, in 1913 it was established that X-rays are indeed electromagnetic waves that can be made to display interference and diffraction patterns by the natural orderly arrays of atoms in crystals.

This property was adapted in 1913 by the English crystallographer Sir William H. Bragg and his son Sir William L. Bragg into a convenient method for measuring the wavelengths of X-rays.

In 1896 the French physicist Henri Becquerel discovered radioactivity in uranium; this discovery was quickly followed by the isolation of other radioactive elements by two French scientists, Pierre Curie and his wife, Marie. By 1900 it was shown by a British physicist, Ernest Rutherford, and a French physicist, Paul Villard, that there are three kinds of radiation of different penetrating power: alpha rays, the least penetrating, identical with helium nuclei; beta rays, identical with electrons; and gamma rays, the most penetrating, of the same nature as X-rays. The most striking result, however, was the discovery in 1902 by Rutherford and a coworker, Frederick Soddy, that radioactivity is associated with a natural transmutation of the elements. Thus, the atom had lost not only its indivisibility but also its immutability.

Discovery of the existence of isotopes. The study of electrical discharges in gases led to still another surprise. If the cathode of the discharge tube is perforated, a beam of radiation passes through the openings away from the anode and can cause visible fluorescence of the walls of the tube. It was found that these rays consist of positively charged particles, with a ratio of charge to mass that depends on the nature of the gas used. In 1912 Thomson devised a method of subjecting the beam of these rays to parallel and coextensive electric and magnetic fields. The deflections produced by the fields in that configuration are such that all particles of the same charge-to-mass ratio strike the wall of the tube (or a photographic plate used as a detector) along a single parabola-shaped curve; the detailed shape of the parabola can be used to evaluate the ratio of the mass (m) of each particle to its charge (c). He found that with neon as the discharge gas, two parabolas were obtained, which indicated singly charged particles of masses 20 and 22, instead of the expected single parabola of mass 20.2 (the atomic weight of neon). The relative intensities of the two were always the same regardless of the purity of the neon.

The conclusion was that there are two varieties (isotopes) of neon atoms of different masses, thereby dethroning still another hypothesis of the original atomic concept, namely, that all atoms of a given element are absolutely identical.

MODELS OF ATOMIC STRUCTURE

Early atomic models. Several models of the atom were proposed in the first few years of the 20th century. The earliest (1902) was originated by an English physicist, Lord Kelvin, but was supported so strongly by Thomson that it became known as the Thomson atom. According to this model, the atom consists of a sphere of uniformly distributed positive charge, about one angstrom in diameter, in which the electrons are embedded like raisins in a pudding. Lord Kelvin himself at least partly abandoned this model and in 1905 proposed another, in which the uniform positive sphere is replaced by alternate positive and negative spherical shells, with a net surplus of positive charge; the electrons are embedded in the positive charge. Another model, proposed in 1903, suggested that all atoms are composed of varying numbers of a single constituent, called dynamids. Each dynamid is conceived as consisting of an intimate association of an electron and a much more massive positive body, having a linear dimension (diameter) of the order of 10^{-4} angstrom with the dynamids held together by unspecified forces. Still another model, proposed in 1904 by a Japanese physicist, Hantaro Nagaoka, considered the positive charge as concentrated at the centre of the atom, with the electrons forming a ring similar to Saturn's rings.

The Thomson model

The Rutherford model. In 1910 two researchers working under Rutherford's direction were studying the scattering of alpha particles—that is, their deflection from straight-line paths—as they passed through thin foils. They noted that a small but significant number were scattered through quite large angles. Rutherford recognized that, if the Thomson model were valid, an alpha particle would interact with many atoms on the way through the foil; no single interaction could produce a large deflection, and the effects of the many interactions would tend to cancel rather than add up. On the other hand, if the positive charge and most of the mass were concentrated in a very small region at the centre, as Nagaoka had suggested, multiple interactions would be rare, but a single interaction could produce a large deviation. Rutherford worked out the theory in 1911 and showed that the scattering should be proportional to the foil thickness t for a nuclear atom but to the square root of the thickness, $\sqrt{t}$, for the Thomson atom; that it should decrease with increasing deflection angle for a nuclear atom but decrease much faster for the Thomson atom. With refined techniques, a complete verification was obtained of the results of the nuclear model.

The Bohr–Sommerfeld atom. Rutherford's nuclear model of the atom also had serious defects. A well-known theorem in electromagnetic theory states that a system of charged bodies cannot be in static (stationary) equilibrium under the action of their mutual electrostatic interactions alone. On the other hand, a dynamic equilibrium would mean that electrons are travelling in some sort of orbits, analogous to the planets in the solar system; then another, equally well-known result of electromagnetic theory would imply that, because of the centripetal acceleration needed to keep the electrons in their curved paths, they would radiate, lose energy, and spiral into the nucleus. This property would produce a continuous spectrum rather than the discrete (line) spectrum characteristic of noninteracting ions; moreover, the resulting lifetimes of atoms could be calculated straightforwardly and turned out to be ridiculously short.

Bohr's theory and the hydrogen spectrum. The difficulty with instability was resolved by a Danish physicist, Niels Bohr, in 1913, essentially by giving up the relationship between the motion of a charged particle and the radiation it emitted. Bohr's theory was based on four postulates, the first three of which are as follows: (1) An atom, consisting of a nucleus together with its system of electrons, possesses certain special dynamical states having the property that as long as the atom remains in one of these states it does not radiate. (2) The dynamical equilibrium of the special states can be treated by ordinary mechanics; but the transitions between them cannot be so treated and indeed are not subject to explicit description. (3) When an atom makes a transition from one state, of energy E_1, to another, of lower energy E_2, the excess energy is emitted as radiation of a single frequency, ν, related to the energy difference by Planck's relationship, $E_1 - E_2 = h\nu$. Bohr gave three forms for the fourth postulate; the simplest, and the one that became the basis for additional developments, was that (4) for a single electron moving in an orbit around the nucleus, the angular momentum, L, is an integer multiple of $h/2\pi$.

Bohr's postulates

These principles can easily be applied to circular orbits in a hydrogen atom—that is, an atom with a single electron of negative charge, $-e$, bound to a nucleus of positive charge, $+Ze$, Z being an integer, in this case equal to one. The total energy of the system consists of the sum of two parts: (1) the kinetic energy (energy of motion) of the electron, given by the formula $\frac{1}{2}mv^2$, in which m is the electron's mass and v its speed; and (2) the electrostatic potential energy, given by the formula $-(Ze)e/r$, in which r is the radius of the orbit. Thus, adding the two gives

$$E = \tfrac{1}{2}mv^2 - \frac{Ze^2}{r}. \tag{5}$$

To keep the electron in its orbit, there must be a centripetal force on it, which is equal to mv^2/r; this force is provided by the electrostatic attraction of the nucleus, $(Ze)e/r^2$, so that

$$\frac{Ze^2}{r^2} = \frac{mv^2}{r};$$

multiplying by r gives

$$\frac{Ze^2}{r} = mv^2, \tag{6}$$

so that equation (5) can be written

$$E = \tfrac{1}{2}\frac{Ze^2}{r} - \frac{Ze^2}{r},$$

$$E = -\tfrac{1}{2}\frac{Ze^2}{r}. \tag{7}$$

The next step is to apply the quantum condition on the angular momentum. For a circular orbit, the angular momentum is just mvr, so that the condition is

$$mvr = \frac{nh}{2\pi}.$$

Dividing this equation by mr gives

$$v = \frac{nh}{2\pi mr},$$

which can be substituted into equation (5) to give

$$\frac{Ze^2}{r} = \frac{n^2h^2}{4\pi^2m^2r^2}m = \frac{n^2h^2}{4\pi^2mr^2}.$$

This equation can be solved for $1/r$, giving

$$\frac{1}{r} = \frac{4\pi^2mZe^2}{n^2h^2};$$

this in turn can be substituted into equation (7) to give

$$E = -\tfrac{1}{2}Ze^2 \cdot \frac{4\pi^2mZe^2}{n^2h^2},$$

$$E = -\frac{2\pi^2mZ^2e^4}{n^2h^2}. \tag{8}$$

If now the atom makes a transition from a state with energy E_1, corresponding to a value n_1 of n, to another state of energy E_2, corresponding to n_2, the frequency of the radiation emitted is

$$\nu = \frac{1}{h}(E_1 - E_2)$$

$$= \frac{1}{h}\left(-\frac{2\pi^2m\,Z^2e^4}{h^2}\right)\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

$$\nu = \frac{2\pi^2m\,Z^2e^4}{h^3}\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right). \tag{9}$$

This equation can be put in terms of the wave number $\tilde{\nu}$, which is related to frequency ν by $\nu = c\tilde{\nu}$, so that

$$\tilde{\nu} = \frac{2\pi^2 mZ^2e^4}{ch^3}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right).$$

This is exactly the form of Balmer's formula, equation (1), and of the formula for the infrared series, if Balmer's constant b (and Rydberg's constant R) satisfies

$$\frac{4}{b} = R = \frac{2\pi^2 me^4}{ch^3},$$

in which Z has been set equal to 1 because these series pertain to hydrogen (the nucleus of hydrogen has a single positive charge). At the time of Bohr's work, the accepted value of R was 1.1×10^5 cm^{-1}; the value of $2\pi^2me^4/ch^3$, as calculated from then known values of m, e, c, and h, was 1.03×10^5 cm^{-1}, in complete agreement, considering the uncertainties in the quantities involved.

Correspondence principle

Bohr, in deriving the Balmer formula, made use of the correspondence principle, according to which, for large quantum numbers, the results of classical theory are approximately valid. From this, Bohr asserted that, for large orbits, the wave frequency emitted by an electron in transition between two successive orbits equalled the frequency of revolution in either.

The theory had other successful applications. In 1896 a set of lines forming a series similar to the Balmer series was discovered in a stellar spectrum. The wavelengths were given by a variation of Balmer's formula in which the integer m was replaced by half an odd integer, and on that basis it was ascribed to hydrogen. Bohr noted that this series of lines (called the Pickering series for its discoverer, a U.S. astronomer, Edward C. Pickering) could be described as arising from ionized helium (the positive charge on the nucleus of helium, Z, equals 2), and proposed an experiment that was carried out in 1913, and that confirmed this conjecture. Bohr also pointed out that certain lines in stellar spectra not yet identified at the time appeared to be due to doubly ionized lithium (for which $Z = 3$). Slight discrepancies between the observed and calculated values in both these cases were shown to be removed by taking account of the motion of the nucleus. Also, he noted that the energy needed to remove the most tightly bound electron corresponds well to the energy needed to excite characteristic X-rays, and finally he was able to account for some of the features of what is called the Stark effect, the splitting of the lines of a spectrum into several components when the source of radiation is subjected to a strong electric field.

Within two years, two additional experimental results lent strong support to the theory. An English physicist, Henry G.J. Moseley, working initially under Rutherford and later with Thomson, made a systematic study of the variation of the frequencies of characteristic X-rays through the periodic table. He found that the atomic property that appeared to govern the wavelength was not the atomic weight (A) but the atomic number (Z). This discovery confirmed a proposal made earlier by a Dutch lawyer and amateur physicist, Antonius van der Broek, on the basis of an analysis of Geiger and Marsden's work on the scattering of alpha particles, and it resolved some discrepancies that had long existed with regard to the Mendeleyev periodic scheme. In particular, Moseley found that the frequency of the K X-ray series satisfied the equation

$$\sqrt{\nu} = \sqrt{(3/4)cR}\,(Z - 1),$$

in which R is Rydberg's constant and c is the speed of light. The significance of this equation (and of a similar equation for the L series) was recognized by Walther Kossel, a German physicist, in 1914. If both sides of the equation are squared,

$$\nu = (3/4)cR\,(Z - 1)^2,$$

and if the factor 3/4 is written as

$$1 - \frac{1}{4} = \frac{1}{1^2} - \frac{1}{2^2},$$

the resulting equation is just of the form of equation (9) above with $n_1 = 2$ and $n_2 = 1$, except for the replacement of Z by $Z - 1$, the explanation being that if there are normally two electrons in the innermost orbit, and the K series is produced by ejecting one of them, then one electron is left to shield the outer electrons from one unit of the nuclear charge.

The other development was a series of experiments by two German physicists, James Franck and Gustav Hertz, to measure ionization potentials; *i.e.*, the energy required to ionize an atom. When an atom was bombarded with electrons, the energy given to the atom by the electron was restricted to certain discrete values related to the frequencies of certain lines in the spectrum of the atom by Planck's law, $E = h\nu$. This corroborated Bohr's first postulate and provided an extension of the third.

Quantum numbers

Bohr's fourth postulate had introduced one integer, or quantum number, as it became known, into the description of atomic behaviour. This first one, n, was called the principal quantum number. In 1915 this postulate was generalized by the German physicist Arnold Sommerfeld and others, who noted that a similar condition should apply to any periodically changing variable. Sommerfeld treated the effects of relativity on an electron in an elliptical orbit and thereby partially accounted for the fine structure in the hydrogen spectrum—the fact that the lines of, say, the Balmer series actually consist of several components, resolvable by a sufficiently delicate instrument. This treatment involved introduction of a second integer, the orbital quantum number k, which could have the values 1, 2, ..., n and which presumably measured the angular momentum in units of $h/2\pi$ (symbolized $\hbar$). Still a third quantum number was associated with the frequency of precession of the plane of the orbit, if the atom were placed in a magnetic field, caused by the interaction of the field with the equivalent magnet produced by the motion of the electron in its orbit. Use of this magnetic quantum number, with values $-k$, $-k + 1$, $-k + 2$, . . ., $k - 1$, k, permitted understanding of the normal Zeeman effect, which is the splitting of a spectral line into three components under the influence of a magnetic field. With the addition of semiclassical considerations about the angular momentum carried by radiation, it was possible to formulate selection rules that accounted for the absence of certain spectral lines on purely energetic considerations. Experiments in 1922 confirmed the restrictions on spatial orientation of atomic magnets implied by the quantum numbers.

Even the extended Bohr theory, however, had serious flaws. The relativistic fine structure was completely unaccounted for; and the "normal" Zeeman effect is actually rarer than the anomalous Zeeman effect, in which the number of components of the split line is different from three. The theory was also incapable of dealing correctly with any system containing more than one electron.

Electron spin and the Pauli principle. The former difficulty was overcome first. Early in 1925 an Austrian physicist, Wolfgang Pauli, suggested that there should be a fourth quantum number, relating to "a characteristic, classically indescribable kind of double valuedness of the quantum-theoretical properties of the state of the electron." Within the year, the Dutch physicists George Uhlenbeck and Samuel Goudsmit pointed out that it was possible to avoid this "nonmechanical constraint" by ascribing to the electron an intrinsic angular momentum or "spin," of magnitude $\frac{1}{2}\hbar$, in contrast to the "orbital" angular momentum of Bohr's postulate, which was a whole multiple of $\hbar$. According to the quantum rules about angular momentum, the spin could be aligned only along or opposite to a specified direction, giving the required two values. If it had associated with it a magnetic moment of appropriate magnitude, an explanation of the hydrogen fine structure and of the anomalous Zeeman effect on a quantitative basis became possible.

Even before this explanation, Pauli made use of the additional quantum number. The Bohr–Sommerfeld theory implied the existence of shells, groups of orbits all partially similar in their spatial configurations and not too widely separated in energy; and, by a certain amount of

juggling with quantum numbers, it was possible to make the number of orbits in a shell agree, apart from a factor of two, with the number of electrons that the shell seemed to be able to accommodate. There was, however, no explanation for the apparent limitation on the number of electrons in a shell; in particular, there was nothing to prevent all the electrons from dropping into the state of lowest energy, that with $n = 1$. Pauli showed that the limitation on the number of electrons in a shell and, with it, at least a qualitative understanding of the periodic structure of Mendeleyev's table followed from a simple rule: no two electrons in an atom can have the same sets of values for the four quantum numbers. This rule, the Pauli exclusion principle, has subsequently been shown to apply to any particle having a spin that is an odd multiple of $\frac{1}{2}\hbar$.

Pauli exclusion principle

Contemporary atomic models. The key to the solution of the other difficulty with the Bohr–Sommerfeld theory came in two forms, almost simultaneously, one fairly simple in concept and the other rather abstract. The latter will be described only briefly here.

De Broglie's wave particle duality. Ever since Planck's theory of blackbody radiation in 1900, the idea that light has some particle-like aspects had been gaining currency. It had been given support by Millikan's verification of Einstein's theory of the photoelectric effect; and the conclusive result was the discovery in 1923 by a U.S. physicist, A.H. Compton, of the Compton effect, the shift in frequency of the X-rays scattered by an atom, and his explanation of it by treating the X-rays as a stream of particles. (It must be noted, however, that, as Compton himself acknowledged, the experiment itself depended on the use of wave properties of the X-rays.) The French physicist Louis-Victor de Broglie accordingly proposed in 1923 that, correspondingly, particles should possess wave properties. Specifically, he suggested that a particle of mass m_0 at rest and thus having (according to relativistic dynamics) an energy m_0c^2 has associated with it a periodic phenomenon of a specific frequency, $\nu_0 = E/h = m_0c^2/h$. If it moves relative to an observer with a certain velocity v, the observer will see the frequency shifted (by the relativistic time dilation defined by the theory of relativity) to a value $\nu = \nu_0(1 - v^2/c^2)^{1/2}$. On the other hand, the energy of the moving particle is such that it would correspond to a frequency $\nu_1 = \nu_0(1 - v^2/c^2)^{-1/2}$. De Broglie proved that these two waves moved so as to be always in phase: the former at the same speed as the particle and the second at a different speed V. Because according to simple mathematical calculations, V is greater than c (the speed of light), it cannot, according to the theory of relativity, represent a transport of energy, and de Broglie called it the phase wave. He recognized the analogy to a phenomenon of ordinary wave motion in a dispersive medium that, when a group of waves of slightly different frequencies travel at slightly different speeds, they produce by interference a group wave that travels at a speed significantly different from all of them, and it is the group velocity that represents the speed of transport of energy by the waves. Making use of this parallel, he deduced the wavelength of his matter waves: if the momentum of the particle is p, the wavelength λ is given by

Particles with wave properties

$$\lambda = \frac{h}{p}. \tag{10}$$

Just as Planck's equation $E = h\nu$ is the fundamental relationship for light quanta, so equation (10) is the fundamental equation for matter waves.

De Broglie's proposal was an act of sheer bold imagination, based on nothing more substantive than a belief in a symmetry of nature. Nevertheless, in four years it had acquired firm experimental support. Experiments showed that electrons could be diffracted by crystal lattices just as if they were waves of wavelength given by de Broglie's formula.

Schrödinger's equation. De Broglie's development had been influenced by the work of a 19th-century Irish mathematician, William Hamilton, who had produced a formulation of Newtonian mechanics completely parallel to the theory of geometrical optics. In 1926 Erwin Schrödinger, a German physicist, took up de Broglie's idea and exploited further the connection with Hamilton's formulation, thus obtaining the equation that must be satisfied by matter waves. This is a partial differential equation containing the energy of the particle as a parameter (a variable fixed as a constant), and acceptable solutions can be obtained only for certain values of the energy. The form of the equation appropriate to the electron in a hydrogen-type atom is particularly tractable, and Schrödinger found the values of the energy giving acceptable solutions: they were just the values given by Bohr's model, equation (8).

The nature of the solutions of Schrödinger's equation often pointed up the deficiencies in the orbital model that had been built up on the basis of the Bohr–Sommerfeld postulates. An outstanding example of this is the problem of angular momentum. It was noted above that the orbital quantum number k is supposed to give the measure of the orbital angular momentum of the electron in units of $\hbar$. It was on the basis of this interpretation, for example, that the value 0 was excluded: a classical orbit with zero angular momentum is a straight line segment, and an electron in such an orbit would collide with the nucleus. Already before 1925, many formulas involving angular momentum had had to be artificially manipulated in order to give quantitatively correct answers. In some cases k had to be replaced by $k + \frac{1}{2}$, in others by $\sqrt{k(k+1)}$, all purely ad hoc; in still others, no change was necessary. In the Schrödinger formulation, the situation is quite different. There is still a quantum number describing "orbital" angular momentum, now conventionally designated by l, but the word "orbital" is here put in quotation marks to emphasize that there is no longer an orbit. Moreover, it is not l itself that gives the magnitude of the angular momentum, but $\sqrt{l(l+1)}$. The allowed values of k were 1, 2, ..., n; those of l, instead, are 0, 1, ..., $n - 1$. It is now possible to have a state of zero angular momentum, but, because there is no orbit in the classical sense, this state no longer presents any difficulty.

Schrödinger's formulation also permits a mathematical expression of the Pauli exclusion principle in terms of the properties of the function obtained as a solution of the equation. This function, customarily denoted by the Greek letter psi (Ψ), depends on the coordinates of the electrons; in the calculation of observable quantities it always appears squared. Now, because all electrons are identical, no physical difference can result from the interchange of two of them. This statement implies that Ψ must either remain the same or change its sign (so that its square is unchanged) upon interchanging the values of the coordinates of any two electrons. If it is unchanged, it is called symmetric; if it changes sign, it is called antisymmetric. In 1926 a German physicist, Werner Heisenberg, and an English physicist, Paul A.M. Dirac, independently showed that the Pauli principle is equivalent to the requirement that electrons must be described by an antisymmetric Ψ. At the time and for many years thereafter, this principle in either form was an extra added postulate; only in the late 1950s was it found to follow from abstruse considerations of relativistic quantum theory.

Symmetry and antisymmetry of the electron wave function

With this tool at hand, the Schrödinger theory was able to score a clear triumph over the orbital model. The old model had failed completely on the simplest system beyond hydrogen, the neutral helium atom. It could not even provide an arrangement of orbits that was more than marginally stable, let alone account for the nature of the spectrum. The Schrödinger equation, on the other hand, with the inclusion of spin and the requirement of antisymmetry, gave a description that was completely satisfactory.

The difficulties with Schrödinger's equation are of a more philosophical character, having to do with the interpretation and significance of the function Ψ. One disturbing feature is that Ψ is a complex function—that is, it involves $\sqrt{-1}$—so that it cannot by itself represent an observable quantity. The conventional interpretation is

that Ψ is a probability amplitude. The measurable quantity is the corresponding intensity, $|\Psi|^2$, which is a probability density and describes the probability that the electron will be found in a given region of space. It is this probability density that is represented by the clouds mentioned earlier. It must be emphasized that the electron itself is *not* to be regarded as spread out into a cloud.

An entirely different approach to the difficulties of Bohr's theory was taken by Heisenberg and developed by Heisenberg and the German physicists Max Born and Pascual Jordan. Heisenberg insisted that the theory should deal only with observable quantities. Because the atom could be observed only by making it change from one stationary state to another, each quantity had to be represented in some fashion by an array of numbers, each number labelled by the initial and final states to which it pertained. The resulting rules of calculation were recognized by Born as just those of a branch of mathematics called matrix algebra, and together with Jordan, they were able to carry out the development.

Most physicists of the time knew very little about matrix algebra, and therefore this method as such was relatively unexploited—especially because Schrödinger proved in 1926 that it was completely equivalent to his wave mechanics. Nevertheless, one very important result came from it: Heisenberg's uncertainty principle, which states that it is inherently impossible to determine simultaneously and with unlimited accuracy the position and the momentum of a particle. This principle is of great significance on the atomic scale but plays no role in ordinary experience because of the very small value of Planck's constant, h, which is an integral part of the mathematical statement.

The uncertainty principle

At this point the stage was basically set for a complete understanding of "a large part of physics and the whole of chemistry," as Dirac expressed it; the fundamental nature of the new quantum theory was demonstrated in 1928, when the physicist George Gamow and, independently, physicists Edward Condon and R.W. Gurney used its principles to explain the relationship, known since 1911, between the half-life of a radioactive element that emits alpha particles and the energy with which the alpha particles are emitted.

Despite the recognized power of the theory, the details of the structure of any but the simplest atoms are not known exactly. The reason is that there are no techniques for solving any but the simplest partial differential equations. Schrödinger's equation for an atom containing just one electron can be solved exactly; but, even for the next simplest case, that of two electrons, approximation methods must be used. The situation with regard to molecules and aggregates of atoms in bulk matter is even worse. Some of the approximation methods, however, can give very accurate results; and in any case certain general features can be deduced.

Relativistic quantum theories. There remained some defects of principle in Schrödinger's theory. For one thing, the spin of the electron had to be included on an ad hoc basis; for another, the equation is not in harmony with the theory of relativity. Several people, including Schrödinger himself, attempted to remedy the second of these defects, but without success. The difficulty arises from the fact that the relativistic expression for the energy (E) of a free particle is

$$E = \pm[(cp)^2 + (mc^2)^2]^{1/2} \qquad (11)$$

in which m and p are the mass and momentum, respectively, of the free particle, and c is the velocity of light. This cannot be used directly, as the square root has no clear meaning once the transition is made to the operators (either Schrödinger's derivatives or Heisenberg's matrices) of quantum theory. Although it would be possible to use the square of the energy as the form to convert to operators, a probability density with the appropriate behaviour would be impossible to define.

The problem was solved by Dirac in 1928 by the expedient of setting

$$E = c(\alpha_x p_x + \alpha_y p_y + \alpha_z p_z) + \beta mc^2 \qquad (12)$$

in which p_x, p_y, and p_z are the three mutually perpendicular components of momentum, satisfying $p_x^2 + p_y^2 + p_z^2 = p^2$. Here α_x, α_y, α_z, and β are determined by the requirements that they be independent of position and time and that $E^2 = (cp)^2 + (mc^2)^2$, the same result that would be obtained by squaring both sides of Equation (11). The α's and β cannot be ordinary numbers, as the restrictions on them show; a typical one is $\alpha_x\beta + \beta\alpha_x = 0$. Rather, they turn out to be operators themselves. Moreover, whereas for a classical free particle the orbital angular momentum L is a constant of the motion, this is not true for a particle the energy of which is of the form of Equation (12); instead, to L must be added a combination of the α's and β. This combination must represent an angular momentum, which does not depend on the motion of the particle and is therefore an intrinsic angular momentum, or spin. When the effect of an electromagnetic field is included in the equation, the spin is found to have associated with it a magnetic moment of just the magnitude that had been proposed by Uhlenbeck and Goudsmit. Thus, Dirac's proposal solved the two difficulties at once. It also implied the existence of particles identical to the electron except for the sign of the electric charge; these particles, called positrons, were discovered in cosmic radiation in 1932 and are produced in some radioactive processes.

The theories of Schrödinger, Heisenberg, and Dirac all pertain to entities that were classically regarded as particles. The quantum of light, the photon, is in a different category, because the classical theory of light is cast in terms of a continuous quantity, a field. Even before his development of the relativistic theory of the electron, Dirac developed a method for bringing electrodynamics within the realm of quantum theory. The significant development of quantum electrodynamics, however, took place in the late 1940s, partly as a result of experimental revelations. According to the Dirac theory, the lowest level of the hydrogen atom is single; the next two levels should have identical energies. Already in the 1930s, there were some indications that this was not quite so, as well as some theoretical arguments why it should not be. In 1947 a delicate microwave experiment by the U.S. physicists Willis Lamb, Jr., and Robert Retherford proved that the two levels do indeed differ in energy, by an amount corresponding to a frequency of about 1,040 megacycles per second. The Dirac theory also ascribes to the electron a magnetic moment of a definite magnitude; but a careful measurement in 1941 by the U.S. physicist Polykarp Kusch showed that the value differs from the prediction by a little over a tenth of a percent. Results were obtained that led several researchers to develop, independently and almost simultaneously in 1949, a relativistic quantum electrodynamics that eliminated some of the difficulties encountered in earlier works.

The problem with photons

The story is far from complete. Relativistic quantum electrodynamics has its own shortcomings; they are primarily of a philosophical nature, but the fact remains that no one yet knows how to write down exactly, in compact form, an equation describing the complete interaction between a proton and an electron, let alone solve it. Fortunately, the approximations that have been discussed are more than adequate for most purposes.

THE ORIGIN OF ATOMIC SPECTRA

That the richest source of clues for the development of atomic structure theory was furnished by atomic spectra is not surprising. It had been recognized early that each chemical element when heated or electrically excited emits a very characteristic array of spectral lines. Elements occurring in stars and in the solar atmosphere had been identified by their spectra. Moreover, the wavelength of each spectral line could be measured with high precision: rather simple equipment could determine these values easily to one part in 10,000. Thus, any proposed atomic model faced the formidable requirement of explaining in minute quantitative detail the origin of the numerous spectral lines emitted by each element. The later extension of spectroscopy into the invisible regions of ultraviolet and infrared radiation added enormously

Usefulness of atomic spectra

to the wealth of data to be explained. From the mid-20th century, new techniques, such as optical pumping, the principle on which the laser works, opened the radio-frequency and microwave regions for atomic spectroscopy, and the precision of all measurements increased tremendously.

A great simplification was recognition that the sets of terms rather than the spectral lines constitute the characteristic property of the atoms. The Bohr model and the experiment of Franck and Hertz have shown that each of these terms represents a sharply defined excited state, or energy level, of the atom. Spectral lines result when the atom goes over from one discrete excited state into another. If the final state of excitation is of lower energy than the initial one, the energy difference between them is emitted as a spectral line. If, on the other hand, incident light "lifts" the atom into a state of higher excitation, a dark absorption line appears on the spectrum of the transmitted light. The most easily observed absorption lines are those in the solar spectrum; the gases in the atmosphere of the Sun absorb lines in the continuous spectrum emitted by its core (these lines were discovered in 1814 and show the presence of a number of elements in the Sun).

The heart of the problem is to relate the terms to the structure of the atom. For details of this process in complex atoms, see SPECTROSCOPY, PRINCIPLES OF.

ATOMIC STRUCTURE AND MOLECULES

The nature of forces between atoms. For full details of this subject see CHEMICAL BONDING. In order to understand the formation of aggregates of atoms—*i.e.*, molecules and bulk matter—it is necessary to examine in somewhat greater detail the nature of the forces between atoms and their relation to atomic structure. It has been pointed out that the details of the Rutherford–Bohr model, of electrons travelling in orbits like those of planets, cannot be taken literally. Nevertheless, some features of that model are still valid. One such feature is the clustering of the permitted orbits into groups, or shells, as described above. In addition to the spatial compactness described there, the shells have the property that there is a special stability associated with the occupancy of all the orbits in a shell by electrons, two to an orbit. The innermost shell contains just one orbit; thus, helium, with two electrons, represents a particularly stable configuration: it is difficult to remove an electron from a helium atom, and there is virtually no tendency for an extra electron to attach itself to a helium atom. Because forces between atoms result primarily from the transfer (at least partial or temporary) of electrons from one to another, the forces between a helium atom and any other atom are very weak. Consequently, helium does not normally form compounds, liquefies only at very low temperature, and solidifies only under the application of pressure.

Electrons in shells

The next shell contains four orbits and thus can accommodate eight electrons and so is filled when the total number of electrons is ten. This arrangement corresponds to neon, another inert gas. The third shell also contains four orbits, and its completion, at a total of 18 electrons, also corresponds to an inert gas, argon.

Formation of a molecular bond. *Van der Waals force.* There are forces even between such inert atoms. An atom composed of filled shells is spherically symmetric on the average, with the positive charge of the nucleus located at the centre of the sphere of negative charge of the electrons. Such a distribution appears electrically neutral from the outside. But the electrons are in continuous motion, and at some instants the negative charge distribution is not centred on the nucleus. At such instants the atom as seen from outside is equivalent to an electric dipole: two charges, one positive and one negative, at opposite ends of a tiny rod. If two of these dipoles approach each other oriented so that like charges are at adjacent ends, they repel each other; but, if the adjacent ends carry the unlike charges, there is a net attractive force between them. In addition, the interaction in that orientation tends to reinforce the formation of the dipoles. The net result is to give a coherent force between any two atoms, known as the van der Waals force. It is very weak, and in atoms without closed shells it is overwhelmed by other forces.

Electrostatic force. The next example is sodium and chlorine. The sodium atom has one electron more than the capacity of the two innermost shells; the chlorine atom, one fewer than needed to complete the third. Consequently, the sodium atom loses one electron very easily, while the chlorine atom holds an extra electron tightly. If a sodium atom and a chlorine atom encounter one another, the electron is readily transferred from the sodium to the chlorine, forming two ions, one positive and one negative, which then are attracted to each other by the electrostatic force between unlike charges, forming a molecule of sodium chloride, common salt. This is called ionic bonding.

Covalent bonding. Pure ionic bonding is relatively rare; it is almost completely limited to compounds between the alkali metals (each of which has just one electron outside a complete shell) and halogens— fluorine, chlorine, bromine, and iodine, each of which lacks one electron from having a complete shell. Other combinations of atoms rely instead, at least partially, on the sharing of electrons. Two or more atoms are likely to share electrons when each atom thereby attains a closed shell, with each shared electron being counted in both atoms. This is called covalent bonding.

Some understanding may be obtained of how this process operates by considering a special but important case, that of two hydrogen atoms. When the two atoms are very far apart, they have no significant effect on one another. The set of possible states is just the combination of the two sets of atomic states, and the total number of states is therefore twice the number of states for a single atom. Each energy value is represented by two states, one localized around each of the atoms; such an energy and its associated states are called (doubly) degenerate. As the atoms get closer together, however, they begin to influence one another, and the nature of the states changes. They lose the property of being localized around a specific atom, and they change their energies. In particular, the degeneracy is removed, each pair of degenerate states separating into one somewhat higher and one somewhat lower, in a fashion shown schematically in Figure 2A. The labels symmetric and antisymmetric refer to properties of the corresponding Schrödinger functions, Ψ. The

Adapted from (A) N.F. Mott and I.N. Sneddon, *Wave Mechanics and Its Applications* (1948); the Clarendon Press, Oxford, (B,C,D) A. Holden, *The Nature of Solids* (1965); Columbia University Press

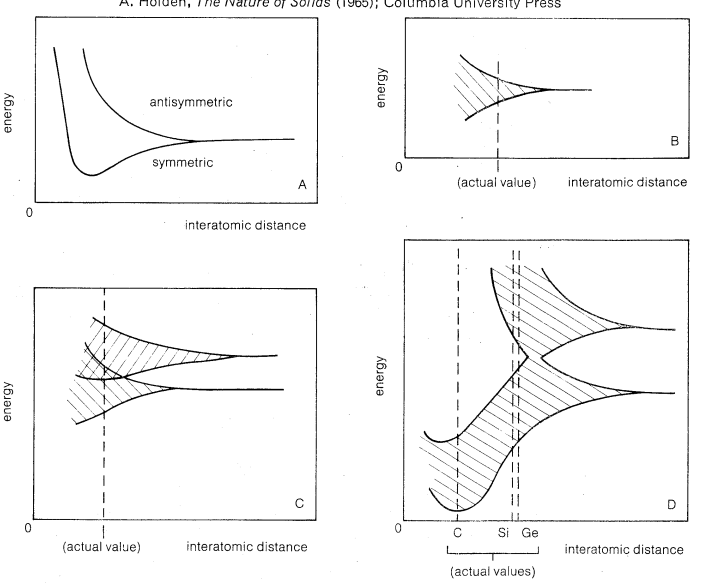


Figure 2: *Energy and distance relationships.*
(A) Variation in energy of the symmetric and antisymmetric states of two hydrogen atoms as the distance between the atoms is varied; (B) spreading of states of a very large number of atoms into a band as the interatomic distance is changed; (C) formation of overlapping bands in, say, zinc; (D) band structure of an insulator (diamond) and two semiconductors.

reason for the difference in behaviour is that an electron in the symmetric state has a high probability of being found between the two nuclei, where it exerts an attractive force on both of them and tends to hold them together. In the antisymmetric state, on the other hand, the electron has a very low probability of being found between the nuclei, so that the positive charges of the nuclei are more exposed to one another, and their electrostatic repulsion comes into play.

Effect of electron spin

These considerations have so far included only the spatial aspects of the states. The inclusion of spin means that each state can accommodate two electrons. In the case of two hydrogen atoms, there are just two electrons present, and so they can both be in the symmetric state, reinforcing the bonding. In the case of two helium atoms, with four electrons, two must go into the antisymmetric state; their effect is to draw the nuclei apart, and this action more than compensates for the binding effect of the two in the symmetric state. The helium molecule-ion, with two electrons in the symmetric state and one in the antisymmetric, is stable.

A basically similar process takes place between other combinations of atoms, although the details will naturally be different. Some examples are shown graphically in Figure 3. Just as purely ionic bonding is rare, however,

Adapted from A. Holden, *The Nature of Solids* (1965); Columbia University Press

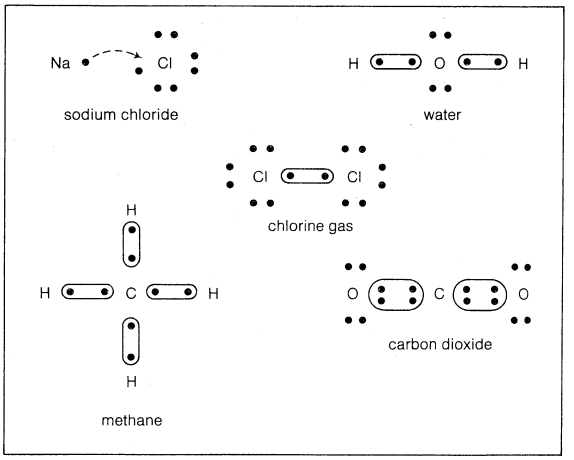


Figure 3: The formation of closed shells in molecules. The dots around each chemical symbol denote the electrons in the outermost incomplete shell. In the case of sodium chloride, the chlorine shell is completed by the transfer of the extra electron from sodium, as indicated by the arrow; in all other cases, the sharing of electrons is indicated by enclosing them together. The hydrogen shell is complete with only two electrons.

so does purely covalent bonding occur only in special cases. A bond between two identical atoms will be purely covalent; but, if the atoms are not identical, there is likely to be a tendency for the shared electrons to be closer on the average to one than to the other. This arrangement results in an average negative charge on the former, positive on the latter, and these average net charges provide some ionic character to the bond.

The origin of molecular spectra. The optical spectra of atoms are produced by transitions between various possible electronic configurations, or states, of the atoms. In molecules, too, the spectra are produced by transitions between different states; but there are two features by which the situation differs from that in atoms. First and most obvious is the fact that the outer electrons are no longer influenced by a single atom but by the whole molecule, so that the states of the molecule are not simply the states of the component atoms; an indication of this effect has been given in the preceding section. Second, it is not only the electrons that contribute to the energy of the molecule. The atoms can move within the molecule to some extent, in a variety of bendings and stretchings of the linkages between them; these motions are known as vibrational modes. Moreover, the molecule as a whole can rotate about various axes. Because the masses involved are those of atoms and molecules rather than electrons, the associated frequencies are much lower than those of purely electronic transitions, giving wavelengths well into the infrared and extending into the microwave range. Especially in large, complicated molecules such as those of organic compounds, many of these motions are characteristic not of the whole molecule but of subordinate parts, such as, say, the group defining an aldehyde, and are very useful for elucidating the structure of the molecule.

The vibrational and rotational modes also affect the spectra associated with the electronic transitions. The molecule may change both its electronic state and, say, its vibrational state simultaneously, and the frequency emitted is then determined by the total change in energy. Because the vibrational and rotational energies are so small, the effect is to associate each electronic transition with not one but a large number of very closely spaced spectrum lines, which under low resolution give the appearance of continuous bands.

ATOMIC STRUCTURE AND BULK MATTER

The nature of bonding in solids. For full details see SOLID STATE OF MATTER; PHASE CHANGES AND EQUILIBRIA. The preceding section described the types of forces that tend to bind atoms or molecules together. Opposing those forces is the motion of the atoms related to temperature: if a gas is at a temperature $T°$ K (kelvin), the atoms composing it are in motion and have an energy of motion (kinetic energy) of the order of kT, k being a constant called Boltzmann's constant and having the value 8.6×10^{-5} electron volt per °K. (At room temperature, about 300° K, this thermal energy is about 1/40 electron volt.) If the thermal energy is less than the energy that can be obtained from the coherent forces, the gas will condense into a liquid or a solid.

Just as there are different kinds of forces between atoms involved in the formation of molecules, so there are different kinds of forces involved in binding bulk matter. A great number of solids, including most organic (compounds of carbon) solids, the solid forms of the inert gases, and the solid forms of such elements as hydrogen and oxygen that are gaseous at ordinary temperatures, are bound by van der Waals forces between the molecules (atoms, in the case of inert gases). These forces are comparatively weak but are stronger for larger molecules, and, moreover, many molecules have permanent dipole moments so that many organic substances remain solid to moderately high temperatures.

Such substances as sodium chloride that form molecules bound by ionic bonding also form solids bound by the electrostatic forces between ions. Each ion tends to surround itself by four, six, or eight ions of the opposite charge, the number depending on the relative sizes of the ions and whether the ions of opposite charges occur in equal numbers in the compound (as sodium chloride, NaCl) of different numbers (as calcium fluoride, CaF_2).

Most inorganic compounds, having some covalent character to their bonds, form solids held together by essentially the same mechanism. Covalent bonds reflect the directional properties of the electron clouds, and solids of this type may have quite complicated structures, which are manifested in complex crystal forms.

Energy bands in the solid state. For full details see SEMICONDUCTORS AND INSULATORS, THEORY OF; ELECTRICITY. The influence of neighbouring atoms on each other's energy states occurs in solids, and the discussion of the covalent bond can easily by carried over. If the number of atoms is N, then the degeneracy of each state when the atoms are isolated is N-fold. Again the degeneracy is removed as the interatomic distance decreases; but the separation between the highest and the lowest energies coming from one originally degenerate level is not affected by the number of atoms. Consequently, if N is very large, as in a piece of bulk matter, the intermediate levels will be so closely spaced as to form an essentially continuous band of energies. Several properties of solids, especially electrical and optical behaviour, are governed by the existence and nature of these bands.

Mechanism of electrical conductivity

The details of the theory of electrical conduction are outside the scope of this article. Generally speaking, however, a substance is an electrical conductor if there is an energy band that contains some electrons but not as many as it could accommodate; it is an insulator if every nonempty band is filled to capacity, with certain exceptions discussed later. Evidently, the electrical properties of a substance depend on the arrangement of the energy bands and the number of electrons that must be accommodated in them. Three cases are shown in Figure 2B–D.

Figure 2B shows the case that applies to the alkali metals. Here the highest occupied band arises from a single set of states. Each of these states in the isolated atoms, however, contains only one electron, so that there are N electrons in a band that can hold $2N$. Thus, the alkali metals are good conductors.

Figure 2C shows another situation that gives rise to metallic conduction, this time the type appropriate to, say, zinc. Here the highest normally occupied atomic state contains two electrons; a band formed from it alone would therefore be filled to capacity. In this case, however, this band overlaps one formed from the next higher atomic state, so that the combined band is only partly filled, and zinc is a conductor.

Two other cases are combined in Figure 2D, which is actually a simplified summary of a much more complicated situation. Here again there is some tendency for the bands to overlap; but here an interaction between the bands causes them to separate again as the interatomic distance decreases still further. Such a picture applies to the elements in the centre of the periodic table; *e.g.*, carbon in the form of diamond, silicon, and germanium. In each case, the lower band is filled by the $4N$ outermost electrons of the N atoms. For diamond, the interatomic distance is well away from the crossover range, the gap between bands is much larger than kT, and diamond is an insulator. For silicon or germanium, however, the interatomic spacing is quite close to the crossover point—close enough so that the gap between the bands is about the same order of magnitude as the thermal energy at ordinary temperatures. Thus, some electrons can absorb enough thermal energy to be raised across the gap. This situation both leaves the lower band incompletely filled and provides some occupancy of the upper, both of which processes give the possibility of electrical conduction. The number of charge carriers, however, that can be made available in this way is several factors of 10 smaller than the number available in a typical metal, so that substances with this sort of band structure are semiconductors (see SEMICONDUCTORS AND INSULATORS, THEORY OF).

Semiconductors

Optical behaviour. Many interesting colour effects in solids have their origin in the structure of the atoms composing the solid. It was pointed out in the section *Origin of atomic spectra* that an atom can absorb radiation only when that radiation is of a frequency corresponding, by Planck's rule, to the energy needed to lift an atomic electron from its normal state to a higher state in the atom. This statement still is true if the atom is part of a solid; now, however, the energy states that must be considered are not those of the isolated atom but those of the atom as affected by its neighbours. Earlier sections have shown that the modification takes two forms: it spreads the originally sharp levels into bands, and it shifts their energies somewhat. The result of the first of these forms is that, whereas an isolated atom will absorb only a relatively few discrete frequencies, a solid will absorb a range of frequencies.

For many crystalline substances, the energy interval from the highest occupied band to the next higher band corresponds to frequencies outside the range of visible light. Such is the case, for example, for pure quartz, sodium chloride, and calcite. In others, however, the energy gap corresponds to visible light. Copper sulfate, for example, absorbs light in the red and yellow region of the spectrum and so has a blue colour.

An especially interesting case of the effect of surroundings occurs in cobaltous chloride. This substance has a tendency to incorporate water into its crystal structure. When it has all the water it can hold, its absorption range is such as to give it a pink colour. As the water is removed, however, the energies of the bands change, and the completely dehydrated form is blue. The water of crystallization can be gained or lost merely by exposure to moist or dry air, so that this substance can be used as a crude humidity indicator.

Luminescence. Every absorption process has corresponding to it an emission process, in which the electron returns from the excited state to its normal state. It would seem at first that these two processes should occur with identical wavelengths, but such is not always the case. An atom in an excited state will generally be different in some spatial respect, such as size, from one in the normal state, with the result that the equilibrium configuration of it and its surroundings may also be altered. When this is so, the crystal usually can readjust itself in a time short compared with the duration of the excited state, leading to changes in the energies of the atomic states involved, usually somewhat different for different states. The return of the electron to its normal state then gives rise to emission of a different (and always lower) frequency than that which was absorbed. This emission is one form of luminescence. Other forms occur when the electron does not return directly to its original state but goes by way of some intermediate state. In some cases, moreover, the excited electron may wander away from its parent atom before it de-excites; it may then spend a considerable time wandering through the crystal before it encounters a vacant lower level to drop into.

It should be noted that the electron may be raised to an excited state by a variety of mechanisms. Not only visible light but also ultraviolet or X-rays may be absorbed —fluorescent lamps use luminescence stimulated by ultraviolet light from a mercury discharge. Also the action of an electric field may be effective in some cases. Even a mechanical strain is sufficent for a few substances.

Stimulated emission. The operation of a laser depends on luminescence of the intermediate-state variety, together with the phenomenon of stimulated emission: an excited atom will be triggered to emit a photon when it is struck by an outside photon of just the energy it would naturally emit; moreover, the emitted photon forms a wave that is in phase—*i.e.*, exactly "in step"—with the triggering wave. In a laser, a large number of electrons are raised to an excited state by an external agency, the pumping light. They then give up some of their energy and drop into an intermediate state the average lifetime of which is typically a few thousandths of a second. (In a ruby laser, this excess energy is given to the crystal as heat; in a gas laser, it may be emitted as a photon of a fairly low frequency.) The laser emission begins when one of these intermediate states emits a photon parallel to the axis of the laser. This photon travels back and forth between the ends, which are silvered so as to act as mirrors, triggering other excited atoms as it passes them; the emission then builds up an avalanche effect.

Magnetism. For full details see MAGNETISM. All atoms possess some magnetic properties. It has already been noted that the orbital motions correspond to current loops that behave like magnets and that the electron spin has a magnetic moment associated with it. Different kinds of atoms, however, behave differently.

The one universal property is called diamagnetism. When the atom is subjected to a magnetic field, the electronic states are altered, and therefore so are the magnetic moments associated with them. The effect is such as to produce an atomic magnetic field opposing the applied field. A bar-shaped piece of material in which this is the only magnetic effect tends to align itself across a magnetic field. Bismuth is the outstanding example of such a substance. Usually, this effect is masked by others.

In most materials, including not only atoms (gases) but also molecules, liquids, and solids, the electronic magnetic fields do not exactly add to zero. This situation leaves a weak magnetic effect called paramagnetism.

In a very few substances, notably iron, cobalt, and nickel but including some special alloys, there are several electrons per atom the spins of which are all in the same

direction. (These are not the outermost, or valence, electrons but are farther in the interior of the atom.) This arrangement produces a large net effect, one that extends to neighbouring atoms, causing the atoms throughout large portions of the resulting solids to arrange themselves with their atomic magnetic moments aligned parallel to one another. These portions, called domains, interact strongly with applied magnetic fields and give the substances involved their characteristic magnetic properties, called ferromagnetism. While ferromagnetism has its ultimate origin in the properties of individual atoms, however, it is basically a phenomenon involving correlation of many atoms.

BIBLIOGRAPHY. GEORGE GAMOW, *Mr. Tompkins in Paperback* (1967), two classic popularizations in a single volume, with atomic-level effects made understandable by tremendous exaggeration; BANESH HOFFMANN, *The Strange Story of the Quantum*, 2nd ed. (1959), a history of the development of quantum theory, with a light touch; MAX BORN, *The Restless Universe*, 2nd ed. (1951), a masterful and clear account of atomic physics and the experiments on which our knowledge is based; SELIG HECHT and EUGENE RABINOWITCH, *Explaining the Atom*, rev. ed. (1954), a layman's description of the atom, plus a discussion of nuclear energy and weaponry; ALAN HOLDEN, *The Nature of Solids* (1965), an elementary treatment of the simpler aspects of solid-state physics, including the necessary wave-mechanics background; FRANCIS O. RICE and EDWARD TELLER, *The Structure of Matter* (1949), an intermediate-level textbook of quantum theory, slanted toward the applications indicated in the title.

(G.L.T./S.A.G.)

Atomic Weight

The term atomic weight is equivalent to "relative atomic mass," the ratio of the average mass of an element's atoms to some standard. Since 1961 the standard unit of atomic mass has been $\frac{1}{12}$ of the mass of an atom of the isotope carbon-12 (an isotope is one of two or more species of atoms of the same chemical element that have different atomic masses). The atomic weight of carbon is 12.011, the average that reflects the typical ratio of natural abundances of its isotopes.

The concept of atomic weight is basic to chemistry, because most chemical reactions take place in accordance with simple numerical relationships among atoms. Since it is almost always impossible to count the atoms directly, chemists measure reactants and products by weighing and reach their conclusions through calculations involving atomic weights.

The quest to determine the atomic weights of elements occupied the greatest chemists of the 19th and early 20th centuries. Their careful experimental work became the key to chemical science and technology.

Variations in atomic-weight values. At the turn of the 20th century, the simple concept of a true atomic weight characteristic of all specimens of any one element remained unchallenged. The idea was that, as precision measurement improved, true values would be increasingly more closely approached.

What followed is typical of the history of the science of measurement. Need for more accurate knowledge of a supposed constant of nature is recognized. Experimenters take up the challenge. As they obtain more precise values, it is suddenly realized that the desired number is not a universal constant. The discovery of isotopes disproved the idea that atomic weights were universal constants, yet the vast majority of changes that occur in the composition of matter are those in which neither the atoms themselves are altered nor significant isotope discrimination is involved. Today, all quantitative operations in chemical reactions, whether aimed at analysis or synthesis of materials, still pivot on atomic and molecular weights. (The molecular weight of a molecule is numerically equal to the sum of the atomic weights of all the atoms that comprise the molecule.)

Atomic masses of individual nuclides (specific isotopic species of one element) can be determined with great accuracy, but only 20 stable elements are limited to one stable nuclide. For all others, the chemist needs the average of the atomic masses appropriate for the mixture of nuclides as he finds them. Basic limitations on precision thus present themselves in the provision and use of atomic weights for the chemist: (1) Even the most precise methods of measurement presently available introduce uncertainties that limit knowledge. (2) There are significant differences in isotopic composition of elements and surprisingly sparse knowledge of the variabilities, especially for minor elemental constituents in rocks and minerals. (3) Modern technology changes the isotopic composition of processed materials. To a small extent, every chemical and every physical process and, to a large extent, every nuclear process differentiates among isotopes of the same element. (4) Extraterrestrial materials, the isotopic compositions of which have not yet been determined, will soon be more widely available.

Problems of measurement

Significance of atomic weights to chemistry. The general acceptance of the early-19th-century atomic theory of the constitution of matter—*i.e.*, as being composed of indivisible atoms, indestructible and merely rearranged by chemical reactions—developed by the English chemist John Dalton, was soon followed by the conclusion that atoms of different elements combine in simple numerical proportions and prepared the ground for a rapid growth of experimentation in chemistry. There were diverse objectives, among them the discovery of new elements, the analysis of rocks and minerals, the synthesis of new substances, and the study of the distribution of the elements in nature. Fundamental tools available for this task were the laboratory balance and volume-measuring devices. The latter served directly, with the aid of Avogadro's hypothesis (which states that equal volumes of gases, at the same pressure and temperature, contain equal numbers of molecules), to deduce the numerical proportions in which atoms of gaseous substances react. For solids and liquids, the primary measurable information was the proportion by weight that characterizes the reactions. On the basic assumption that atoms combine in integral numerical proportions, it could be concluded that relative combining weights were either identical with, or simple submultiples of, atomic weights.

In the early history of chemistry, chemists sought to measure the masses of atoms solely as a means of counting atoms, not out of interest in atomic mass as such. As long as the atom was considered immutable, its mass was of only secondary interest. The relative values of atomic masses, on a common scale, were of the greatest importance. Such ratios provided the means of deriving numerical relationships among reacting atoms, and, by simple calculations, the formulas of molecules could be derived.

Work toward continual improvement of atomic-weight values yielded important by-products. Since the end result of such efforts was considered to be a true constant of nature, variations in experimental results obviously revealed errors. Uncovering sources of errors and determining their magnitudes eventually resulted in significant advances in chemical knowledge. Areas of special interest were the purity and stoichiometric (combining proportions) composition of substances, prevention of loss of substance and contamination during chemical reactions, accurate weighing, and improved knowledge of the chemical reactions themselves. A marked advance in laboratory procedures resulted when fused-silica reaction vessels and electrically heated furnaces became available to the chemist. The use of these devices greatly decreased the likelihood of contamination of reactants and their product by container materials and products of combustion.

Key discoveries

As the accuracy of atomic-weight values increased, their significance to the understanding of the laws of chemistry increased also. Anomalies in the periodic classification of the elements (the periodic table), based on their atomic weights, were confirmed even if not immediately explained. It became clear that iodine should follow tellurium and that nickel should follow cobalt in the table, despite the fact that, according to their atomic weights, the order should be reversed.

The answer to these riddles had to await the discovery of isotopes, a discovery that was itself hastened by the

reluctant conclusion of T.W. Richards, an American chemist and Nobel Prize recipient for his researches on atomic weights, that a group of lead samples of differing geological origins were identical chemically but differed in atomic weight. Thus the way was prepared for the discovery, soon to follow, that most elements occur in nature as mixtures of atomic species to which the name isotopes was given.

Because the rationale of organic compounds depends so strongly on the atomic weight of carbon, much effort was expended on this element. An unusual number of difficulties was encountered. As late as 1937 an attempt to elucidate the structure of the hydrocarbon that constitutes natural rubber failed to yield meaningful results when the then accepted value for the atomic weight of carbon (12) was used in calculating the numerical ratio of carbon and hydrogen atoms. A year later, a new determination yielded the value 12.01, which was later further revised to 12.011.

Reliable values for atomic weights serve an important purpose, in a quite different way, when chemical commodities are bought and sold on the basis of the content of one or more specified constituents. Ores of costly metals such as chromium or tantalum and the industrial chemical soda ash are examples. Content of the specified constituent must be determined by quantitative analysis. The computed worth of the material depends on the atomic-weight values used in the calculations.

Atomic-weight scales. The original standard of atomic weight, established in the 19th century, was hydrogen, with a value of 1. From about 1900 until 1961, oxygen was used as the reference standard, with an assigned value of 16. The unit of atomic mass was thereby defined as 1/16 the mass of an oxygen atom. In 1929, it was discovered that natural oxygen contains small amounts of two isotopes slightly heavier than the most abundant one and that the number 16 represents a weighted average of the masses of the atoms of the three isotopic forms of oxygen as they occur in nature. This situation was considered undesirable for several reasons, and, since it is possible to determine the relative masses of the atoms of individual isotopic species, a second scale was soon established with 16 as the value of the principal isotope of oxygen rather than the value of the natural mixture. This second scale, preferred by physicists, came to be known as the physical scale, and the earlier scale continued in use as the chemical scale, favoured by chemists, who generally worked with the natural isotopic mixtures rather than the pure isotopes.

Although the two scales differed only slightly, the ratio between them could not be fixed exactly, because of the slight variations in the isotopic composition of natural oxygen from different sources. It was also considered undesirable to have two different but closely related scales dealing with the same quantities. For both of these reasons, chemists and physicists established a new scale in 1961. This scale, based on carbon-12, required only minimal changes in the values that had been used for chemical atomic weights.

Methods of determining atomic weights. *Chemical methods.* In the late 19th century, chemists differed among themselves on the merits of two scales of atomic weights, one based on hydrogen with the assigned relative mass of 1 and the other on oxygen with the assigned integral value of 16. Values on these two scales differed by nearly 1 percent, and much effort was directed toward fixing the relationship between the scales more exactly by determining the combining-weight ratio of the two elements. The first results on this ratio had been published in 1821, and other measurements had followed at intervals, until the American chemists E.W. Morley in 1895 and W.A. Noyes in 1907 published their elaborate and definitive studies. Their results did not quite agree with each other, but the mean of the two investigations agrees well with that derived from more recent physical determinations of the relative atomic masses of the two elements.

The oxygen–hydrogen ratio

A good example of the directly measured ratio of an element to oxygen is found in the work of two American chemists, G.P. Baxter and C.R. Hoover, who in 1912 reduced weighed amounts of carefully prepared ferric oxide in a current of hydrogen and weighed the residue of pure iron. Their investigation yielded 55.8456 as the atomic weight of iron, in excellent agreement with the currently accepted value, 55.847 ± 0.003.

The relative ease of preparing highly pure metallic silver, together with the stability of silver chloride and silver bromide, led many investigators to determine the equivalence ratios (relationship between quantities of combining chemical species) of various soluble chlorides and bromides with silver and the corresponding silver salts. The experimental procedures involved in such measurements were brought to a high degree of perfection in the laboratories of T.W. Richards and G.P. Baxter at Harvard University and of O. Hönigschmid at the University of Munich, and it can now be said that the observed ratios of silver to chlorine, bromine, and oxygen were in error by no more than 0.001 percent, 0.002 percent, and 0.003 percent, respectively. Such determinations made major contributions during the period up to 1940 to the reliability of the International Table of Atomic Weights.

Halide–silver ratios

It seems probable that a principal source of the errors now known to have affected the halide–silver ratios was the virtual impossibility of achieving true equilibrium between a precipitate of silver chloride or bromide and the solution in which it was formed (equilibrium: a state of balance between all reactants and products in a reversible chemical reaction, attained when two opposing reactions go on at equal rates). If the equivalence point (where each reaction has the same rate) can be determined for two substances that react in solution without forming a precipitate, attainment of equilibrium is more assured. An example is provided by combining the ratio of sodium carbonate to iodine pentoxide (which forms periodic acid when dissolved in water) with that resulting from the dissociation of iodine pentoxide to iodine and oxygen. Combining these ratios cancels an apparent minute deviation of the iodine pentoxide from its nominal proportionate composition, and values then derived for sodium, carbon, and iodine are consistent within 0.001 percent of presently accepted values.

Physical methods. Physical measurements are rapidly increasing in diversity and precision. Probably the gradual change from atomic-weight values based on chemical determinations to values based on physical measurements will run its course to completion, but it is difficult as yet to predict which physical methods will be the most accurate.

Types of physical measurement

Basically, two types of physical measurements should be distinguished. First, there are those in which the atomic weight of an element is determined directly as an average appropriate for the isotopic composition of that element. Historically the most important technique of this kind is that of gas-density ratios depending on Avogadro's rule. The most promising method of this first type is the X-ray-diffraction method in which, ideally, the macroscopic density of a pure, perfect crystal is compared with the density of the atomic-scale-pattern unit, dimensionally determined by X-ray diffraction.

In the second type of physical determinations the atomic masses of nuclides are first measured. This can now be done with great accuracy by mass spectroscopy, the technique in which charged particles are separated according to their atomic masses and from the energy changes in nuclear reactions. For elements composed of only one nuclide, extremely accurate atomic-weight values are thus directly available. The isotopic composition must be separately measured, however, for the great majority of the elements before their atomic weights can be calculated from the nuclidic masses.

To date, mass spectroscopy is the only technique used for measuring isotopic compositions quantitatively, but the technique of resonance spectroscopy (in which the character of atomic nuclei placed in a static magnetic field is observed through resonant response to a high-frequency magnetic field) may become competitive in the future. Mass spectroscopic measurements, calibrated by

Atomic Weights, 1971

(based on the assigned relative atomic mass of $^{12}C = 12$)

The values given here apply to elements as they exist in materials of terrestrial origin and to certain artificial elements. When used with due regard to the footnotes they are considered reliable to ± 1 in the last digit, or ± 3 if that digit is underscored. Atomic weights of elements with no stable isotope are given as mass numbers, in parentheses, of the most stable isotope.

name	symbol	atomic number	atomic weight
Actinium	Ac	89	(227)
Aluminum	Al	13	26.98154*
Americium	Am	95	(243)
Antimony	Sb	51	121.75
Argon	Ar	18	39.948†‡§‖
Arsenic	As	33	74.9216*
Astatine	At	85	(210)
Barium	Ba	56	137.34
Berkelium	Bk	97	(247)
Beryllium	Be	4	9.01218*
Bismuth	Bi	83	208.9804*
Boron	B	5	10.81‡§♀
Bromine	Br	35	79.904‡
Cadmium	Cd	48	112.40
Calcium	Ca	20	40.08
Californium	Cf	98	(251)
Carbon	C	6	12.011†§
Cerium	Ce	58	140.12
Cesium	Cs	55	132.9054*
Chlorine	Cl	17	35.453‡
Chromium	Cr	24	51.996‡
Cobalt	Co	27	58.9332*
Copper	Cu	29	63.546†§
Curium	Cm	96	(247)
Dysprosium	Dy	66	162.50
Einsteinium	Es	99	(254)
Erbium	Er	68	167.26
Europium	Eu	63	151.96
Fermium	Fm	100	(257)
Fluorine	F	9	18.99840*
Francium	Fr	87	(223)
Gadolinium	Gd	64	157.25
Gallium	Ga	31	69.72
Germanium	Ge	32	72.59
Gold	Au	79	196.9665*
Hafnium	Hf	72	178.49
Helium	He	2	4.00260†‡
Holmium	Ho	67	164.9304*
Hydrogen	H	1	1.0079†§
Indium	In	49	114.82
Iodine	I	53	126.9045*
Iridium	Ir	77	192.22
Iron	Fe	26	55.847
Krypton	Kr	36	83.80
Lanthanum	La	57	138.9055†
Lawrencium	Lr	103	(257)
Lead	Pb	82	207.2§‖
Lithium	Li	3	6.941‡§♀
Lutetium	Lu	71	174.97
Magnesium	Mg	12	24.305‡
Manganese	Mn	25	54.9380*
Mendelevium	Md	101	(258)
Mercury	Hg	80	200.59
Molybdenum	Mo	42	95.94
Neodymium	Nd	60	144.24
Neon	Ne	10	20.179‡
Neptunium	Np	93	237.0482†¶
Nickel	Ni	28	58.71
Niobium	Nb	41	92.9064*
Nitrogen	N	7	14.0067†‡
Nobelium	No	102	(255)
Osmium	Os	76	190.2
Oxygen	O	8	15.9994†‡§
Palladium	Pd	46	106.4
Phosphorus	P	15	30.97376*
Platinum	Pt	78	195.09
Plutonium	Pu	94	(244)
Polonium	Po	84	(209)
Potassium	K	19	39.098
Praseodymium	Pr	59	140.9077*
Promethium	Pm	61	(145)
Protactinium	Pa	91	231.0359*¶
Radium	Ra	88	226.0254*¶‖
Radon	Rn	86	(222)
Rhenium	Re	75	186.2
Rhodium	Rh	45	102.9055*
Rubidium	Rb	37	85.4678‡
Ruthenium	Ru	44	101.07
Samarium	Sm	62	150.4
Scandium	Sc	21	44.9559*
Selenium	Se	34	78.96
Silicon	Si	14	28.086§
Silver	Ag	47	107.868‡
Sodium	Na	11	22.98977*
Strontium	Sr	38	87.62‖
Sulfur	S	16	32.06§
Tantalum	Ta	73	180.9479†
Technetium	Tc	43	98.9062¶
Tellurium	Te	52	127.60
Terbium	Tb	65	158.9254*
Thallium	Tl	81	204.37
Thorium	Th	90	232.0381*¶
Thulium	Tm	69	168.9342*
Tin	Sn	50	118.69
Titanium	Ti	22	47.90
Tungsten (wolfram)	W	74	183.85
Uranium	U	92	238.029†‡♀
Vanadium	V	23	50.9414†‡
Xenon	Xe	54	131.30
Ytterbium	Yb	70	173.04
Yttrium	Y	39	88.9059*
Zinc	Zn	30	65.38
Zirconium	Zr	40	91.22

*Mononuclidic element. †Element with one predominant isotope (about 99–100% abundance). ‡Element for which the atomic weight is based on calibrated measurements. §Element for which variation in isotopic abundance in terrestrial samples limits the precision of the atomic weight given. ‖In some geological specimens this element has a highly anomalous isotopic composition, corresponding to an atomic weight significantly different from that given. ¶Most commonly available long-lived isotope. ♀Element for which users are cautioned against the possibility of large variations in atomic weight due to inadvertent or undisclosed artificial isotopic separation in commercially available materials.

synthetic isotope mixtures, are generally superior to chemical and to other physical methods for elements with two or at most three isotopes. Thus for some time to come mass spectroscopy is likely to remain the technique on which a large number of atomic-weight values are based.

Summary. The atomic weight of a chemical element is the ratio of the mass of a large number of its atoms (randomly selected from naturally occurring material) to the mass of the same number of atoms of carbon with six neutrons in their nucleus; *i.e.*, the isotope carbon-12.

Atomic-weight tables. The table of atomic weights here reproduced, with minor editorial changes, is published and biennially revised by the Commission on Atomic Weights of the International Union of Pure and Applied Chemistry. The Table sets out values intended to be as accurate as possible, so accurate as not to discard reliable knowledge; however, atomic-weight values are not carried to more decimal places than are justified by the variability of isotopic composition and uncertainty in determination. The Table is also intended to warn of some kinds of processed materials that need to be more closely characterized before published atomic-weight values can be applied with confidence; *e.g.*, lithium, which may have been subjected to isotopic separation in commercially available materials.

BIBLIOGRAPHY. The basic concepts involved in atomic weight measurement and use are discussed in all elementary textbooks on chemistry. The historic development and detailed considerations during the last 70 years are best followed through commission reports. The responsible commission since 1923 has been that on Atomic Weights of the International Union of Pure and Applied Chemistry. Its reports are now published in *Pure and Applied Chemistry*. They are usually translated or reprinted in national chemical literatures. Of recent reports the most important are those of 1961 and 1969, for which the following are among the best references: A.E. CAMERON and E. WICHERS, "Report of the International Commission on Atomic Weights 1961," *J. Am. Chem. Soc.*, 84:4175–4179 (1962); and the INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY, COMMISSION ON ATOMIC WEIGHTS, "The Atomic Weights of the Elements 1969," *Pure Appl. Chem.*, 21:95–108 (1970).

(H.S.P./E.Wi.)

Atomism

In the broadest sense the term Atomism refers to any doctrine that explains complex phenomena in terms of aggregates of fixed particles or units. This philosophy has found its most successful application in natural science: according to the atomistic view, the material universe is composed of minute particles, which are considered to be relatively simple and immutable and too small to be visible. The multiplicity of visible forms in nature, then, is based upon differences in these particles and in their configurations; hence any observable changes must be reduced to changes in these configurations.

THE BASIC NATURE OF ATOMISM

Atomism is in essence an analytical doctrine. It regards observable forms in nature not as intrinsic wholes but as aggregates. In contrast to holistic theories, which explain the parts in terms of qualities displayed by the whole, Atomism explains the observable properties of the whole by those of its components and of their configurations.

Core concepts

In order to understand the historical development of Atomism and, especially, its relation to modern atomic theory, it is necessary to distinguish between Atomism in the strict sense and other forms of Atomism. Atomism in the strict sense is characterized by three points: the atoms are absolutely indivisible, qualitatively identical (*i.e.*, distinct only in shape, size, and motion), and combinable with each other only by juxtaposition. Other forms of Atomism are less strict on these points.

Atomism is usually associated with a "realistic" and mechanistic view of the world. It is realistic in that atoms are not considered as subjective constructs of the mind employed for the sake of getting a better grip upon the phenomena to be explained; instead, atoms exist in actual reality. By the same token, the mechanistic view of things, which holds that all observable changes can be reduced to changes of configuration, is not merely a matter of employing a useful explanatory model; the mechanistic thesis holds, instead, that all observable changes are caused by motions of the atoms. Finally, as an analytic doctrine Atomism is opposed to organismic doctrines, which teach that the nature of a whole cannot be discovered by dividing it into its component parts and studying each part by itself.

VARIOUS SENSES OF ATOMISM

The term Atomism is derived from the Greek word *atoma*—"things that cannot be cut or divided."

Two basic types of Atomism. The history of Atomism can be divided into two more or less distinct periods, one philosophical and the other scientific, with a transition period between them (from the 17th to the 19th century). This historical fact justifies the distinction between philosophical and scientific Atomism.

Philosophical Atomism. In philosophical Atomism, which is as old as Greek philosophy, attention was focussed not on the detailed explanation of all kinds of concrete phenomena but on some basic general aspects of these phenomena and on the general lines according to which a rational explanation of these aspects was possible. These basic aspects were the existence in nature of a manifold of different forms and of continuous change.

Forces and change

In what way could these features be explained? Philosophical Atomism offered a general answer to that question. It did not, however, strictly confine itself to the general problem of explaining the possibility of change and multiplicity—not even in ancient Greek Atomism, for in Greek thought philosophy and science still formed a unity. Consequently, Atomists also tried to give more detailed explanations of concrete phenomena, such as evaporation, though these explanations were meant more to endorse the general doctrine of Atomism than to establish a physical theory in the modern sense of the word. Such a theory was not yet possible, because a physical theory must be based upon indirect or direct information about the concrete properties of the atoms involved, and such information was not then available.

Modern atomic theory. With the development of a scientific atomic theory, the general philosophical problems gradually disappear into the background. All attention is focussed on the explanation of concrete phenomena. The properties of the atoms are determined in direct relationship with the phenomena to be explained. For this reason the chemical atomic theory of the 19th century supposes that each identified chemical element has its own specific atoms and that each chemical compound has its own molecules (fixed combinations of atoms). What particles act as unchanged and undivided units depends upon what kind of process is involved. Some phenomena, such as evaporation, are explained by a process in which the molecules remain unchanged and identical. In chemical reactions, however, the molecules lose their identity. Their structures are broken up, and the composing atoms, while retaining their own identity, are rearranged into new molecules. With nuclear reactions a new level is reached, on which the atoms themselves are no longer considered as indivisible: more elementary particles than the atoms appear in the explanations of nuclear reactions.

Extensions to other fields. Whereas classical Atomism spoke mainly of material atoms (*i.e.*, of particles of matter), the success of the atomic doctrine encouraged the extension of the general principles of Atomism to other phenomena, more or less removed from the original field of application. Rather plausible, for example, was the extension of Atomism to the phenomena of electricity. There were reasons to suppose the existence of an elementary charge of electricity associated with an elementary material particle: the electron (19th century). A second fruitful extension concerned energetic processes (20th century). Some experimental data suggested the hypothesis that energy can exist only in amounts that are whole multiples of an elementary quantity of energy. Extensions of the idea of an atomic structure to amounts of gravitation and even to time have been attempted but have not been sufficiently confirmed.

Logical Atomism, monadology, associationism

More removed from the original field of application of Atomism is a theory known as Logical Atomism (developed by the eminent philosopher and logician Bertrand Russell and by the philosopher of language Ludwig Wittgenstein), which supposes that a perfect isomorphism exists between an "atom" of language (*i.e.*, an atomic proposition) and an atomic fact; *i.e.*, that for each atomic fact there is a corresponding atomic proposition. An atomic proposition is one that asserts that a certain thing has a certain quality; *e.g.*, "this is red." An atomic fact is the simplest kind of fact and consists in the possession of a quality by some individual thing.

Another application of Atomism (albeit in a moot sense) lies in the monadology of the philosopher-scientist G.W. Leibniz. According to Leibniz the atoms of Democritus, who provides the paradigm case of ancient Greek Atomism, were not true unities; possessing size and shape, they still were divisible in principle. The ultimate constituents of things must, therefore, be points, said Leibniz—not mathematical but metaphysical points, *i.e.*, points of real existence. They are indeed a kind of soul, which he came to call "monads."

In psychology, Atomism is a doctrine about perception. It holds that what man perceives is a mosaic of atomic sensations, each independent and unconnected with any other sensation. According to the early modern Empiricist David Hume and the pre-World War I father of experimental psychology Wilhelm Wundt, the fact that man nevertheless experiences an ordered whole formed from the unordered "atoms" of perception is caused by the mind's capacity to combine them by "association."

DIVERSE PHILOSOPHICAL CHARACTERIZATIONS OF ATOMISM

The intrinsic nature of the atoms. In 1927 the Belgian astronomer Georges Lemaître formulated the hypothesis that the present high degree of differentiation of matter in space and the complexity of forms displayed by the various astronomical objects must have resulted from a violent explosion and subsequent dispersal of an origi-

One being: Lemaître and Parmenides

nally highly compressed homogeneous material, a kind of "primitive atom," containing all of the matter that exists. From the philosophical viewpoint this hypothesis is interesting. By its attempt to reduce the manifold to unity it recalls the beginning of Greek philosophy, which was also inspired by a thesis of the unity of being, propounded by the Eleatic Parmenides. Even apart from their respective contexts, there is, of course, a great difference between Lemaître's and Parmenides' conceptions of the unity of being, for the latter combined the thesis of the unity of being with that of the immutability of being.

Although it would be wrong to classify Parmenides among the Atomists, it is nonetheless appropriate that in an introduction to the diverse forms of Atomism, his conception of reality as just one being should be mentioned. Parmenides' thesis is not only historically but also logically the cornerstone of atomistic thought. Any atomic theory can be interpreted as an attempt to reconcile the thesis of the unity and immutability of being with the fact that the senses observe multiplicity and change. The different ways in which the unity and immutability are understood characterize the different forms of Atomism.

Atoms as lumpish corpuscles. As corpuscles (minute particles), atoms can either be endowed with intrinsic qualities or be inherently qualityless.

Inherently qualitied atoms

The most striking basic differences in the material world, which lead to a first classification of substances in nature, are those between solids, liquids, gases, and fire. These differences are an observed datum that must be accounted for by every scientific theory of nature. It is, therefore, only natural that one of the first attempts to explain the phenomena of nature was based upon these differences and proclaimed that there are four qualitatively different primitive constituents of everything, namely, the four elements: earth, water, air, and fire (Empedocles, 5th century BC)—a theory that dominated physics and chemistry until the 17th century.

Although the theory of the four elements is not necessarily an atomistic theory, it obviously lends itself to interpretation in atomistic terms, namely, when the elements are conceived as smallest parts that are immutable. In this case, all observable changes are reduced to the separation and commingling of the primitive elementary substances. Thus Parmenides' thesis that being is immutable is maintained, whereas the absolute unity of being is abandoned. Yet, the fact that the infinite variety of forms and changes in nature is reduced to just one type of process between only four elementary kinds of atoms shows its affinity with the thesis of the unity of all being.

Notwithstanding the great disparity between the theory of the four elements and modern chemistry, it is clear that modern chemistry with its approximately 100 qualitatively different atoms falls into the same class of atomic theories as that of Empedocles. There are differences, of course, but these will be deferred for later discussion.

More removed from the original thesis of Parmenides was the theory of his contemporary Anaxagoras of Clazomenae, which assumed as many qualitatively different "atoms" as there are different qualitied substances in nature. Inasmuch as these atoms, which Anaxagoras called "seeds," were eternal and incorruptible, this theory still contains an idea borrowed from Parmenides. A special feature of Anaxagoras' theory was that every substance contains all possible kinds of seeds and is named after the kind of seed that predominates in it. Since the substance contains also other kinds of seed, it can change into something else by the separation of its seeds.

Another interesting form of Atomism with inherently qualitied atoms, also based on the doctrine of the four elements, was proposed by Plato. On mathematical grounds he determined the exact forms that the smallest parts of the elements must have. Fire has the form of a tetrahedron, air of an octahedron, water of an icosahedron, and earth of a cube. Inasmuch as he characterized the atoms of the four elements by different mathematical forms, Plato's conception can be considered as a transition between the qualitative and quantitative types of Atomism.

Qualityless atoms

The most significant system of Atomism in ancient philosophy was that of Democritus (5th century BC). Democritus agreed with Parmenides on the impossibility of qualitative change but did not agree with him on that of quantitative change. This type of change, he maintained, is subject to mathematical reasoning and therefore possible. By the same token, Democritus also denied the qualitative multiplicity of visible forms but accepted a multiplicity based on purely quantitative differences. In order to reduce the observable qualitative differences to quantitative differences, Democritus postulated the existence of invisible atoms, characterized only by quantitative properties: size, shape, and motion. Observed qualitative changes are based upon changes in the combination of the atoms, which themselves remain intrinsically unchanged. Thus Democritus arrived at a position that was defined above as Atomism in the strict sense. In order to make the motion of atoms possible, this Atomism had to accept the existence of the void (empty space) as a real entity in which the atoms could move and rearrange themselves. By accepting the void and by admitting a plurality of beings, even an infinite number of them, Democritus seemed to abandon—even more than Empedocles did—the unity of being. Nevertheless, there are sound reasons to maintain that, in spite of this doctrine of the void, Democritus' theory remained close to Parmenides' thesis of the unity of being. For Democritus' atoms were conceived in such a way that almost no differences can be assigned to them. First of all, there are no qualitative differences; the atoms differ only in shape and size. Secondly, the latter difference is characterized by *continuity;* there are no privileged shapes and no privileged sizes. All shapes and sizes exist, but they could be placed in a row in such a manner that there would be no observable difference between successive shapes and sizes. Thus not even the differences in shape and size seem to offer any ground explaining why atoms should be different. By accepting an infinite number of atoms, Democritus retained as much as possible the principle that being is *one.* With respect to the acceptance of the void, it must be stressed that the void in the eyes of Democritus is more nonbeing than being. Thus even this acceptance does not seriously contradict the unity of being.

Atoms as sheer extension. Democritus had declared quantitative differences to be intelligible, because they were subject to mathematical reasoning. Precisely this relationship between quantitative differences and mathematics made it impossible for Descartes (17th century) to think along the atomistic lines of Democritus. If the only thing that is clearly understandable in matter is mathematical proportions, then matter and spatial extension are the same—a conclusion that Descartes did not hesitate to draw. Consequently, he rejected not only the idea of indivisible atoms but also that of the void. In his eyes the concept "void" is a contradiction in terms. Where there is space, there is by definition extension and, therefore, matter.

Descartes's vortices

Yet, however strange it may seem in view of his identification of matter with extension, Descartes offered nonetheless a fully developed theory of smallest particles. To the questions that arise immediately as to how these particles are separated and distinct from each other, Descartes answered that a body or a piece of matter is all of that which moves together. In the beginning of the world all matter was divided into particles of equal size. These particles were in constant motion and filled all of space. As, however, there was no empty space for moving particles to move into, they could only move by taking the places vacated by other particles that, however, were themselves in motion. Thus the motion of a single particle involved the motion of an entire closed chain of particles, called a vortex. As a result of the original motion, some particles were gradually ground into a spherical form, and the resulting intermediary space became filled with the surplus splinters or "grindings." Ulti-

mately, three main types of particles were formed: (1) the splinter materials, which form the finest matter and possess the greatest velocity; (2) the spherical particles, which are less fine and have a smaller velocity; and (3) the biggest particles, which originated from those original particles that were not subject to grinding and became united into larger parts.

Thus Descartes could construct an atomic theory without atoms in the classical sense. Although this theory as such has not been of great value for the scientific atomic theory of modern times, its general tendency was not without importance. However arbitrarily and speculatively Descartes may have proceeded in the derivation of the different kind of corpuscles, he finally arrived at corpuscles characterized by differences in mass, velocity, amount of motion, etc.—properties that could be treated mathematically.

Boscovich's point-centres and Leibniz's monads

Atoms as centres of force: dynamic particles. Most systems of Atomism depict the action between atoms in terms of collision—*i.e.*, as actual contact. In Newton's theory of gravitation, however, action between bodies is supposed to be action at a distance—which means that the body in question acts everywhere in space. As its action is the expression of its existence, it is difficult to confine its existence to the limited space that it is supposed to occupy according to its precise shape and size. There is, therefore, no reason for a sharp distinction between occupied and empty space. Consequently, the mind finds it natural to consider the atoms not as extended particles but as point-centres of force. This conception was worked out by the Dalmatian scientist R.G. Boscovich (1711–87), who attempted to account for all known physical effects in terms of action at a distance between point-particles, dynamic centres of force.

Atoms as psychophysical monads. The idea of applying the atomistic conceptions not only to material but also to psychical phenomena is as old as Atomism itself. Democritus had spoken of the atoms of the soul. According to the principles of his doctrine these atoms could differ only quantitatively from those of the body: they were smoother, rounder, and finer. This made it easy for them to move into all parts of the body. Basically, however, the atoms of the soul were no less material than other atoms.

In Leibniz's monadology the situation was quite different. Leibniz did not first conceive of material atoms and then only later interpret the soul in terms of these atoms; from the beginning he conceived his "atoms," the monads, in terms of an analogy with the soul. A monad is much more a spiritual than a material substance. Monads have no extension; they are centres of action but not, first of all, in the physical sense. Each of the monads is gifted with some degree of perception; each mirrors the universe in its own way. Monads differ from each other, however, in the degree of perception of which they are capable.

The immutability of atoms. By their nature all atomic theories accept a certain degree of immutability of the atoms. For without any fixed units no rational analysis of complex phenomena is possible. At least with respect to the stable factors in the analysis involved, the atoms have to be considered as immutable. According to Atomism in the strict sense, this immutability had to be interpreted in an absolute way.

Absolute vs. relative

The same absolute interpretation appeared in classical chemistry, although its atomic theory deviated from Atomism in the strict sense by assuming qualitatively different atoms and by assuming molecules (rather stable aggregates of atoms). The decisive point, however, is that molecules were formed by mere juxtaposition of atoms without any intrinsic change of the qualities of the atoms. Modern atomic theory, in contrast, gives a less rigid interpretation of the immutability of elementary particles: the particles that build up an atom do not retain their identity in an absolute way.

In some philosophical atomistic theories, the immutability of the atoms has been understood in a highly relative sense. This interpretation arose mainly in the circles of those Aristotelian philosophers who tried to combine atomistic principles with the principle of Aristotle that elements changed their nature when entering a chemical compound. The combination of both principles led to the doctrine known as the *minima naturalia* theory, which holds that each kind of substance has its specific *minima naturalia*, or smallest entities in nature. *Minima naturalia* are not absolutely indivisible: they can be divided but then become *minima naturalia* of another substance; they change their nature. In a chemical reaction the *minima* of the reagents change into the *minima* of the substances that result from the reaction.

Other differences. Atomisms also differ regarding the number of atoms, whether they occupy a void, and how they relate to one another.

Number of atoms. As has already been mentioned, Democritus introduced the hypothesis that the atoms are infinite in number. Although one may question whether the term infinite has to be taken in its strict sense, there is no doubt that by using this term Democritus wanted not merely to express the triviality that, on account of their smallness, there had to be an enormous quantity of atoms. Democritus also had a strong rational argument for postulating a strictly infinite quantity of atoms: only thus could he exclude the existence of atoms that specifically differed from each other (see above *Qualityless atoms*).

When in modern science the problem of the number of atoms arises, the situation is quite different from that of the Greek Atomists. There is now much more detailed information about the properties of the atoms and of the elementary particles, and there is also in astrophysical cosmology some information about the universe as a whole. Consequently, the attempt to calculate the total number of atoms that exist is not entirely impossible, although it remains a highly speculative matter. In a time (around 1930) when all chemical atoms were supposed to be composed of electrons and protons, the pioneering joint-relativity-quantum astrophysicist A.S. Eddington calculated the number of these elementary particles to be $2 \times 136 \times 2^{256}$, or approximately 10^{79}, arguing that, since matter curves space, this is just the number of particles required barely to close the universe up into a hypersphere and to fill up all possible existence states.

Eddington's cosmical number

The existence of the void. To Democritus the existence of the void was a necessary element in atomistic theory. Without the void the atoms could not be separated from each other and they could not move. In the 17th century Descartes rejected the existence of the void, whereas Newton's conception of action at a distance was in perfect harmony with the acceptance of the void and the drawing of a sharp distinction between occupied and nonoccupied space.

The success of the Newtonian law of gravitation was one of the reasons that atomic theories came to prevail in the 18th century. Even with respect to the phenomena of light, the corpuscular and hence atomic theory of Newton, which held that light is made of tiny particles, was adopted almost universally, in spite of Huygens' brilliant development of the wave hypothesis (see LIGHT).

When in the beginning of the 19th century the corpuscular theory of light in its turn was abandoned in favour of the wave theory, the case for the existence of the void had to be reopened, for the proponents of the wave theory did not think in terms of action at a distance; the propagation of waves seemed to presuppose, instead, a medium with not only geometrical properties but with physical ones as well. At first the physical properties of the medium, the ether, were described in the language of mechanics; later they were described in that of the electromagnetic field theory of J.C. Maxwell. Yet, to a certain extent the old dichotomy between occupied and nonoccupied space continued to exist. For, according to the ether theory, the atoms moved without difficulty in the ether, whereas the ether pervaded all physical bodies.

In contemporary science this dichotomy has lost its sharpness, owing to the fact that the distinction between material phenomena, which were supposed to be discon-

tinuous, and the phenomena of light, which were supposed to be continuous, appears to be only a relative one. In conclusion it can be claimed that although modern theories still speak of space and even of "empty" space, this "emptiness" is not absolute: space has come to be regarded as the seat of the electromagnetic field, and it certainly is not the void in the sense in which the term was used by Democritus.

Atoms in external aggregation versus in internal relationship. In most forms of Atomism it is a matter of principle that any combination of atoms into a greater unity can only be an aggregate of these atoms. The atoms remain intrinsically unchanged and retain their identity. The classical atomic theory of chemistry was based upon the same principle: the union of the atoms into the molecules of a compound was conceived as a simple juxtaposition. Each chemical formula (*e.g.*, H_2O, H_2SO_4, NaCl, etc.) reflects this principle through the tacit implication that each atom is still an H, O, or S, etc., even when in combination to form a molecule.

Chemistry had a twofold reason for adopting this principle. One reason was an observational, the other a philosophical one. The fact that some of the properties of a chemical compound could, by simple juxtaposition, be derived from those of the elements (the molecular weight, for example, equals the simple sum of the respective atomic weights) was a strong factual argument in favour of the principle. Many properties of the components, however, could not be determined in this way. In fact, most chemical properties of compounds differed considerably from those of the composing elements. Consequently, the principle of juxtaposition could not be based on factual data alone. It was in need of a more general support. This support was offered by the philosophical idea that inspired all Atomism, viz., that if complex phenomena cannot be explained in terms of aggregates of more elementary factors, they cannot be explained at all.

Holistic tendencies

For the evaluation of this idea, the development of the scientific atomic theory is highly interesting, especially with respect to the interpretation of the concept of an aggregate. Is the only interpretation of this concept that of an assemblage in which the components preserve their individuality—like, for instance, a heap of stones?

Modern atomic theory offers an answer to this question. This theory still adheres to the basic principle that a complex structure has to be explained in terms of aggregates of more elementary factors, but it interprets the term "aggregate" in such a way that it is not limited to a mere juxtaposition of the components. In modern theories atomic and molecular structures are characterized as associations of many interacting entities that *lose* their own identity. The resulting aggregate originates from the converging contributions of all of its components. Yet, it forms a new entity, which in its turn controls the behaviour of its components. Instead of mere juxtaposition of components, there is an internal relationship between them. Or, expressed in another way: in order to know the properties of the components, one has to study not only the isolated components but also the structures into which they enter. To a certain extent modern atomic theory has bridged the gap between atomistic and holistic thought.

HISTORY AND MAJOR REPRESENTATIVES OF THE VARIOUS ATOMISMS

Philosophical Atomism. From the ancient Greeks through the 16th century, Atomism remained mainly philosophical.

Ancient Greek Atomism. It is characteristic of the importance of Greek philosophy that, already in the foregoing exposition of the different aspects of Atomism, several Greek philosophers had to be introduced. Not only the general idea of Atomism but also the whole spectrum of its different forms originated in ancient Greece. As early as the 5th century BC Atomism in the strict sense (Leucippus and Democritus) is found, along with various qualitative forms of Atomism: that of Empedocles, based on the doctrine of the four elements and that of Anaxagoras, with as many qualitatively different atoms as there are different substances.

Criticisms by Plato and Aristotle

Yet, in spite of its successful start, Atomism did not gain preeminence in Greek thought. This is mainly because Plato and Aristotle were not satisfied with the atomistic solution of the problems of change as a *general* solution. They refused to reduce the whole of reality, including man, to a system that knew nothing but moving atoms. Even with respect to the problems of the material world, Atomism seemed to offer no sufficient explanation. It did not explain the observable fact that, notwithstanding continual changes, a total order of specific forms continued to exist. For this reason Aristotle, with Plato, was more interested in the principle of order than in that of the material elements. In his own analysis of change, which resulted in the matter-form doctrine, Aristotle explicitly rejected the thesis of Democritus that in a chemical reaction the component parts retain their identity. According to Aristotle, the elements that entered into a composite with each other did not remain what they were but became a compound. Although there is some indication that in Aristotle's chemical theory smallest particles played a role, it was certainly not a very important one.

Meanwhile, atomistic ideas remained known in Greek thought. Their opponents paid much attention to them, and there were also a few adherents of Democritean Atomism in later times, such as the Greek hedonist Epicurus (*c.* 341–279 BC) and the Roman poet Lucretius Carus (*c.* 95–55 BC) who, through his famous didactic poem *De rerum natura* ("On the Nature of Things"), introduced Atomism into the Latin world.

The elachista of the early Aristotelian commentators. Empedocles had suggested an Atomism with qualitatively different atoms, based upon the doctrine of the four elements. Aristotle adopted the latter doctrine but without its atomistic suggestion. Certain Greek commentators on the works of Aristotle, however, viz., Alexander of Aphrodisias (2nd century AD), Themistius (4th century AD), and Philoponus (6th century AD), combined the Aristotelian theory of chemical reactions with atomistic conceptions. In their systems the atoms were called *elachista* ("very small" or "smallest"). The choice of this term is connected with the Aristotelian rejection of the infinite divisibility of matter. Each substance had its own minimum of magnitude below which it could not exist. If such a minimum particle were to be divided, then it would become a minimum of another substance.

The minima naturalia of the Averroists. The Latin commentators on Aristotle translated the term *elachista* into its Latin equivalent *minima* or also into *minima naturalia; i.e., minima* determined by the nature of each substance. In fact, for most medieval Aristotelians the *minima* acquired little more reality than the theoretical limit of divisibility of a substance; and in their descriptions of physical and chemical processes, they paid no attention to the *minima*. With the Averroists—followers of the Arab Aristotelian Averroës (1126–98)—an interesting development occurred. Agostino Nifo (1473–1538), for example, explicitly stated that in a substance the *minima naturalia* are present as *parts;* they are physical entities that actually play a role in certain physical and chemical processes. Because the *minima* had acquired more physical reality, it then became necessary to know how the properties of the *minima* could be connected with the sensible properties of a substance. Speculations in this direction were developed by the Italian physician, philosopher, and litterateur Julius Caesar Scaliger (1484–1558).

Modern scientific Atomism: Early pioneering work. Modern Atomism arose with the flowering of science in the present sense of the word.

The 17th century. In the history of Atomism the 17th century occupies a special place for two reasons: it saw the revival of Democritean Atomism and it saw the beginning of a *scientific* atomic theory.

The revival of Democritean Atomism was the work of the ambiguous Epicureo-Christian thinker Pierre Gas-

Gassendi, Sennert, and Boyle

sendi (1592–1655), who made his contemporaries not only better acquainted with Atomism but also succeeded in divesting it of the materialistic interpretation with which it was hereditarily infected. This reintroduction of Democritus was well timed. Because of its quantitative character Democritus' Atomism invited for its elucidation the application of mathematics and mechanics, which in the 17th century were sufficiently developed to answer this invitation. In point of fact, the 17th century was more interested in the possibilities that Atomism offered for a physical theory than it was in the philosophical differences between the different atomistic systems. For this reason it saw, for example, hardly any difference between the systems of Gassendi and Descartes, although the latter explicitly rejected some of the fundamentals of Democritus, such as the existence of the void and the indivisibility of the atoms, as noted above (see *Atoms as sheer extension*).

In the case of scientists mainly interested in the chemical aspects, the same shift of emphasis from philosophical to scientific considerations can be discerned. According to the physician and philosopher of nature Daniel Sennert (1572–1637), Democritus' Atomism and the *minima* theory really amounted to the same thing. As far as philosophy was concerned, Sennert was only interested in the general idea of Atomism; the precise content of an atomic doctrine in his view ought to be a matter of chemical experimentation. His own experience as a chemist taught him the specific differences existing between the atoms. In this respect Sennert continued the *minima* tradition. His own contribution to the chemical atomic theory lay in the clear distinction that he made between elementary atoms and the *prima mista*, or atoms of chemical compounds.

The early modern experimentalist Robert Boyle (1627–91) followed the same line of thought as Sennert, but he was much more aware of the discrepancy between Democritus' Atomism and an atomic theory suitable for chemical purposes. Boyle's solution to this problem was the thesis that the atoms of Democritus were normally associated into primary concretions, which did not easily dissociate and which acted as elementary atoms in the chemical sense. These primary concretions can combine to form compounds of a higher order, which may be compared to Sennert's *prima mista* and to the molecules of modern chemistry.

Founding of modern Atomism. The 17th century had laid the theoretical foundations for a scientific atomic theory. For its further development it was in need of better chemical insights, especially concerning the problem of what substances should be considered as chemical elements. Boyle had shown conclusively that the traditional four "elements" were certainly not elementary substances, but at the same time he confessed that he did not yet see any satisfactory method to determine which substances were true elements. This method was provided by another of the principal founders of modern chemistry A.-L. Lavoisier (1743–94): a chemical element is a substance that cannot be further analyzed by known chemical methods.

Lavoisier and Dalton

John Dalton (1766–1844), usually regarded as the father of modern atomic theory, applied the results of Lavoisier's chemical work to atomistic conceptions. When Dalton spoke of elementary atoms, he did not have a merely theoretical idea in mind but the chemical elements as determined by Lavoisier. Dalton held that there are as many different kinds of elementary atoms as there are chemical elements. The atoms of a certain element are perfectly alike in weight, figure, etc.; and the same point applies to the atoms of a certain compound. As weight was the decisive characteristic in Lavoisier's theory, Dalton stressed the importance of ascertaining the relative weights of atoms and the number of elementary atoms that constituted one compound "atom." As to the question of the way in which the atoms are combined in a compound, Dalton conceived this combination as a simple juxtaposition with each atom under the influence of Newtonian forces of attraction. The atoms retain their identity through a chemical reaction. In this one point the founder of the chemical atomic theory did not differ from Democritus.

Recent and contemporary scientific Atomism. Until its development in the third decade of the 20th century, the scientific atomic theory did not differ philosophically very much from that of Dalton, although at first sight the difference may appear large. Dalton's atoms were no longer considered to be immutable and indivisible; new elementary particles sometimes appeared on the scene; and molecules were no longer seen as a mere juxtaposition of atoms—when entering into a compound atoms became ions. Yet, these differences were only accidental; the atoms revealed themselves as composed of more elementary particles—protons, neutrons, and electrons—but these particles themselves were considered then as immutable. Thus the general picture remained the same. The material world was still thought to be composed of smallest particles, which differed in nature and which in certain definite ways could form relatively stable structures (atoms). These structures were able to form new combinations (molecules) by exchanging certain component parts (electrons). The whole process was ruled by well-known mechanical and electrodynamic laws.

In contemporary atomic theory the differences from Dalton are much more fundamental. The hypothesis of the existence of immutable elementary particles has been abandoned: elementary particles can be transformed into radiation and vice versa. And when they combine into greater units, the particles do not necessarily preserve their identity; they can be absorbed into a greater whole.

Atomism in the thought of India. It is interesting to note that atomistic conceptions are not restricted to Western philosophy and science. Examples of qualitative Atomism, based upon the doctrine of the four elements, are also found in Indian philosophy. In some Indian systems the atoms are not absolutely indivisible but only relatively so. In certain aspects Indian Atomism is, therefore, more related to the *minima* doctrine than to the Atomism of Democritus. Indian Atomism has, however, not developed into a scientific theory.

FOUNDATIONAL ISSUES POSED BY ATOMISM

Atomism as a metaphysical system. In discussing Atomism, one particular system, that of Democritus, has been here distinguished as Atomism in the strict sense because of the fact that in no other system have the foundational issues of Atomism been so clearly expressed. Atomism in the strict sense is not merely one of the historical forms of Atomism, one of the many possible scientific attempts at explaining certain physical phenomena; it is, first of all, a metaphysical system: it has been presented as the only possible explanation of change and multiplicity. And as a metaphysic it is rationalistic, mechanistic, and realistic.

Rationalistic, mechanistic, realistic

It is *rationalistic* because it has so much confidence in reason that, in order to explain observed phenomena, it does not hesitate to postulate the existence of unobservable atoms: *i.e.*, of things that are in principle unobservable by the human senses and can be known only by a process of reasoning. Atomists go even further, for they not only are convinced of the existence of atoms but they also think it possible to deduce in a rational way their fundamental properties. Moreover, the description of these properties in *mechanistic* terms is not just a matter of convenience; it is supposed to be the adequate expression of *reality*.

This rationalistic and mechanistic metaphysics is not only characteristic of Democritus' Atomism but also of the early forms of scientific Atomism. The clearest expression of this metaphysics is found in Descartes. For Democritus mechanistic concepts are clear and distinct ideas, so that any further experimental investigation is superfluous. It should be stressed that the atomistic assumption that the human mind, by mere reasoning, can know the properties of the atoms is a necessary consequence of the idea that atoms are not subject to internal change; for the changeless can never be a subject of ex-

perimentation. The great weakness of the mechanistic concept of immutable atoms was that it forced the analyzing experiments to stop at the atoms; but this weakness could reveal itself only after, in the course of the further development of science, the fundamentally experimental character of human knowledge had become evident.

This weakness, in fact, was precisely one of the reasons why Aristotle rejected the Atomism of Democritus, viz., that the latter had postulated atoms that were not subject to change. For Aristotle the very essence of matter was its being subject to change; hence to him the concept of immutable atoms was a contradiction in terms.

Kant's anti-realism and scientific models

Aristotle's criticism of Atomism was clearly directed against its mechanistic metaphysics, not against its realism. This latter characteristic was the target, however, of an attack launched by the incomparable 18th-century epistemologist Immanuel Kant. In a famous argument, known as the antinomy of the continuum, Kant tried to prove that the acceptance of the reality of spatial extension, the cornerstone of Atomism, led to contradictions. His argument can be summarized as follows: It is possible to prove that any compound must be composed of simple things (for if not, there would be nothing but composition). On the other hand, it also is possible to prove that no material thing can be simple, for the very reason that a part of an extended being is always extended and is thus open to division. Hence, every allegedly simple part is at once simple and nonsimple. Consequently, spatial extension cannot be real. Extension is therefore not a property of the material world itself; it is a form imposed upon reality by man's perception.

By his argument Kant did not intend to reject atomic theories as such; he rejected only their realistic pretensions. Kant was deeply convinced that man had to *think* of nature by way of analogy with a mechanism, but he denied that knowledge construed in such a fashion could reach reality itself.

In the 19th century, scientists were, as a rule, hardly impressed by Kant's attack on the realistic pretensions of human knowledge. Scientists had already learned to go their own way and no longer worried about philosophical considerations. Only when an internal crisis in science itself arose were they prepared to reflect upon their presuppositions. Such a crisis occurred in the 20th century when science was forced to accept the relativity of both classical models: the wave and the particle model.

To a certain extent, the problem of whether a scientific model is nothing but a subjective construct in which the scientist unites his experience is the same as the problem that Kant had in mind. One of the differences, however, is that in Kant's time science was still rather exclusively *theory*. Its close connection with praxis (practice, doing) had not yet been discovered. For this reason the Kantian epistemological (or human knowledge) problem could centre on such a question as: what guarantee does the *knowing* subject have that his "models" of reality reflect reality itself? Inasmuch as, in an exclusively theoretical science, the only contact that one has with reality is afforded by means of his knowledge, the problem seems to be insoluble.

The development of science from a theoretical to an experimental discipline forces philosophy to view the epistemological problem in a new way. For in an experimental science man is in a twofold contact with reality, viz., by his knowledge and by his experimental praxis. Modern atomic theory is one of the best examples to illustrate this point. It was this theory that was most directly confronted with the problem of the realistic value of its models. It could take up this challenge because of the theory's effectiveness for experimental praxis, which is the final judge of the realistic value of the theoretical models. It has confirmed the audacious rational speculations of ancient Atomism; but at the same time it has revealed that, in order to be really effective, reason is in need of experimental assistance.

Ancient Greek Atomism versus contemporary scientific Atomism. In comparing Greek Atomism and modern atomic theories, it should be recalled that in Greek thought philosophy and science still formed a unity. Greek Atomism was inspired as much by the desire to find a solution for the problems of mutability and plurality in nature as by the desire to provide scientific explanations for specific phenomena. While it is true that some of the Greek Atomists' ideas can rightly be considered as precursors of later physics, the main importance of the old atomistic doctrines for modern science does not lie in these primitive anticipations. Much more important is the attempt to take seriously the variety and mutability discerned by sense experience and yet to reconcile it with the thesis of Parmenides about the unity and the immutability of matter. In its search for universal and unchangeable laws, modern science is to a great extent inspired by the same idea as Parmenides, since universal laws presuppose a certain unity in the material world and unchangeable laws cannot be established without the presupposition that something unchangeable must be hidden behind all changes. By the same token, without this latter presupposition experiments would not make any sense at all. For if the diversity of reactions occurring under different conditions is to reveal anything, these reactions must be the expression of an immutable nature. The differences have to indicate something about that which remains the same. The great achievement of the Greek philosophers was, therefore, that they took a general view of nature as a whole that made a scientific attitude possible. To this, both the quantitative and the qualitative forms of Atomism contributed, the former drawing attention to the mathematical aspects of the problem, the latter to the observational.

Parmenides and the modern quest for laws

A comparison of ancient Greek Atomism with scientific Atomism merely on the basis of their respective scientific contents would therefore do a great injustice to Greek Atomism; it would in fact misjudge its main value. Such a comparison would, however, also take too narrow a view of modern scientific Atomism. It would imply the philosophical irrelevance of the latter. It has here been shown, however, that the later development of the scientific atomic theory has clarified many philosophical problems that, as basic issues, divided Atomism in the strict sense from other forms of Atomism. To mention only a few examples: the development of the scientific atomic theory has deepened man's insights into the relationship between a whole and its components, into the relative character of ultimate particles, and into certain fundamental epistemological problems.

Evaluation of Atomism. The success of the atomic theory shows the value of the general idea of Atomism: the explanation of the complex in terms of aggregates of fixed particles or units. Its history also shows, however, the inherent danger of this idea, viz., that of absolutism. History has corrected this absolutism: the unitary factors have to be conceived as ultimate only with respect to the complex under consideration, and their union into aggregates need not occur only by way of juxtaposition.

BIBLIOGRAPHY. E. CANTORE, *Atomic Order: An Introduction to the Philosophy of Microphysics* (1969); F. COPLESTON, *A History of Philosophy*, 8 vol. (1950–66); E.J. DIJKSTERHUIS, *Die Mechanisierung des Weltbildes* (1956; Eng. trans., *The Mechanization of the World Picture*, 1961), a history of science from antiquity to the 17th century; A.S. EDDINGTON, *The Philosophy of Physical Science* (1939); K. LASSWITZ, *Geschichte der Atomistik vom Mittelalter bis Newton*, 2 vol. (1890, reprinted 1963), a 19th century classic; A.G.M. VAN MELSEN, *Van atomos naar atoom* (1949; Eng. trans., *From Atomos to Atom: The History of the Concept Atom*, 2nd ed. 1960), including references for the primary sources; L.K. NASH, *The Atomic-Molecular Theory* (1950), a discussion of the first phase of the chemical atomic theory; E.T. WHITTAKER, *History of the Theories of Aether and Electricity*, rev. ed., 2 vol. (1951–54), only for readers with a solid background in science; L.L. WHYTE, *Essay on Atomism: From Democritus to 1960* (1961), a brief introduction to the idea of Atomism and its history.

(A.G.M.v.M.)

Atrophy

Atrophy is a decrease in size of a cell, organ, tissue, or of a part of the body such as a limb. The term implies that

the atrophied part was of a size normal for the individual, considering age and circumstance, prior to the diminution. In atrophy of an organ or body part, there may be a reduction in the number or in the size of the component cells, or in both.

NORMAL ATROPHY

Certain cells and organs normally undergo atrophy at certain ages or under certain physiologic circumstances. In the human embryo, for example, a number of structures are transient and at birth have already undergone atrophy. The adrenal glands become smaller shortly after birth because an inner layer of the cortex has shrunk. The thymus and other lymphoid tissues atrophy at adolescence. The pineal organ tends to atrophy about the time of puberty; usually calcium deposits, or concretions, form in the atrophic tissue. The widespread atrophy of many tissues that accompanies advanced age, although universal, is influenced by changes of nutrition and blood supply that occur during active mature life.

The normal cyclic changes of female reproductive organs are accompanied by physiologic atrophy of portions of these organs. During the menstrual cycle, the *corpus luteum* of the ovary atrophies if pregnancy has not occurred. The muscles of the uterus, which enlarge during pregnancy, rapidly atrophy after the delivery of the child; and after completion of lactation the milk-producing acinar structures of the breast diminish in size. After the menopause the ovaries, uterus, and breasts normally undergo a degree of atrophic change.

ABNORMAL ATROPHY

Whole body atrophy. Atrophy in general is related to changes in nutrition and metabolic activity of cells and tissues. A widespread or generalized atrophy of body tissues occurs under conditions of starvation, whether because food is unavailable or because it cannot be taken and absorbed due to the presence of disease. The unavailability of certain essential protein components and vitamins disturbs the metabolic processes and leads to atrophy of cells and tissues. Under conditions of protein starvation, the body protein is broken down into constituent amino acids, which serve to provide energy and help maintain the structure and cells of the most essential organs. The brain, heart, adrenals, thyroid, pituitary, gonads, and kidneys show less atrophy, relatively, than the body as a whole; whereas the fatty stores of the body, liver, spleen, and lymphoid tissues diminish relatively more than the body as a whole. The brain, heart, and kidneys, organs with abundant blood supply, appear to be the least subject to the wasting effects of starvation.

Associated with the widespread atrophy due to lack of protein is the atrophy of certain tissues that is due to deficiencies of specific vitamins. Atrophic changes of the skin increase because of the lack of vitamin A, and atrophy of muscle increases because of the unavailability of vitamin E.

Senile atrophy

After a growth period of human metabolism, there sets in a gradual decline: slow structural changes other than those due to preventable diseases or accidents occur. Aging eventually is characterized by marked atrophy of many tissues and organs, with both a decline in the number of cells and an alteration in their constitution. This is reflected eventually in the changed, diminished, or lost function characteristic of old age and eventuates in death. The changes in senescence are affected by both inherited constitution and environmental influences, including disease and accident.

Atrophic changes of aging affect almost all tissues and organs, but some changes are more obvious and important. Arteriosclerosis—the thickening and hardening of arterial walls—decreases the vascular supply and usually accentuates aging processes.

Atrophy in old age is especially noticeable in the skin, characteristically flat, glossy or satiny, and wrinkled. The atrophy is due to aging changes in the fibres of the true skin, or dermis, and in the cells and sweat glands of the outer skin.

Wasting of muscle accompanied by some loss of muscular strength and agility is common in the aged. In a somewhat irregular pattern, there is shrinkage of many individual muscle fibres as well as a decrease in their number. Other changes have been observed within the muscle cells.

Increase of the pigment lipofuscin is also characteristic in the muscle fibres of the heart in the aged in a condition known as brown atrophy of the heart. Wasting of the heart muscle in old age may be accompanied by increase of fibrous and fatty tissue in the walls of the right side of the heart and by increased replacement of elastic tissue with fibrous tissue in the lining and walls of coronary arteries within the heart muscle. Abnormal deposits of the protein substance amyloid also occur with greater frequency in the atrophic heart muscle in old age.

Atrophy of the liver in the aged is also accompanied by increased lipochrome pigment in the atrophied cells.

The bones become progressively lighter and more porous with aging, a process known as osteoporosis. The reduction of bone tissue is most marked in cancellous bone —the open-textured tissue in the ends of the long bones— and in the inner parts of the cortex of these bones. In addition to changes in and loss of osteocytes, or bone cells, there is decreasing mineralization, or calcium deposit, with enhanced fragility of the bones.

Atrophy of the brain

Atrophy of the brain in old age is shown by narrowing of the ridges, or gyri, on the surface of the brain and by increased fluid in the space beneath the arachnoid membrane, the middle layer of the brain covering. There is shrinkage of individual nerve cells, with an increase in their lipochrome pigment content, as well as a decrease in their number. Sometimes the nerve fibrils have degenerated, and deposits called senile plaques may be found between the nerve cells, particularly in the frontal cortex and hippocampus (a ridge in the wall of an extension, or horn, of the lateral ventricle, or cavity, of the brain). Similar atrophic changes are seen in the brain in Alzheimer's disease, a condition of unknown cause most likely to occur in older patients. The mental deterioration (senile dementia) of the aged is the clinical manifestation of these changes. Senile atrophy may be increased and complicated by the presence of arteriosclerosis.

Simmonds' disease is a chronic deficiency of function of the pituitary gland that leads to atrophy of many of the viscera including the heart, liver, spleen, kidneys, thyroid, adrenals, and gonads.

A destructive or atrophic lesion affecting the pituitary glands with loss of hormones leads to atrophy of the thyroid, adrenal glands, and gonads and in turn brings atrophic changes to their target organs and the viscera. The decrease in size of the endocrine glands may be extreme.

Atrophy of muscle or of muscle and bone. Local atrophy of muscle, bone, or other tissues results from disuse or diminished activity or function. Although the exact mechanisms may not be completely understood, decreased blood supply and diminished nutrition occur in inactive tissues. Disuse of muscle resulting from loss of motor nerve supply to the muscle (*e.g.*, as a result of poliomyelitis) leads to extreme inactivity and corresponding atrophy. Muscles become limp and paralyzed if there is destruction of the nerve cells in the spinal cord that normally activate them. The shrinkage of the paralyzed muscle fibres becomes evident within a few weeks. Eventually, after some months, fragmentation and disappearance of the muscle fibres occurs with some replacement by fat cells and a loose network of connective tissue. Some contracture may result.

The skeletal tissues forced to inactivity by paralysis (*e.g.*, of a limb as a result of poliomyelitis) also undergo disuse atrophy. If there is a tendency for bone to become lighter and more porous in some particular area, a condition known as local osteoporosis, this can be recognized by X-rays within a few weeks. The cortex of the long bones becomes considerably thinned or atrophic, with decreased mineral content. Disuse as a result of painfully diseased joints, as in rheumatoid arthritis, results in a

Rheumatoid arthritis

similar but lesser degree of atrophy of muscles concerned with movement of the involved joint; and local atrophy may also occur in the bone in the neighbourhood of the joint. A local osteoporosis of bone known as Sudeck's atrophy sometimes develops rapidly in the area of an injury to bone.

Severe or prolonged deficits of blood sugar deprive the nervous system of needed sources of energy and as a rare event result in degeneration of cells of the brain and peripheral nerves. The disuse atrophy of muscle or bone that may result is fundamentally similar to the other disuse atrophies of these tissues.

Persistent pressure will cause atrophy of a compressed cell, organ, or tissue, presumably because of interference with the nutrition and metabolic activity of the affected part. Cells in a local area (*e.g.*, in the liver) atrophy from the pressure of materials such as amyloid deposited around them. The pressure of an expanding benign tumour causes atrophy of adjacent normal structures. The pressure of a localized dilatation of an artery (aneurysm) will cause atrophy of tissues, even bone, on which it impinges.

Bulging of an intervertebral disk or growth of a tumour sometimes brings pressure on nerves near their point of exit from the spinal cord; if the pressure is prolonged, the muscles normally controlled by these nerves may atrophy. Most often the calf muscles are affected. Pressure as a result of involvement of the vertebrae at the level of the neck, or from compression of the network of nerves called the brachial plexus by the *scalenus anticus* muscle, produces similar effects in the upper chest and arms.

Simple disuse of muscle or bone, as, for example, from the immobilization produced when a limb is put in a cast or sling, results in atrophy of these tissues. In the case of muscle, the degree of atrophy is generally less severe than that caused by injury to a nerve, although the nature of the change is similar.

Localized atrophies of leg and arm muscles may result from hereditary or familial diseases in which the nerves of the spinal cord that supply them are inactivated or destroyed. In Charcot-Marie-Tooth disease, the atrophy involves mainly the peroneal muscles, at the outer side of the lower legs, and sometimes the muscles of the hand as well. It commonly begins in childhood or adolescence. Peroneal muscle atrophy is also seen in the hereditary spinal cord degenerative disease known as Friedreich's ataxia.

Atrophy due to injury

Atrophy of nerve tissue. Atrophy of brain or spinal cord tissue may be brought about by injuries that directly affect a localized area or that interfere with the blood supply to an area. When peripheral nerves are severed, degenerative and eventually atrophic changes ensue in the part beyond the injury. This type of atrophy is known as Wallerian degeneration. If conditions do not allow regeneration of nerve fibres from the proximal fragment of the cut nerve, atrophy is the eventual fate of the nerve tissue distal to the injury. Retrograde atrophy also occurs from disuse and affects the ganglion cells of the injured nerve.

Prolonged pressure brings about atrophy in the central nervous system as elsewhere. The pressure of an expanding tumour of the membranes covering the brain results in localized atrophy of the adjacent brain substance on which it impinges. In hydrocephalus more widespread atrophy of brain tissue results from the abnormal amounts of fluid confined within the rigid bony compartment of the skull. Increased pressure within the skull may force a portion of the brain through the *foramen magnum*, the bony opening at the base of the skull, and, if prolonged, results in a localized atrophy of cerebellar tissue pressed against the bony wall.

The late stages of chronic infections may be characterized by atrophy of the brain. A striking example of this is the variety of syphilitic infection of the nervous system known as general paresis in which the brain is shrunk and reduced in weight, the atrophy affecting mainly the cortex of the brain, particularly or most markedly in the frontal area. Occasionally the atrophy is local or affects only one side of the brain. The shrinkage of the brain tissue is mainly due to loss of many nerve cells of the cortex.

Atrophy of fatty tissue. Atrophy of adipose tissue of the body occurs as a part of the generalized atrophy of prolonged undernutrition. Localized atrophy of adipose tissue—lipodystrophy—may be the result of injury to the local area; *e.g.*, repeated insulin injections cause atrophy of fatty tissue at the site of the injections. Progressive lipodystrophy is a disease of unknown cause in which the fatty tissue atrophies only in certain regions of the body. It occurs mainly in women and often begins in childhood; the progressive wasting of adipose tissue affects mainly the face, arms, and trunk. In the affected areas, the specialized fat-holding cells of adipose tissue disappear.

Atrophy of skin. A widespread atrophic change in the skin has been noted as a prominent part of the aging process. Similar atrophic changes in the skin appear to be brought about or enhanced by excessive exposure to sunlight. While a number of abnormal conditions of the skin may include localized atrophic changes in the epidermis or dermis as a part of their lesions, certain generalized diseases of the skin are particularly characterized by such changes. The hardening of the skin known as scleroderma may occur in a localized, or circumscribed, form called morphea or as a more diffuse and severe disease. Advanced stages of scleroderma are characterized by marked atrophy of the tissue and appendages of the true skin. Atrophic thinning of the overlying epidermis also may occur, and the underlying fatty tissue and muscle may atrophy as well. The chronic form of the disease discoid lupus erythematosus also is characterized by atrophy. In advanced stages atrophy occurs particularly in the epidermis in focal areas. The thinned layer of epidermis may be a prominent feature of the microscopic appearance of the skin.

Atrophy of glands. Endocrine glandular tissues may undergo atrophy when an excess of their hormonal product is present as a result of disease. An example is seen in connection with a hormone-producing tumour of the cortical tissue of one adrenal gland, which may be accompanied by marked atrophy of the cortical tissue of the opposite adrenal gland. This probably results from disturbance of the delicate mechanism of hormonal stimulation via the pituitary gland.

Various endocrine organs (thyroid, adrenals, gonads) depend for their activity on endocrine stimulation by hormones of the pituitary gland. A severe general failure of production of the pituitary hormones results in the widespread endocrine atrophy of Simmonds' disease, as has been noted. Lesser degrees of pituitary functional disturbance may disturb a delicate balance, involving mainly one type of stimulating hormone of the pituitary, and may result in selective atrophy of the adrenal cortical tissue or of the gonads.

Glands that release their secretions through a duct (*e.g.*, salivary glands, pancreas) may become atrophic as a result of obstruction of the duct. In the case of the pancreas, a complete obstruction of its duct results in atrophy of the glandular tissue, except for the insulin-producing islets of Langerhans, the secretion of which is absorbed into the bloodstream. Factors of both disuse and increased pressure may be present in the atrophy resulting from obstruction of the outlet channel. Similarly, rapid and complete obstruction of a ureter is followed by atrophy of the corresponding kidney.

Chemical-induced atrophy. Atrophy brought about by chemical injury is not common. In chronic arsenic poisoning, degenerative changes occur in peripheral nerves, resulting in weakness and atrophy in the tissues (usually legs or arms) to which the nerves are distributed. Similar results may follow the peripheral neuropathy of chronic lead poisoning.

BIBLIOGRAPHY. H. UNGER, "Diseases of Aging," in S.L. ROBBINS (ed.), *Pathology*, 3rd ed., ch. 14 (1967), a useful review of the cellular and tissue changes associated with the aging process; J.F.A. MCMANUS, *General Pathology: Biological*

Aspects of Disease (1966), a comprehensive discussion of general pathologic processes; G.R. CAMERON, *Pathology of the Cell* (1952), an extensive review of pathologic changes, with emphasis on those at the cellular level; R.D. ADAMS, D. DENNY-BROWN, and C.M. PEARSON, *Diseases of Muscle* (1953), a comprehensive review of the abnormal conditions affecting muscle; G.P. WRIGHT and W.S. SYMMERS (eds.), *Systemic Pathology* (1967), a good discussion of the atrophic changes occurring in the nervous system; W.A.D. ANDERSON, *Pathology*, 2 vol., 5th ed. (1966), an inclusive text and reference book of pathology that includes a discussion of atrophy as a process and of its effects in various tissues and organs.

(W.A.D.A.)

Attention

Aldous Huxley wrote in the opening chapters of his novel *Island:* ". . . you forget to pay attention to what's happening. And that's the same as not being here and now." If attention is awareness of here and now, how are we aware?

Sense organs and their associated brain structures obviously participate in consciousness, but one may be acutely aware of one's inner state apart from sensing the outer world. Indeed, internal aspects of attention are very rarely absent, since awareness of external events implies recognition, recall, and expectation derived from previous experience. The experience "now" is not instantaneous but includes traces of awareness of past and future. Similarly, the subjective "here," although it hinges on local stimulation on and in the body, involves the full range of the distance receptors of seeing and hearing as well.

The sensory implications of such words as gazing and listening accompany an element of conscious intention, selection, and direction. These features are less evident for other senses, although savouring is distinguished from tasting, and feeling from touching.

Problems of definition

In defining attention, the subjective, colloquial, scientific, and linguistic overtones and ambiguities tend to confound meaning.

Thus, in everyday waking life the feeling of being here and now is discontinuous, efforts to concentrate on immediate sensory data being obscured by periods of absentmindedness, reminiscence, and fantasy. Such states may merge into sleep and dreams in which attention is directed to experiences arising primarily from internal (endogenous) sources. Similarly, severely painful experience may divert attention from almost all else for long periods. Pain is likely to be regarded as a dominant here-and-now experience; recall the expression "I had to pinch myself" to validate a strange or incredible external (exogenous) event. Yet, even the dominance of pain can be overcome by more potent competition for attention; a seriously wounded soldier may not experience pain until the urgencies of battle are over. Such subjective aspects of attention as distraction and motivation are difficult to specify; the soldier himself may be hard put to say whether he was motivated by determination to survive, by political fanaticism, or what.

Definitions of attention as the state of "selective awareness" tend to be circular, as in William James's statement "*My experience is what I agree to attend to,*" implying a prior decision as to what is worthy of attention. Whatever the reasons, volitional or not, only a very small portion of external events that stimulate the sense organs are noticed. Also, the nerve signals generated in response to exogenous and endogenous events are corrupted or distorted by the physiologic mechanisms of the neural structures themselves. The eyes and ears, organs of touch, taste, smell, and limb position show specialized and limited response to changes, differences, and movement; steady or slowly changing states become less noticeable through adaptation in the sense organs, nerve fibres, and central nervous system. With such constraints on attention, verbal reports of external events are often untrustworthy guides to fact, however confidently made.

Early notions of attention. Although closely related to mind-body controversies, the term attention has been somewhat neglected by philosophers.

Gottfried Wilhelm Leibniz (1765) suggested that one's loss of awareness of the constant sound of a waterfall illustrates how events cease to be apperceived (represented in consciousness) without specific attention. He suggested that attention determines what will and will not be apperceived. Although the term is now little used, Immanuel Kant also considered apperception, and it found favour with Wilhelm Wundt, one of the founders of psychology in the 19th century. Wundt wrote of the wide field of awareness (the *Blickfeld*) within which lay the more limited focus of attention (the *Blickpunkt*). He suggested that the range of the *Blickpunkt* was about six items or groups and speculated that attention is a function of the frontal lobes of the brain. W.B. Pillsbury (1908) compared this notion with retinal activity in the eye, in which the central region of high definition is surrounded by a wide field of sensitive but coarse-grained vision.

Attention was equated with "sensible clearness" by Edward Bradford Titchener (1867–1927), while William James (1842–1910) conceived of attention as the active selection of events or stimuli. In combination, these views indicate a process of focalization or concentration of awareness leading to increased clarity of conscious content.

Introspective reports of attention can be supplemented by quantitative study of the effects of selected stimuli on bodily movement or on glandular secretion among laboratory animals and people. Thus, Ivan Petrovich Pavlov (1849–1936) found that secretion of such digestive juices as saliva in hungry dogs often was initiated by auditory signals that were routinely given to precede the ingestion of food. Copious salivary flow suggested that the animals attended to the signals as much as to the food itself. Similarly, movements of head, eyes, ears, and limbs were reliably produced by signals that repeatedly had been followed by such stimuli as electric shocks or bits of food. In such experiments the warning signal is called a conditional stimulus, and the punishment or reward an unconditional stimulus. That is, response to the warning, indicating the animal's attentiveness, is conditional on its past association with the punishment or reward that elicits the salivation or movements unconditionally (*i.e.*, without special training). Since such studies require that the dog be hungry beforehand or "dislike" or "like" the punishment or reward, effects of unconditional stimuli are inconsistent, in a real sense being conditional on the prior state of the animal, depending on what may be called its set, attitude, expectancy, or drive. Related difficulties in appraising attention are variations among individuals and the need to postulate a basic alertness, irrespective of specific drives.

Pavlov reported what is usually termed the orienting response today; in a dog this includes such signs of attention as pricked-up ears, movement of the head and eyes, increased muscular tension, and physiological changes detectable with instruments. The orienting response can be evoked by novel or unexpected events; subjectively identified in human beings with surprise, it is sometimes called the novelty response. In the main, this complex form of behaviour probably represents involuntary, automatic components of attention.

Incorporating the observations of Pavlov, John Broadus Watson (1878–1958) largely was concerned with stimulus-response relations; attention seemed an unnecessary concept in Watson's system (behaviourism), which rejected mentalistic notions such as volition, free will, introspection, and consciousness. Attention was operationally defined, if at all, in terms of discriminative responses to external stimuli. Explanations in terms of single-stimulus situations, however, were found inadequate. Further hypotheses were necessary to account for the effects of competing stimuli, leading to such notions as set, attitude, and expectancy and to renewed interest in attention.

Classification of attentive phenomena. Although attention has been treated here, thus far, as if it were a unitary phenomenon, the term already has been used in sev-

eral senses. Some of these relate to the background state of alertness and arousal rather than to the selective and immediate processes of attending to the here and now. Consider the word attention in everyday usage: familiar motor acts, for example, are performed in response to complex and changing patterns of stimulation with only fleeting attention. A man may be driving his automobile adequately while attending primarily to his own thoughts or his radio. At his destination he will be unable to give a full account of traffic conditions to which, at the time, he responded appropriately.

Pre-attentive processes

This effect suggests so-called preattentive processes similar to those involved in visual search, as in scanning a list for instances of a specific name. Such mechanisms are neither very accurate nor selective; the driver, for example, quickly must become selectively alert when confronted with an emergency. Preattentive processes pass to focal attention, leading to less routine response patterns that are more likely to produce learning. Thus, preattentive processes tend to be limited to the immediate present, whereas more permanent appreciation and storage of information is likely to require an act of well-focussed attention.

Responses tend to become preattentive with practice; once a skill is well learned, less attention is needed than while it was being acquired. Consider the painful degree of attention given to each act during a first driving lesson compared with the experienced driver's almost automatic performance. Clearly this transition is adaptive; in neural terms it seems to free parts of the brain for other demands, while retaining an emergency system that quickly will mobilize these brain centres as required. The mechanism seems to control the balance between width of field and detail of the incoming signals, like that of a zoom lens on a camera. As new, demanding stimuli enter the field, refocussing may be said to narrow the field of attention and sharpen the detail.

Attention in children

Jean Piaget has shown how children tend to anchor attention on dominant objects and distort the stimulus field. This effect (centration) prompts them to overestimate the size of objects, such as geometric figures, in the centre of the field. As a child grows up he learns to deploy attention in ways that compensate for these misperceptions and is able to improve his coordination.

Progressive accumulation of experience also acts to decrease the frequency with which attention is attracted to isolated events. The short attention span (distractibility) of young children is related to their lack of skill in discerning rare from commonplace events. To hold their attention it helps to isolate them in a classroom, even from events that are ordinary by adult standards.

Attention to two (or more) events that tend to occur together is related, within limits, to the improbability of any particular combination. The objective chances of association are reflected in signs of attention that imply a corresponding subjective estimate of probability. One general proposition is that brain mechanisms underlying attention constitute a computer of probabilities and contingencies. In young children and lower animals these mechanisms are held mainly to reconcile direct sensory experience with simple internal states to achieve survival. Among human adults, social influences such as language play a larger part in attracting attention. Thus, a speaker may preface a statement with an imperative ("Look!" or "Listen!") or use such rhetorical devices as irony and provocative questions. Attention to what is said implies that the brain of the listener already has set a high estimate on the probability that the speaker will say something that yields an appreciable reward. Attention will lapse or wander if what is said is too strange, excessively familiar, incoherent, or highly repetitive.

Apparently attention is determined not so much by the intensity or sense modality of signals as by their context, significance, and information content. In the limit, attention may be attracted by the absence, cessation, or omission of a signal or by a slight change in its character. A person may have become accustomed to the ticking of a clock but often is alerted at once should it stop.

In sustained periods between events, vigilance wanes and attention wanders intermittently. Thus an airport observer may fail to notice radar echoes that occur rarely, briefly, and unexpectedly toward the end of his watch and be quite unaware of his inattention. The shorter the signal, the more likely it is to be overlooked during one of these lapses.

Orientation reaction

Orientation reactions (mentioned before in relation to Pavlov) correspond in part to the concept of attention and mobilize action that may be required in a given situation. The pupils of the eyes dilate and allow in more light, hearing becomes more sensitive, ongoing muscular activity is arrested, and muscle tonus rises in readiness. Blood supply is increased to the heart, head, and muscles; respiration and heart rate accelerate; and there are changes in skin resistance to electrical flow. If they are brief, these effects are reported by people as "surprise"; if prolonged, as "excitement." Electroencephalographic (EEG) activity changes in the direction of increased arousal; *i.e.*, desynchronized brain waves of lower amplitude replace usually dominant rhythmic patterns.

Significantly, the orientation reaction becomes progressively weaker and eventually disappears in habituation to regularly repeated presentations of the same stimulus. From such evidence, the reaction seems to be concerned with attentive preparation for action in a novel situation, rather than with routine stimulation.

Determinants of attention. Great variations among individuals in attending to stimuli drove Pavlov to allow for temperament in his general theories of behaviour. A loud sound merely attracts the attention of one dog, while another (similarly reared) animal may scurry to the nearest cover. There is evidence that these differences may be constitutional rather than acquired; dog breeders are familiar with gun-shy strains, even among breeds originally selected for hunting. A noise that merely arouses the interest of one person may panic another, even if he be otherwise courageous or relaxed. These differences depend not only on a lifetime of past experiences but on inborn tendencies as well.

The organic condition of a person also may determine the stimulus to which he will most readily respond at any given moment. He is likely to attend selectively in terms of his immediate biological needs; in hunger he will be especially alert to stimuli related to food.

Social suggestion plays a strong part in determining attention. Two or three persons watching a hole being dug in the road have positively magnetic qualities in attracting a larger crowd. The particular social role played by a person strongly directs his attention. A mother will sleep through considerable noise, but may waken immediately at the slightest cry from her baby. A trained musician will rivet attention on one wrong note that may escape notice by most of the audience. An experienced soldier will duck only when the whistle of an enemy shell rises in pitch.

Novelty as an attractive property of a stimulus indicates the importance to attention of the statistical rarity of a new event. Similarly, it may be said that a complex stimulus tends to attract attention through its greater information content or improbability. Given the alternative, rats tend to choose a longer, complex path over a simple route in a maze. Chimpanzees spend the most time examining the most complex objects around them. Human infants viewing pairs of cards with patterns of different complexity seem to attend most to the more complex; adults also are observed to devote more time to viewing unusual shapes, arrangements, or sequences. Complexity can be defined in this context in terms of the improbability of association (incongruity) of the elements of the pattern or sequence of events. Again, subjective estimates of improbability are based on expectancies developed from past experience and personal temperament.

Personal factors that influence estimation of significance, and therefore thresholds of perception and degrees of attention, may be conscious and cultivated interests, unconscious prejudices, or repressed desires and fears. A person who knows Morse code may be very attentive

to a series of faint random buzzes on the chance that they may convey a coherent message; attention to this possibility even may distract him from other more important aspects of the situation. Another person having "forgotten" an attack by a swarm of bees in infancy may be equally distracted by faint buzzing sounds.

Unconscious factors

The importance of such unconscious influences is emphasized in psychoanalytic theories, particularly in relation to the anomalies of attention and behaviour found among persons with neurotic disturbances (see PSYCHONEUROSES). It is suggested that the subjective significance of commonplace events and objects may be amplified and distorted when they symbolize personal inadequacies or disturbing experiences in early life. These effects are seen in the clinic as compulsive washing rituals, obsessive fears of confinement or of open spaces, and depressive suicidal attempts. Reasons for these disturbances rarely can be elucidated by direct inquiry, but, it is claimed, sometimes they can be deduced from the content of dreams and free association during psychoanalysis.

Physiological mechanisms of attention. Twitchings of the tail or whiskers in a cat give evidence of attention, but there may be little outward indication of degree of attention in a person. With suitable techniques, however, an increase in a person's pulse rate, arrested respiration, increase in muscle tension, and a drop in skin resistance usually can be recorded following a novel or surprising stimulus, these effects varying widely among people. When the signals are familiar, peripheral signs of attention usually are diminished, and often the only signs of attention are in the brain itself.

Bioelectric aspects. Stimulation of any sense organ sets up nerve impulses that constitute the sensory messages relayed to various brain regions. For example, visual signals activate specific receiving areas of the cerebral cortex in the hindmost parts of the brain (occipital lobes), tactile signals reach strips of cortex in the lateral central regions, and auditory signals are relayed to patches on the brain surface just above each ear. There are also large areas of cortex in the frontal and temporal lobes (so-called silent areas) that receive signals from all sense organs. These are nonspecific and receive their information through relays branching from direct pathways in the brain stem and thalamus. The projection to the nonspecific cortex is a diffuse network called the ascending reticular activating system. Access to the nonspecific cortex is also provided by another widely dispersed system (arising in the thalamic relay centres) called the diffuse thalamic projection system.

Integrity of the reticular activating system is essential for wakefulness; reduced activity there (for example, by anesthesia or concussion) results in confusion, somnolence, or unconsciousness, even though exogenous sensory messages still pass to the brain over the direct pathways. The activating system appears to mediate rapid general arousal and attention to novel stimuli, while the diffuse thalamic projection system, which branches more directly from specific relay centres, may convey specific information about the details of signals to which attention has been attracted. The arrangement is more complex than this account would suggest. In human beings, diencephalic structures, particularly the hypothalamus, are involved in regulating states of sleep and wakefulness, and limbic structures, such as the hippocampus, take part in arousal reactions when reward or punishment are involved. Interactions among these systems and their relation to the extraction of information from memory remain obscure.

Brain waves

Responses of the cerebral cortex evoked by signals from the sense organs can be detected as very weak electric pulses through the scalp. This electroencephalographic (EEG) technique involves amplification of the neuroelectric signals, often followed by computer analysis and display. Types of evoked EEG response include: (1) those in the specific receiving areas, almost invariably found, even in sleep; (2) related responses in adjacent (association) zones that habituate with repetition and are maintained by attention to details of sensory events; and (3) responses in nonspecific cortex, mainly in the frontal lobes, that very closely reflect the degree of attention and engagement of the person.

Beyond these evoked responses, intrinsic brain rhythms often are modified by attention to external events and to thinking, imagining, and similar internal activity. The clearest effect of this kind is the inhibition (blocking) of so-called alpha rhythms, usually when the eyes are opened or when the person is thinking about a task, especially one involving visual imagery. Alpha rhythms are more or less regular electric oscillations at a frequency of about 10 Hz (cycles per second) and are most pronounced in the posterior (visual) brain regions. The amplitude and responsiveness of these rhythms vary considerably from person to person. While they tend to be maximized when a person is blank-minded and has his eyes closed, in some persons they persist even during visual activity; in others, there is no sign of alpha rhythms even with the eyes shut. Absence of rhythmic features in the EEG generally is regarded as evidence of arousal as long as there are signs of less rhythmic (asynchronous) activity; total lack of electric discharge is a serious sign of brain morbidity. Desynchronization of the EEG also is typical of sleep when dreams occur. Rapid eye movements at these times are controversially considered signs of attention to dream images (see DREAMS).

Nonspecific cortex responses to external signals are significantly related to overt signs of selective attention. When a person attends to such signals as clicks or flashes by counting them, by noticing changes of intensity, or by pressing a button when they occur, the nonspecific responses in the EEG are augmented in voltage and persist over hundreds of trials. If a subject attends to some other stimulus, the responses evoked by the clicks or flashes are attenuated and may disappear. The person's speed of reaction to the signals tends to increase when the amplitude and consistency of the nonspecific responses are enhanced by focussing attention on the signals; reaction speed is slower when evoked responses are diminished by distraction. Habituation, the decline in attention that results from regular, monotonous stimuli, is accompanied by a progressive decline in the voltage of the evoked brain potentials.

In tasks requiring vigilance, fluctuations in efficiency are accompanied by corresponding changes in evoked potentials. Signals that escape attention evoke lower amplitudes than do detected signals and sometimes show increased latencies (take longer to appear).

When paired stimuli (*e.g.*, two clicks) are presented, the subject's task may be complicated by having him respond to the second signal. In effect, the first signal becomes a warning or conditional stimulus (as in Pavlov's experiments) although the subject may not realize this for several trials. Since the response to the second stimulus is conditional on the instructions of the experimenter, that stimulus may be called imperative rather than unconditional. EEG responses from nonspecific cortex are evoked by both conditional and imperative stimuli in such cases. In addition, a slow cortical voltage change appears that links the responses evoked by the two stimuli. This slowly changing, electrically negative state of the surface of the brain (the contingent negative variation, or CNV) is contingent on the association of the two stimuli and the subject's intention to respond to the imperative second stimulus.

Contingent negative variation

The CNV is a most clear-cut objective correlate of attention. Its characteristic wave form rises slowly to a voltage peak at the moment the imperative signal is expected, dropping sharply to zero as the relevant response or decision is made. When the interval between the two stimuli is less than about half a second, the wave does not develop fully, and reaction time is shortened less than is usual when a longer period (say, up to ten seconds) intervenes. The CNV can be sustained over periods of about ten seconds when enough trials lead the subject to appreciate the association between stimuli. When the interval between conditional and imperative signals is longer, the CNV wave form often shows several peaks and

troughs before the subject's final response, probably a reflection of how attention wanders during the prolonged wait. In normal people the CNV emerges quite consistently and correlates very well with other measures of attention, subjective and objective; it appears regardless of the sense organ stimulated or the intensity of stimulation (*e.g.*, cessation or absence of stimuli, pictorial or verbal events, or situations selected by the subject).

The CNV may function to prime (increase the sensitivity of) the frontal cortex to expected stimuli. Indeed the responsiveness of the cortex to electrical stimulation does, in fact, peak when the CNV wave does. This suggests that direction of attention involves selective control of cortical excitability.

A similar effect (perhaps almost identical in origin) is seen when a person *intends* to perform some action. It is also a slow increase in the electronegativity of the frontal cortex and precedes a spontaneous voluntary act or decision by about one second. In German it has been called the *Bereitschaftspotential* (readiness potential or intention wave). This effect supports an interpretation of intention as a form of attention in which the active component is particularly explicit. Thus, an intention wave even can be produced when the person merely wills to perform an act but does not in fact do so.

Together with bioelectric changes, the rate of brain metabolism increases when a person attends to external events. When attention is drawn to sensory stimuli that require no action, brief rises in blood flow are detected in the appropriate cortical sensory receiving zones. When an act is performed, a more prolonged increase in energy exchange is observed that also involves the frontal lobe and brain parts that control bodily movement.

When attention is directed to a long sequence of related events, such as a conversation, a protracted negative variation in voltage often develops over the frontal cortex. This effect presumably is connected with the maintenance of attention to the transition probabilities and significance of the signals in the conversation.

If attention is concentrated on a given set of stimuli and the decision or action required by them, the CNV is at its highest voltage. Under distraction from a variety of competing stimuli, the wave is at its smallest amplitude. The distraction may come from an external source or from such inner experiences as stomach pain or ideational activities. Unlike the nonspecific evoked responses, the CNV does not appear to habituate. In a very long series of trials, however, it can be seen to weaken as attention begins to wander from the central task.

The distribution of the evoked responses, CNVs, and intention waves over the frontal cortex is very wide but patchy. The limited brain zones in which they arise are remarkably constant for the same person when a given situation recurs, but when attention is divided, each of several simultaneous electrical signals involves its distinctive mosaic of cortical precincts. In EEG recordings from the scalp, the separate potentials from the various patches summate to produce a larger potential than that seen when only a single signal draws attention. This suggests that the frontal cortex can carry on several simultaneous attentive operations with some adequacy. This capacity, however, is sharply limited and varies considerably from person to person. It is very difficult to attend to certain actions in combination. For example, most people must struggle to rotate an arm smoothly clockwise while a leg on the same side moves anticlockwise; attention seems to fluctuate from one to the other, the neglected limb tending to follow the one on which attention is concentrated.

Competition for attention

Microelectrodes implanted in the visual cortex of laboratory animals reveal cells that uniquely respond to specific patterns of visual stimulation and others that respond to movement in a certain direction. Some groups of cells fire more impulses when the animal is alerted; in other groups the rate of firing is inhibited when there are signs of wakefulness and alertness. Some assemblies of cells in the auditory cortex respond as a unit only when the animal seems to attend to sound waves. Near these so-called attention units are other cell groupings that respond whether or not auditory stimuli appear to draw attention. Among humans, also, there are specific regions and cell groups fairly close together in the brain that respond quite differently to apparent attentional and distracting effects and to habituation. In the human thalamus, for example, some cells respond persistently to a touch on the skin, however frequently presented and whatever the distraction from other stimuli. In a nearby part of the thalamus, other cells seem to detect only novel, nonrepetitive stimuli.

Brain responses also relate attention to the conditional and semantic aspects of an event. Thus a television picture may be used as a conditional stimulus to precede an imperative signal. If the scene presented is especially interesting, however, the conditional effect may be impaired while the subject attends to the intriguing details. The appearance of the CNV is delayed, its amplitude is diminished, and reaction time to the imperative stimulus is prolonged.

Subliminal stimuli

Brain responses are of particular value in measuring the hearing acuity of small children or others who may not be able to communicate clearly. The stimulus intensity at which brain responses can be detected usually coincides with the subjective threshold (with properly adjusted EEG equipment). EEG responses sometimes can be detected below the level of subjective awareness when the signals are exciting or provocative. This relates to effects of so-called subliminal stimuli; *i.e.*, signals for which the subject reports no awareness but that sometimes can influence behaviour. As evidenced by brain responses, the threshold of perception is related not merely to stimulus intensity but to the subjective significance of the stimulus as derived from the unconscious effects of personal experience, temperament, and social factors.

Among patients suffering from chronic anxiety, brain responses to all stimuli generally are pronounced, widely distributed, and slow to habituate. Such patients have difficulty in initiating and in maintaining selective attention; they report most events as menacing and use such language as "Everything keeps coming in on me." Symptoms sometimes can be relieved by selective surgical interruption (leucotomy) of nerve tracts connecting the orbital frontal cortex (just above the eye sockets) with deeper brain structures called the thalamus and limbic system. That system seems to involve emotional components of attention (see EMOTION). When these components are exaggerated the person shows fear or even panic. In earlier versions of the leucotomy (lobotomy) operation, large lesions were made in the nerve tracts, and reduction of fear sometimes went so far as to interfere grossly with ability to sustain attention.

Interpersonal influences on attention are seen most clearly in hypnosis. Hypnotized deeply, a person (on the suggestion of the hypnotist) may show no signs of attending to a stimulus that normally produces excruciating pain, may report hallucinatory experiences of people and other objects, or deny seeing selected objects around him. Evoked brain responses during hypnosis are not distinctive, but the CNV can be controlled by suggesting to the subject that given stimuli are more (or less) important, regardless of the objective significance of such stimuli.

Biochemical aspects. Complex chemical changes underlie the electrical correlates of attention in the nervous system. In mediating attention, the brainstem reticular formation, hypothalamus, limbic system, and their respective projections all depend on the neurochemical called noradrenaline for transmission of impulses from one nerve cell to the next. The thalamic arousal system similarly functions through another chemical (acetylcholine), while cortical activity is associated with release of amino acids, particularly glutamic acid and aspartic acid during prolonged attentiveness (see NERVOUS SYSTEM, HUMAN).

The delicate balance of structures that serve the attentive process is easily disturbed by changes in body chemistry. A lowered level of sugar in the blood after fasting,

decreased carbon dioxide from excessive deep breathing, lack of oxygen in diseases of the heart or lungs, or from asphyxia may cloud consciousness, increase suggestibility, and even make the person hallucinate. Stimulants such as caffeine (in coffee, tea, cocoa, and some cola drinks) or amphetamine drugs enhance alertness, but at the same time may increase distractibility. Ethyl alcohol (as in whiskey and beer) in small doses sometimes seems to assist attention by relieving anxiety but in larger doses blunts sensory and motor selectivity. Most dramatic effects are seen with such hallucinogenic drugs as mescaline and lysergic acid diethylamide (LSD). Users of these drugs report spectacular intensification of visual experiences. Hallucinations of vivid colours and majestic scenery may so rivet attention that an LSD-intoxicated person may walk out of a high window or wander about in dangerous street traffic. The less dramatic effects of marijuana (*Cannabis sativa*) also may heighten attention to detail and shift the person's assessment of its importance; sometimes there is distortion of the time sense (see TIME PERCEPTION). As far as these drugs now are understood, they seem to be toxic, poisoning brain mechanisms involved in signal selection and evaluation; the word intoxicated literally means poisoned.

Information theory. Dependence of attention both on the unexpectedness of events and on their familiar association may seem paradoxical. This may be resolved, or at least redefined, by considering events (occurring as sequences in time, or as features of a spatial pattern) in relation to the information they may convey.

The formal approach called information theory suggests that the significance of any event only can be estimated in terms of what else might have happened. Its significance (and tendency to attract attention) is considered a function of its statistical improbability. Novelty, as estimated through the number of times an event has been experienced before (compared with all possible alternatives), provides a measure of so-called surprise value. Thus an event that never has been experienced before has a high surprise value and should attract attention, even without specific associations or consequences.

Research suggests that when sensory stimulation exceeds a given level (called the channel capacity) transmission becomes restricted. One perceptual channel may transmit at the expense of others, or there may be partial transmission of information from more than one source, with attendant error and uncertainty. (A listener may attend almost entirely to one of two simultaneous conversations or shift momentarily from one to the other.)

Investigators often use a unit of information called the bit (short for binary digit), reflecting a primitive choice between two alternatives; *i.e.*, yes or no. The limited success of efforts to apply such measures to human beings reflects the incalculable effect of past experience on the information carried by any objectively discernible bit.

Humans compared with computing devices

One view is that people can deal with data only from one source at a time. Stimulation arriving while a person is reacting to a previous signal theoretically has to wait until his so-called decision mechanism becomes free, as in a telephone line. This single-channel hypothesis arose when it was noticed that people do not track a moving target smoothly but make discrete corrections for misalignment spaced about half a second apart. They seemed to be performing as intermittent-correction servomechanisms (*e.g.*, thermostats in home heating systems). Some interpreted this to reflect a time lag in central (brain?) processes, in crude analogy to the refractory phase of a nerve.

The single-channel model accounts for evidence that only one of a set of signals that require different actions gets attention at a time. If the interval between successive signals is less than about 0.08 of a second, however, grouping seems to occur and movements can be coordinated, as in violin playing or typing.

D.E. Broadbent postulated some form of filter in the central nervous system that admits only selected data, deferring those that cannot receive immediate attention. Deferred information seems to be stored temporarily, being irrecoverable if barred from awareness too long. Simultaneous sensory messages are said to be processed in parallel initially and then to converge on a perceptual or decision channel of limited capacity. The information (in bits per second) carried by incoming messages is believed to be a critical element in the number of stimuli that can be perceived. It has been suggested that the filter serves to weaken rather than to eliminate signals, these still being detectable by appropriately attuned structures. Sleep is depicted as having a similar (presumably more general) attenuating effect on response to external stimuli.

Delays and intermittency of reaction

Another hypothesis posits an absolute refractory period following one signal during which others cannot be accepted. Others suggest that perceiving proceeds in discrete time units, incoming data having to wait until the next time sample begins. The exact length of the unit or refractory period variously has been estimated between 0.1 and 0.5 of a second (see TIME PERCEPTION). Still another account invokes the so-called temporal uncertainty effect. People take longer to react when a warning signal appears less than one-fifth of a second before the signal that requires the response. This is attributed to interference by the warning signal with the preparatory process; in single-channel theory the warning signal is said to capture the decision mechanism.

Reaction-time studies also reveal an intermittency effect. When pressing a switch in response to each of two auditory or visual stimuli, people become slower when the interval between signals is less than 0.25 of a second. This holds even when reaction is required only for the second signal of each pair. Perhaps the effect depends on attention to the first signal (rather than on preparation to respond), since people are not slowed in reacting to a single signal that follows one of their own spontaneous responses. The interval between signals can be varied from 0.05 to 0.5 of a second without significant effect of temporal uncertainty on the delays.

Intermittency has been attributed to control of data processing in the central nervous system by some kind of internal clock that "ticks" off about once every 0.05 of a second. Basically this is a time-sharing variation of single-channel and filter theories. Attention is held to be aligned with just one incoming channel, messages on unattended channels being delayed for the time required to switch attention. Additional lag, seen as the intermittency effect, is said to result from a need to wait for the next "tick." Some suggest that the frequency of EEG alpha waves might be an outward sign of the hypothetical clock. Attentional gating (control of the flow of incoming information, or input) might occur in phase with these brain waves.

In contrast to the view of input overloading is the suggestion that a person acts not through a limited, fixed-capacity channel but as a limited-capacity processor. Indeed, prolonged practice does abolish initial differences in reaction time involving two, four, and eight choices; little practice is needed if stimulus and response are highly compatible. There may be a central processor of limited capacity that calculates ways to serialize input.

Attention also may be explained simply in terms of selective output, placing no restrictions on perceptual processing at all. There is little evidence for a peripheral filter, and selection is more likely to occur in the central nervous system (brain and spinal cord). For example, the content of two competing messages seems to be taken into account before one of them is selected for transmission onward.

A number of so-called perceptual analyzers have been posited, each providing a set of mutually exclusive descriptions. Perhaps there are different levels of analyzers, lower levels dealing with primitive perceptual qualities and higher levels mediating recognition of complex percepts; *e.g.*, those of letters, faces, or words.

Despite all these theoretical uncertainties, there does seem to be evidence of specialized attentive functions for specific parts of the nervous system. Some brain regions are activated by all stimuli in a given sensory modality,

others respond only to novel stimuli, others to a shift from monotony to novelty, others when one stimulus is associated with another, still others when a motor response is intended. Some brain systems seem programmed for particular functions related to situational details. This would suggest continuous neural monitoring of experience in terms of novelty, familiarity, association, and significance. Since attention is subject to voluntary direction, there also must be processes for emphasizing some features and playing down others, as in balancing musical instruments in an orchestra. Although the diffusely interacting structures of the brain provide ample instrumentation, they give no evidence of a physical location for anything resembling a baton-wielding conductor.

BIBLIOGRAPHY. D.E. BROADBENT, *Perception and Communication* (1958), a communication theory approach to the psychology of attention with particular reference to auditory attention; D.N. BUCKNER and J.J. MCGRATH (eds.), *Vigilance: A Symposium* (1963), an account of different theoretical positions and methodological approaches to the psychology of sustained attention and signal detection; C. CHERRY, *On Human Communication* (1957), a constructive review of the applications of communication theory in the human field; C.R. EVANS and T.B. MULHOLLAND (eds.), *Attention in Neurophysiology* (1969), proceedings of an international conference with contributions to the concept of attention from the fields of psychology, physiology, medicine, anatomy, mathematics, and physics; R. LYNN, *Attention, Arousal and the Orientation Reaction* (1965), an examination of the psychophysiological mechanisms of attention and arousal in the light of developments of the Soviet concept of the orienting response; J.F. MACKWORTH, *Vigilance and Habituation* (1969), a neuropsychological approach dealing with attentional factors that influence performance by facilitating the detection and discrimination of signals from their background, and *Vigilance and Attention* (1970), a companion volume specifically concerned with a signal-detection approach to attention; M.J. MELDMAN, *Diseases of Attention and Perception* (1970), an examination of the importance of attention in a wide range of mental and perceptual disorders; A. MCGHIE, *Pathology of Attention* (1969), an account of attentional processes in psychiatric patients and in patients with brain damage; W.B. PILLSBURY, *Attention* (1908), an early, but still much-quoted, text offering a comprehensive account of the philosophy and psychology of attention, a later work being D.I. MOSTOFSKY (ed.), *Attention: Contemporary Theory and Analysis* (1970); A.F. SANDERS (ed.), *Attention and Performance* (1967), a series of papers, mainly by experimental psychologists, on attention in relation to information processing, reaction time, performance, and short-term memory; T. TRABASSO and G.H. BOWER, *Attention in Learning: Theory and Research* (1968), a psychological treatment of attention, with particular reference to its role in learning; W. GREY WALTER, *The Living Brain* (1953), a view of the brain, containing a number of examples of electrophysiological phenomena relevant to attention; A.R. WHITE, *Attention* (1964), a brief account of the philosophy of attention from the viewpoint of one modern philosopher.

(W.G.W./W.C.McC.)

Attila

King of the Huns from 434 to 453, Attila, known in western Europe as the "Scourge of God," was one of the greatest of the barbarian rulers who assailed the Roman Empire. The Huns, a nomadic people from north central Asia, had begun to conquer much of Europe in the 4th century AD. The empire that Attila and his elder brother Bleda inherited seems to have stretched from the Alps and the Baltic in the west to somewhere near the Caspian Sea in the east. Their first known action on becoming joint rulers was the negotiation of a peace treaty with the Eastern Roman Empire, which was concluded at the city of Margus (Požarevac). By the terms of the treaty the Romans undertook to double the subsidies they had been paying to the Huns and in future to pay 700 pounds (300 kilograms) of gold each year.

From 435 to 439 the activities of Attila are unknown, but he seems to have been engaged in subduing barbarian peoples to the north or east of his dominions. The Eastern Romans do not appear to have paid the sums stipulated in the treaty of Margus, and so in 441, when their forces were occupied in the west and on the eastern frontier, Attila launched a heavy assault on the Danubian frontier of the Eastern Empire. He captured and razed a number of important cities, including Singidunum (Belgrade). The Eastern Romans managed to arrange a truce for the year 442 and recalled their forces from the West. But in 443 Attila resumed his attack. He began by taking and destroying towns on the Danube and then drove into the interior of the empire toward Naissus (Niš) and Serdica (Sofia), both of which he destroyed. He next turned toward Constantinople, took Philippopolis, defeated the main Eastern Roman forces in a succession of battles, and so reached the sea both north and south of Constantinople. It was hopeless for the Hun archers to attack the great walls of the capital; so Attila turned on the remnants of the empire's forces, which had withdrawn into the peninsula of Gallipoli, and destroyed them. In the peace treaty that followed, he obliged the Eastern Empire to pay the arrears of tribute, which he calculated at 6,000 pounds (2,700 kilograms) of gold, and he trebled the annual tribute, henceforth extorting 2,100 pounds (950 kilograms) of gold each year.

Attacks on the Eastern Roman Empire

Attila's movements after the conclusion of peace in the autumn of 443 are unknown. About 445 he murdered his brother Bleda and thenceforth ruled the Huns as an autocrat. He made his second great attack on the Eastern Roman Empire in 447, but little is known of the details of the campaign. It was planned on an even bigger scale than that of 441–443, and its main weight was directed toward the provinces of Lower Scythia and Moesia in southeastern Europe—*i.e.*, farther to the east than the earlier assault. He engaged the Eastern Empire's forces on the Utus (Vid) River and defeated them, but he himself suffered serious losses. He then devastated the Balkan provinces and drove southward into Greece, where he was only stopped at Thermopylae. The three years following the invasion were filled with complicated negotiations between Attila and the diplomats of the Eastern Roman emperor Theodosius II. Much information about these diplomatic encounters has been preserved in the fragments of the *History* of Priscus of Panium, who visited Attila's headquarters in Walachia in company with a Roman embassy in 449. The treaty by which the war was terminated was harsher than that of 443; the Eastern Romans had to evacuate a wide belt of territory south of the Danube, and the tribute payable by them was continued, though the rate is not known.

Attila's next great campaign was the invasion of Gaul in 451. Hitherto, he appears to have been on friendly terms with the Roman general Aetius, the real ruler of the west at this time, and his motives for marching into Gaul have not been recorded. He announced that his objective in the west was the kingdom of the Visigoths (a Germanic people who had conquered parts of the two Roman empires) centred on Tolosa (Toulouse) and that he had no quarrel with the Western emperor, Valentinian III. But in the spring of 450, Honoria, the Emperor's sister, sent her ring to Attila, asking him to rescue her from a marriage that had been arranged for her. Attila thereupon claimed Honoria as his wife and demanded half the Western Empire as her dowry. When Attila had already entered Gaul, Aetius reached an agreement with the Visigothic king, Theodoric I, to combine their forces in resisting the Huns. Many legends surround the campaign that followed. It is certain, however, that Attila almost succeeded in occupying Aurelianum (Orléans) before the allies arrived. Indeed, the Huns had already gained a footing inside the city when Aetius and Theodoric forced them to withdraw. The decisive engagement was the Battle of the Catalaunian Plains, or, according to some authorities, of Maurica (both places are unidentified). After fierce fighting, in which the Visigothic king was killed, Attila withdrew and shortly afterward retired from Gaul. This was his first and only defeat.

Invasion of Gaul

In 452 the Huns invaded Italy and sacked several cities, including Aquileia, Patavium (Padua), Verona, Brixia (Brescia), Bergomum (Bergamo), and Mediolanum (Milan); Aetius could do nothing to halt them. But the famine and pestilence raging in Italy in that year compelled the Huns to leave without crossing the Apennines.

In 453 Attila was intending to attack the Eastern Empire, where the new emperor Marcian had refused to pay the subsidies agreed upon by his predecessor, Theodosius II. But during the night following his marriage, Attila died in his sleep. Those who buried him and his treasures were subsequently put to death by the Huns so that his grave might never be discovered. He was succeeded by his sons, who divided his empire among them.

Priscus, who saw Attila when he visited his camp in 448, described him as a short, squat man with a large head, deep-set eyes, flat nose, and a thin beard. According to the historians, Attila was, though of an irritable, blustering, and truculent disposition, a very persistent negotiator and by no means pitiless. When Priscus attended a banquet given by him, he noticed that Attila was served off wooden plates and ate only meat, whereas his chief lieutenants dined off silver platters loaded with dainties. No description of his qualities as a general survives, but his successes before the invasion of Gaul show him to have been an outstanding commander.

BIBLIOGRAPHY. The only comprehensive history of Attila in English is that of E.A. THOMPSON, *A History of Attila and the Huns* (1948); but the sources for his life and times are translated from the Latin and Greek by C.D. GORDON in *The Age of Attila* (1960). For what appear to be the archaeological remains of the Huns, see J. WERNER, *Beiträge zur Archäologie des Attila-Reiches*, 2 vol. (1956).

(E.A.T.)

Attitudes

Social psychologists of the 1920s focussed on attitudes as a central concept in seeking to develop their field as a scientific discipline. Indeed, attitudes seem so pervasive that their study might be said to cover the full range of human behaviour and experience. People tend to develop attitudes toward whatever they experience—toward other people, toward political and religious institutions, toward moral and philosophical systems, apparently toward everything. Yet often enough, attitudes fail to stand the test of logical scrutiny; each person seems to be (as the essayist Charles Lamb, 1775–1834, said of himself) a bundle of prejudice.

DEFINITIONS

The concept of attitude arises from attempts to account for observed regularities in the behaviour of individual persons. One tends to group others around him into common classes; he may assign people of a given skin colour to a single class and behave similarly toward all of them. In such case he is said to hold an attitude specific to that ethnic or racial group. He may lump together the rich or the pious or the lame and so is assumed to bear a particular attitude toward each group. Individuals also classify such objects as paintings or such events as battles and therefore may be considered to have distinctive attitudes toward nonobjective art or toward war.

The quality of one's attitudes is judged from the observable, evaluative responses he tends to make. He might react to everyone of the same ethnic background with expressions of dislike, with derogatory comments about their honesty or intelligence, or he may advocate repressive, exclusionary public policies against them. On the evidence of such negative responses he is said to have an unfavourable attitude toward that ethnic group. Someone who uniformly praises nonobjective paintings, who frequently attends museums that exhibit them, and who hangs their reproductions on his walls is judged to hold a favourable attitude toward nonobjective art.

Inferred attitudes

Attitudes held by others are not directly observable; they must be inferred from behaviour. While one might consult his inner experiences as evidence of his own attitudes, only his public behaviour can receive objective study. Thus, investigators heavily depend on behavioural indices of attitudes—*e.g.*, on what people say, on how they respond to questionnaires, or on such physiological signs as changes in heart rate.

By way of definition, then, attitudes are predispositions to classify sets of objects or events and to react to them with some degree of evaluative consistency. While attitudes logically are hypothetical constructs (*i.e.*, they are inferred but not objectively observable), they are manifested in conscious experience, verbal reports, gross behaviour, and physiological symptoms.

Distinctions. Some authorities see the critical distinction between attitudes and a number of other terms to reside in their relative inclusiveness. All are predispositions to group objects and respond to them in a similar evaluative way. They can be arranged in a hierarchy based on their degree of specificity or exclusiveness. "Values" are said to represent very broad tendencies of this type, "interests" being slightly less inclusive and "sentiments" narrower still; "attitudes" are viewed as still more narrow predispositions, with "beliefs" and "opinions" being progressively the most specific members of this hierarchy. According to this terminology the difference is one of degree rather than of kind.

Other investigators consider one's attitude toward any class to be the intensity with which he expects that group to serve his own values. For example, he may be asked to rate the extent to which he prizes given values (such as health, safety, independence, justice). Then he estimates the degree to which that class (say, politicians) tends to facilitate or impede each value. The sum of the products of these two ratings provides a measure of the individual's attitude toward the group. Thus, if he highly prizes justice and judges that politicians severely interfere with it, his attitude toward that class is taken to be negative.

Attitudes sometimes are regarded as underlying predispositions, and opinions as their overt manifestations. A rarer distinction equates attitudes with unconscious and irrational tendencies and opinions with conscious and rational activities. Others refer to attitudes as meaningful and central and to opinions as more peripheral and inconsequential. A still more popular distinction refers attitudes to matters of taste (*e.g.*, liking a certain country or type of music) and opinions to questions of fact (*e.g.*, whether Zeus exists).

Some apply the term knowledge to what are held to be certainties and attitudes to what is uncertain, even using them to mean "true" and "false" beliefs, respectively. Another suggestion is that attitudes refer to beliefs that impel action and that knowledge is more intellectual and passive.

There are many confusing alternative conventions for distinguishing attitudes from such related concepts as values, opinions, and knowledge. This tends to generate unnecessary dispute and mere proliferation of language. Generally accepted terminology is lacking, and investigators often accept or discard distinctions as they judge them to be useful.

COMPONENTS

The most common analysis invests attitudes with perceptual, emotional, and motivational attributes. These three dimensions or components of human activity were suggested at least as early as Plato and have been applied in a variety of psychological contexts ever since. Traditionally the three aspects are called cognitive (having to do with perceiving, knowing, believing); affective (emotional); and conative (motivational, striving, acting).

Cognitive, affective, and conative attributes

Thus the informational content of any attitude is cognitive. For example, in one's attitude toward an ethnic group the cognitive aspect embraces the stereotyped beliefs (valid or misinformed) he may hold about the group's ability, appearance, habits, and so on.

His feeling of like or dislike for the ethnic group represents the affective aspect of the attitude. The affective component most epitomizes the evaluative nature of attitudes. Such gross physiological signs as accelerated heart rate (sometimes taken as indices of attitudes) reflect affective activity in terms of intensity (either positive or negative). Physiological distinctions between positive and negative attitudes are more difficult to measure (see EMOTION).

One's tendency to exhibit overt behaviour toward the object of an attitude is conative. Depending on the direc-

tion of his attitude, one would be motivated to seek members of a given ethnic group as congenial companions or as targets for his hostility, or he might strive to avoid them (see MOTIVATION).

As a rule, favourable attitudes are characterized by positive directions for all three attributes; unfavourable attitudes tend to involve the reverse. Nevertheless, one may experience hearty dislike without the usual cognitive basis (*i.e.*, without knowing why). Or he may behave aggressively (conative) without feeling hostile (affective). Such unbalanced attitudes often are said to be symptomatic of psychiatric difficulties; frequently they are seen to reflect one's own inner conflicts rather than his objective experiences. So-called prejudices are often attitudes of this sort.

FUNCTIONS OF ATTITUDES

Instrumental functions: attitudes as means to other ends. In seeking social acceptance one may make a show of hostile attitude toward some religious minority held in low esteem by his neighbours; or he may go through the motions of dancing as instrumental to entering desired social groups. Attitudes thus function to facilitate the achievement of goals, retrospectively on the basis of past pleasant experiences or in prospective anticipation of future reward.

Goal facilitation

To the extent that one's attitude serves as a means to an end, the instrumental function, it derives intensity from its perceived effectiveness in goal facilitation. A person who first only acts as if he shares a popular antiminority attitude to facilitate his acceptance in the community may experience no animosity against the minority; should this community prejudice change, he readily could exhibit favourable behaviour without embarrassment. If his overt hostility continued for an extended period, however, he might begin (through learning) to experience the subjective attributes of full-blown prejudice (*e.g.*, a feeling of hate with cognitive beliefs to justify it).

Noetic functions: attitudes as ways of thinking and understanding. Attitudes function in one's daily life in the way theories serve the work of science. One develops attitudes by teasing out regularities in his experience, by learning appropriate reactions in a systematic way, and by stating these relationships as generalizations. As theories, attitudes can be judged true or false, uplifting or degrading, provocative or banal.

Attitudes have the same cognitive role in everyday life that theories have in one's broader philosophical or scientific understanding. One's experiences are so diverse and the range of available responses is so extensive that the simplifications inherent in attitudes are urgent if one is to avoid chaos. If one did not systematically associate classes of objects and events with consistent sets of responses, his life would be an uninterrupted sequence of strange, new problems. To generalize is to simplify and abstract by selectively ignoring many aspects of concrete experience. Nevertheless, despite potential errors, generalizations such as those embodied in attitudes permit the only approach to rational activity.

Expressive functions: attitudes as means for emotional release. Some theorists view human beings as energy systems in which excessive tension can build up (crudely analogous to steam boilers). Thus, the expressive aspect of attitudes is held to relieve psychological pressure as if it were some sort of safety valve. Expressive theorists stress conative-affective attributes of attitudes (noetic theorists emphasize the cognitive) and consider attitudes as useful tension reducers in their own right. People seem to enjoy their attitudes as some are said to enjoy their infirmities. Gratifying feelings of tension release through expressive attitudes are epitomized in such phrases as "It is lovely to be in love" or "The man you love to hate. . . ."

In their expressive function, attitudes provide a number of alternatives to overt action. It is held that, if a person can release hostile tensions through relatively safe attitudinal activity (*e.g.*, fantasy or daydreaming), he doesn't have to act them out with aggressive behaviour. The posited role of attitudes as substitutes for action is reminiscent of the Aristotelian theory of tragedy. According to Aristotle, one can vicariously experience emotional release or purgation (catharsis) of his pity and fear by witnessing a theatrical production of a tragedy.

Ego-defensive functions: attitudes as symptoms of psychiatric disturbance. According to Freudian theorists, one's conscience, his primitive drives, and the demands of the environment all may come into conflict. People are said to adapt to these conflicts by repressing (putting out of awareness) unwelcome urges. One is held to adopt those attitudes that serve to bolster repression; ego functions (those concerned with environmental "realities") thus are thought to be defended against anxiety.

For example, according to one theory, one's prejudice toward groups of nonconformists (*e.g.*, those with a bohemian life style) can be traced to occasions early in his life when he was punished for manifesting any hostility toward his father. Theoretically, he learns to repress angry feelings toward his father and, by extension, toward any figures of authority (*e.g.*, police, politicians, military leaders). To bolster this repression, he is said to cultivate a directly contrasting attitude ("reaction formation") tending to idealize all authorities. He has become an "authoritarian personality" who despises atypical social groups that threaten orthodox social institutions on which his precarious psychiatric adjustment depends. It is held that changes in such ego-defensive attitudes depend on the individual's gaining insight into his own personality difficulties.

DETERMINANTS OF ATTITUDES

Hereditary factors. One has attitudes toward many objects or events that were unknown to his ancestors, and so they could hardly have derived from biological inheritance. The notion that anyone genetically can transmit his learned attitudes to his children seems particularly untenable. Nevertheless, genetic factors do represent the biological potential for all human activities, including the development of attitudes.

Thus, hereditary factors do seem to play some role in determining the general intensity with which attitudes are held or expressed. In dogs, aggressive or hostile predispositions evidently are transmitted genetically. The statistical correspondence is incomplete, however; individual littermates produced by vicious dogs have been observed to be gentler than some bred from more benign animals.

The Swiss psychiatrist Carl Jung (1875–1961) even argued that significant ancestral experiences are communicated genetically as attitudes universal to the human race ("racial unconscious"). While his formulation may have poetic or metaphoric appeal, its literal validity remains very much in doubt.

Bodily states. A host of drugs (including alcoholic beverages) can make one's attitude generally more euphoric, more aggressive, less concerned with others, or whatever. Comparable general alterations in attitude appear when such anesthetic drugs as nitrous oxide (laughing gas) are used; or even when one breathes deeply (hyperventilates) for a few minutes. Transient but spectacular effects on very specific (sexual) attitudes are induced with such aphrodisiac substances as cantharis (obtained from dried beetles). Direct electrical stimulation of the brain as well as its surgical manipulation (lobotomy) can have attitudinal effects.

Role of the brain

Relatively spontaneous physiological changes as in hunger or thirst can sharply alter attitudes toward food or water. Abruptly heightened sexual appetites during adolescence seem to represent a maturational change in the physiological basis for attitudes. Attitudinal disturbances are observed among aged people with hardening of the arteries and during the course of such disorders as tuberculosis, epilepsy, anemia, and encephalitis at any age. Suspicious (paranoid) attitudes toward others are classical long-term symptoms among sufferers of brain syphilis (paresis).

Direct experience. Attitudes are also affected by direct experience. This obvious determinant seems so reason-

able and objective that its importance may tend to be exaggerated.

Effects of frequent interaction with objects of attitudes have been observed in studies of racial integration in the U.S. The evidence suggests that interaction between groups of people initially intensifies pre-existing attitudes; amiably disposed people become even friendlier, while those with negative feelings grow more hostile. With increased familiarity, however, hostility tends to yield to more favourable attitudes in the long run. Amelioration seems most likely when integrated groups share equal social and economic status, when they mutually facilitate their attainment of goals, and when they are able to discover that they share beliefs on many issues.

A single salient encounter also may play a role in determining one's attitude toward an object. Such dramatic confrontations include cases of so-called love at first sight; unreasonable fears (phobias) attributable to single terrifying experiences (*e.g.*, being locked in a closet); and apparently sudden religious conversion after some strikingly moving event (*e.g.*, St. Paul on the road to Damascus).

Most people spend at least part of their lives under the substantial control of others. A nearly universal and usually benign example is one's home environment during the early years of childhood; usually somewhat briefer (and less happy) are adult experiences in such highly controlled environments as hospitals, military installations, jails, schools, and concentration camps.

These so-called total institutions obviously have an important effect on individual attitudes. Typically one undergoes what variously is called rehabilitation, re-education, or brainwashing. Even when institutional concern is only with overt behaviour rather than beliefs, it almost inevitably shapes attitudes to accord with the induced behaviour.

Communication. The major determinant of most of the individual's attitudes is communication from other people. While face-to-face experience is limited to a narrow range of objects, indirect experience with objects through communication from other people is virtually unlimited. There is growing evidence that one's attitudes are affected by nonverbal communication; for example, it has been shown that parents (sometimes unconsciously) can communicate their fears, likes, and dislikes to children through bodily movements and facial expression. But language constitutes the dominant factor in determining attitudes toward objects and events both in direct discussion with others and through mass media (*e.g.*, television). Determination of attitudes through verbal communication constitutes the process of persuasion (*q.v.*).

MEASUREMENT OF ATTITUDES

As a problem in psychological measurement, attitude assessment is subject to the usual question of validity, or in this case, the degree to which any measure accurately reflects an underlying attitude.

Instruments for measuring attitudes differ in terms of the degree to which their purpose is disguised. Surreptitious methods include the use of indices of a person's physiological arousal, for example, dilation of the pupils of his eyes may be photographed while he is viewing pictures of members of different ethnic groups. Or he may be asked to learn different attitudinal statements, the inference being that he will most readily learn those statements that agree with his own attitudes. Undisguised instruments include questionnaires that ask one explicitly to state his own attitudes toward specific issues.

Psychological measurement

Attitude-measuring instruments (disguised or undisguised) also vary in terms of the response options they allow. In a highly constrained format one may be asked to state his attitude by responding to specific questions (or statements) with only one of two alternatives: agree or disagree. Or he may be asked to communicate an attitude under relatively little constraint by writing a paragraph or essay outlining his position.

By far the most commonly used instrument is the undisguised, highly constrained questionnaire. In printed form it is usually called an opinionnaire and consists of a series of evaluative statements to which the person is asked to respond by indicating his position along some scale of agreement or disagreement. It is considered good practice to use several different items for measuring any specific attitude. Wording should be objective, subdued, and moderate; the language of social desirability should be avoided: (*e.g.*, "Chairman Mao says . . . ," "The Queen believes . . . ," "All good Americans prefer . . ."). Items should be so constructed that some demand agreement while others require disagreement if the respondent is to show a consistent attitude. Otherwise the instrument may measure the individual's acquiescent or negativistic tendencies rather than his specific attitudes.

Special care needs to be taken when attitudes are to be measured through interviews. The evidence is that respondents tend to show "interviewer bias" by stating attitudes they perceive to be in agreement with the interviewer's own. A relatively cooperative, compliant person may respond to even the most subtle cues from an interviewer. Careful training of interviewers may reduce such biases; more indirect techniques (*e.g.*, telephone interviews) also may be helpful. Or the person may be asked to fill out a written questionnaire without the interviewer watching.

When assessing attitudes among a relatively large group of people (as in measuring public opinion) it may not be feasible to consult each member of the population under study. A relatively small sample of people, however, if well chosen, can give an accurate estimate of the whole population's attitude. The sample commonly is constituted by randomly selecting people or, better still, by systematically selecting from representative subgroups in the population (*e.g.*, by age, sex, and income). Increasing the size of a random sample tends to increase accuracy, but it may be economically prohibitive. For example, according to statistical estimates, to double the accuracy of the estimate requires quadrupling the size of the sample. Sampling

In using a highly constrained instrument such as an opinionnaire, final scoring of responses is a clear-cut (if sometimes complex) mathematical procedure. In the simplest case (such as agree or disagree) the respondent may be given one unit of score for each item through which he responds favourably toward a given object, the simple sum being taken as an index of favourable attitude. When response scales permit degrees of agreement or disagreement to be recorded, each item might be assigned a different weight in proportion to the intensity of attitude exhibited. More elaborate statistical analyses may be performed on the full set of responses to yield an understanding of specific opinions that comprise the general attitude.

Less constrained instruments (*e.g.*, "open-ended" questions requiring an essay or paragraph) may need to be scored by one or more people serving as judges. In such a preliminary procedure (called content analysis) judges read the responses, attempt to break them down into components, and may score each in terms of favourable quality. While unconstrained instruments are held to allow more freedom and subtlety of expression to respondents, they have the disadvantage of being difficult to score; results depend on the judgments of investigators, just as they do in the construction of opinionnaires. Despite careful attention to technical details attitude measurement remains an intricate and imperfect process.

BIBLIOGRAPHY. A succinct overview of this subject is found in W.J. MCGUIRE, "The Nature of Attitudes and Attitude Change," in G. LINDZEY and E. ARONSON (eds.), *Handbook of Social Psychology*, 2nd ed., vol. 3 (1969). A fuller discussion of the nature of attitudes and their effects on human behaviour and conscious experience may be found in M.J. ROKEACH, *Beliefs, Attitudes and Values* (1968); and in D.T. CAMPBELL, "Social Attitudes and Other Acquired Behavioral Dispositions," in S. KOCH (ed.), *Psychology: A Study of a Science*, vol. 6 (1963). Techniques for measuring attitudes are reviewed extensively in A.N. OPPENHEIM, *Questionnaire Design and Attitude Measurement* (1966); while a wide variety of standardized scales for measuring specific attitudes are pre-

sented in M.E. SHAW and J.M. WRIGHT, *Scales for the Measurement of Attitudes* (1967). A discussion of more covert behavioral measures is contained in E.J. WEBB *et al.*, *Unobtrusive Measures* (1966). J. HARDING *et al.*, "Prejudice and Ethnic Relations," in *Handbook of Social Psychology*, vol. 5 (1969), reviews attitude research concerning interracial relations; R.P. ABELSON *et al.*, (eds.), *Theories of Cognitive Consistency: A Sourcebook* (1968), deals with the role of attitudes in maintaining consistent inter- and intrapersonal relations; C.W. and M. SHERIF (eds.), *Attitude, Ego-Involvement, and Change* (1967), relates attitudinal and perceptual processes.

(W.J.McG.)

Auden, W.H.

In the early 1930s W.H. Auden was acclaimed prematurely by some as the foremost poet then writing in English, on the disputable ground that his poetry was more relevant to contemporary social and political realities than that of T.S. Eliot and William Butler Yeats, who previously had shared the summit. By the time of Eliot's death in 1965, however, a convincing case could be made for the assertion that Auden was indeed Eliot's successor, as Eliot had inherited sole claim to supremacy when Yeats died in 1939.

Horst Tappe



Auden, 1965.

Auden was a counterpart to Eliot not only in transatlantic migration (Auden assuming United States citizenship as Eliot had British) but in his long and controversial career as critic and man of letters as well as poet. He was, as poet, far more copious and varied than Eliot and far more uneven. He tried to interpret the times, to diagnose the ills of society and deal with intellectual and moral problems of public concern. But the need to express the inner world of fantasy and dream was equally apparent, and, hence, the poetry is sometimes bewildering. If the poems, taken individually, are often obscure—especially the earlier ones—they create, when taken together, a meaningful poetic cosmos with symbolic landscapes and mythical characters and situations. In his later years Auden ordered the world of his poetry and made it easier of access; he collected his poems, revised them, and presented them chronologically in two volumes: *Collected Shorter Poems 1927–57* (1966) and *Collected Longer Poems* (1968). A religious poet who is also a clown, a virtuoso who is incorrigibly didactic, a satirist who is also a supreme lyricist is a problem for critics; and most useful criticism of Auden is recent.

Influences and early fame. Wystan Hugh Auden was born in York, England, February 21, 1907. In the next year the family moved to Birmingham, where his father became medical officer and professor in the university. Since the father was a distinguished physician of broad scientific interests and the mother had been a nurse, the atmosphere of the home was more scientific than literary. It was also devoutly Anglo-Catholic, and Auden's first religious memories were of "exciting magical rites." The family name, spelled Audun, appears in the Icelandic sagas, and Auden inherited from his father a fascination with Iceland.

Education and early influences

His education followed the standard pattern for children of the middle and upper classes. At eight he was sent away to St. Edmund's preparatory school, in Surrey, and at 13 to a public (private) school: Gresham's, at Holt, in Norfolk. Auden intended to be a mining engineer and was interested primarily in science; he specialized in biology. By 1922 he had discovered his vocation as a poet, and two years later his first poem was published in *Public School Verse*. In 1925 he entered Oxford (Christ Church), where he established a formidable reputation as poet and sage, having a strong influence on such other literary intellectuals as C. Day Lewis (named poet laureate in 1968), Louis MacNeice, and Stephen Spender, who printed by hand the first collection of Auden's poems in 1928. Though their names were often linked with his as poets of the so-called Auden generation, the notion of an "Auden Group" dedicated to revolutionary politics was largely a journalistic invention. Upon graduating from Oxford in 1928, Auden, offered a year abroad by his parents, chose Berlin rather than the Paris by which the previous literary generation had been fascinated. He fell in love with the German language and was influenced by its poetry, cabaret songs, and plays, especially those by Bertolt Brecht. He returned to become a schoolmaster in Scotland and England for the next five years.

In his *Collected Shorter Poems* Auden divides his career into four periods. The first extends from 1927, when he was still an undergraduate, through *The Orators* of 1932. The "charade" *Paid on Both Sides*, which along with *Poems* established Auden's reputation in 1930, best reveals the imperfectly fused but fascinating amalgam of material from the Icelandic sagas, Old English poetry, public-school stories, Karl Marx, Sigmund Freud and other psychologists, and schoolboy humour that enters into all these works. The poems are uneven and often obscure, pulled in contrary directions by the subjective impulse to fantasy, the mythic and unconscious, and the objective impulse to a diagnosis of the ills of society and the psychological and moral defects of the individuals who constitute it. Though the social and political implications of the poetry attracted most attention, the psychological aspect is primary. The notion of poetry as a kind of therapy, performing a function somehow analagous to the psychoanalytical, remains fundamental in Auden.

Political views in the 1930s

The second period, 1933–38, is that in which Auden was the hero of the left. Continuing the analysis of the evils of capitalist society, Auden also warned of the rise of totalitarianism. In *On This Island* (1937; in Britain, *Look, Stranger!*, 1936) his verse became more open in texture and accessible to a larger public. For the Group Theatre, a society that put on experimental and noncommercial plays in London, he wrote first *The Dance of Death* (a musical propaganda play) and then three plays in collaboration with Christopher Isherwood, Auden's friend since preparatory school: *The Dog Beneath the Skin* (1935), *The Ascent of F 6* (1936), and *On the Frontier* (1938). He also wrote commentaries for documentary films, including a classic of that genre, *Night Mail* (1936); numerous essays and book reviews; and reportage, most notably on a trip to Iceland with MacNeice, described in *Letters from Iceland* (1937), and a trip to China with Isherwood that was the basis of *Journey to a War* (1939). He visited Spain briefly in 1937, his poem *Spain* (1937) being the only immediate result; but the visit, according to his later recollections, marked the beginning both of his disillusion with the left and of his return to Christianity. In 1936 he married Erika Mann, the daughter of the German novelist Thomas Mann, in order to provide her with a British passport. When he and Isherwood went to China, they crossed the United States both ways, and on the return journey they both decided to settle there. In January 1939, both did so.

Change of citizenship

Reorientation of thought and faith. That move begins the third period, 1939–46, during which Auden became a U.S. citizen and underwent decisive changes in religious

and intellectual perspective. *Another Time* (1940) contains some of his best songs and topical verse, and *The Double Man* (containing "New Year Letter," which provided the title of the British edition; 1941) embodies his position on the verge of commitment to Christianity. The beliefs and attitudes that are basic to all of Auden's work after 1940 are defined in three long poems: religious in the Christmas oratorio *For the Time Being* (1944); aesthetic in the same volume's *Sea and the Mirror* (a quasi-dramatic "commentary" on Shakespeare's *Tempest*); and social–psychological in *The Age of Anxiety* (1947), the "baroque eclogue" that won Auden the Pulitzer Prize in 1948. Auden wrote no long poems after that, perhaps because these three together make a kind of triptych.

The fourth period began in 1948, when Auden established the pattern of leaving New York each year to spend the months from April to October in Europe. From 1948 to 1957 his summer residence was the Italian island of Ischia; in the latter year he bought a farmhouse in Kirchstetten, Austria, where he then spent his summers. In *The Shield of Achilles* (1955), *Homage to Clio* (1960), *About the House* (1965), and *City Without Walls* (1969), Auden's only longer works, are sequences of poems arranged according to an external pattern (canonical hours, types of landscape, rooms of a house). With Chester Kallman, a U.S. poet and close friend, he rehabilitated the art of the opera libretto. Their best known are *The Rake's Progress* (1951), with Igor Stravinsky, and *Elegy for Young Lovers* (1961) and *The Bassarids* (1966), with Hans Werner Henze. In 1962 Auden published a volume of criticism, *The Dyer's Hand,* and in 1970 a commonplace book, *A Certain World.* He spent much time on editing and translating, tasks that he performed with unfailing intelligence and zest. Of the numerous honours conferred on Auden in this last period, the Bollingen Prize (1953), the National Book Award (1956), and the professorship of poetry at Oxford (1956–61) may be mentioned.

Assessment

While some hostile critics regarded his migration as a betrayal and his whole development after the 1930s as a decline, most critics and innumerable readers were grateful for the long presence on the literary scene of one so entertaining, gifted, variously instructive, and wise. Robert Lowell, a leading U.S. poet of a younger generation, wrote:

> Auden's work and career are like no one else's, and have helped us all. He has been very responsible and ambitious . . ., constantly writing deeply on the big subjects, and yet keeping something wayward, eccentric, idiosyncratic, charming, and his own. . . . I am most grateful for three or four supreme things: the sad Anglo-Saxon alliteration of his beginnings, his prophecies that seemed the closest voice to our disaster, then the marvellous crackle of his light verse and broadside forms, . . . and finally for a kind of formal poem that combines a breezy baroque grandeur with a sophisticated Horatian simplicity.

In 1972 Auden transferred his winter residence from New York to Oxford, where he was an honorary fellow at Christ Church College. He died in Vienna on September 29, 1973.

MAJOR WORKS

POETICAL WORKS: *Poems* (1928, privately printed by Stephen Spender); *Poems* (1930), including the verse play *Paid on Both Sides,* subtitled "A Charade"; *The Orators* (1932); *Look, Stranger!* (1936; U.S. title, *On this Island,* 1937); *Spain* (1937); *Another Time* (1940); *New Year Letter* (U.S. title, *The Double Man;* includes sonnet sequence "The Quest," 1941); *For the Time Being* (1944), containing *The Sea and the Mirror,* subtitled "A Commentary on Shakespeare's The Tempest" and *For the Time Being,* subtitled "A Christmas Oratorio"; *The Age of Anxiety,* subtitled "A Baroque Eclogue" (1947); *Nones* (1951); *The Shield of Achilles* (1955); *Homage to Clio* (1960); *About the House* (1965); *City Without Walls and Other Poems* (1969).

PLAYS: *The Dance of Death* (1933); written in collaboration with Christopher Isherwood: *The Dog Beneath the Skin* (1935); *The Ascent of F 6* (1936); and *On the Frontier* (1938).

LIBRETTI (in collaboration with Chester Kallman): Igor Stravinsky's *Rake's Progress* (1951); Hans Werner Henze's *Elegy for Young Lovers* (1961) and *The Bassarids* (1966).

TRAVEL BOOKS AND CRITICISM (in collaboration with Louis MacNeice): *Letters from Iceland* (1937), contains the long piece of light verse "Letter to Lord Byron"; with Christopher Isherwood: *Journey to a War* (1939), includes the sonnet sequence "In Time of War," later called "Sonnets from China"; *The Enchafèd Flood,* subtitled "The Romantic Iconography of the Sea" (1950); *The Dyer's Hand and Other Essays* (1962); *Secondary Worlds* (T.S. Eliot Memorial Lectures) (1968); *A Certain World: A Commonplace Book* (1970).

BIBLIOGRAPHY. Full information about criticism of Auden (as well as about the publication of Auden's own works) may be found in B.C. BLOOMFIELD, *W.H. Auden: A Bibliography* (1964); and a selection, with introduction, in M.K. SPEARS (ed.), *Auden: A Collection of Critical Essays* (1964). The most useful recent books are JOHN FULLER, *A Reader's Guide to W.H. Auden* (1970); the interpretations by HERBERT GREENBERG, *Quest for the Necessary* (1968); and JOHN G. BLAIR, *The Poetic Art of W.H. Auden* (1965). M.K. SPEARS, *The Poetry of W.H. Auden* (1963, in paperback with added preface, 1968), remains the most comprehensive account. Earlier books still worth consulting are RICHARD HOGGART, *Auden: An Introductory Essay* (1951), the first full-length study (together with the same author's pamphlet of 1957, "W.H. Auden"); and JOSEPH W. BEACH, *The Making of the Auden Canon* (1957), an analysis of Auden's procedure in compiling his *Collected Poetry* (1945) and *Collected Shorter Poems* (1950).

(M.K.Sp.)

Augustine of Hippo, Saint

St. Augustine, bishop of Hippo in Roman Africa from 396 to 430, and the dominant personality of the Western Church of his time, is generally recognized as having been the greatest thinker of Christian antiquity. His mind was the crucible in which the religion of the New Testament was most completely fused with the Platonic tradition of Greek philosophy; and it was also the means by which the product of this fusion was transmitted to the Christendoms of medieval Roman Catholicism and Renaissance Protestantism.

Alinari—Art Reference Bureau



St. Augustine, fresco by Sandro Botticelli, 1480. In the Church of Ognissanti, Florence.

This unique significance would have belonged to Augustine had he never written the famous *Confessions,* in which at the age of about 45 he told the story of his own restless youth and of the stormy voyage that had ended, as he believed, 12 years before he put it in writing, in the haven of the Catholic Church. It is easy to forget that the real work of Augustine's life did not begin until the last scene of the *Confessions* was already receding for him into a remembered past. Moreover, the *Confessions* themselves are not so much autobiography as they are devotional outpourings of penitence and thanksgiving.

Augustine's conscientious memory generally can be trusted for the facts: his reflections upon them are those of the bishop on his knees. This is not to say that, in any attempt to understand or appreciate the mind of the bishop, the *Confessions* can be neglected. The picture must, however, be drawn in proper proportion; it is essential to avoid giving undue prominence to what should be no more than its background.

Youth and conversion. Hippo Regius is the modern Annaba on the Algerian coast, in what was then the Roman province of Numidia. Augustine, named Aurelius Augustinus, was born on November 13, 354, of middle class parents at Tagaste (modern Souk-Ahras), a small town about 45 miles (72 kilometres) to the south. His father, Patricius, was and remained until late in life a pagan; his mother, Monica, was a Christian of intense but simple piety, from whose early teaching Augustine retained a reverence for the "name of Christ" that never left him. But he was not baptized in infancy. He went through primary and secondary schooling and soon displayed such intellectual promise that the modest family funds were banked upon securing him an academic career that would qualify him for government service. As a 19-year-old student at Carthage he was stirred by the reading of a treatise of Cicero—the now lost *Hortensius*—and was filled with an enthusiasm for "philosophy," which meant not only a devotion to the pursuit of truth but a conviction of the superiority of a life devoted to that pursuit (the *vita contemplativa*) over any aims of secular ambition. The faith of the Catholic Church seemed to him too hopelessly unphilosophical for any man of culture to entertain; and he was easily carried away by the discovery in Manichaeism of a religion that professed to appeal to reason rather than authority.

Enthusiasm for philosophy

Influence of Manichaeism. The Manichaean system as propagated in the Western Roman Empire was a materialistic dualism that accounted for the creation of the world as the product of a conflict between light and dark substances and for the soul of man as an element of the light entangled in the dark. Manichaeism claimed to be the true Christianity, preaching Christ as the Redeemer who enables the imprisoned particles of light to escape and return to their own region. In the Manichaean Church the higher order of "elect" were strictly ascetic and celibate, all physical generation being held to serve the realm of darkness. After an adolescence that probably was no more licentious than was common in his time and country, Augustine had formed a liaison with a woman of low birth by whom he had a son and to whom he remained loyally attached throughout the nine years of his association with the Manichaeans, and he was therefore allowed to join that sect's lower order as one of the "hearers," to whom marriage was permitted as a concession to human weakness.

His first zeal for this "religion of enlightenment" did not last long, however, for the Manichaean experts were intellectually second rate and proved incapable of dealing with the questions he put to them. He became increasingly disillusioned and was already falling into a general agnosticism when, at the age of about 28, he left Carthage, where he had worked as a free-lance teacher of rhetoric, and went to Rome in search of more satisfactory pupils. There he made connections that led to an official professorship at Milan, where the Western emperor then resided. The bishop of Milan was Ambrose, the most eminent Christian churchman of the day. Augustine was introduced to Ambrose but never came to know him well. He went to hear him preach, however, and this, his first contact with the mind of a Christian intellectual, was enough to shake Augustine's prejudice against Catholic teaching. Although he had abandoned the doctrines of Manichaeism, he retained its materialistic presuppositions, which left him still a skeptic with no satisfying alternative to Manichaean notions of ultimate reality. The being of God and the nature and origin of evil remained for him problems as insoluble as they had ever been.

Contact with St. Ambrose

Influence of Neoplatonism. The solution of both problems was given to him by a chance introduction to Neoplatonic writings, for which he may well have been prepared by Ambrose's use of them in some of his sermons. Neoplatonism, in the work of the 3rd-century philosopher and mystic Plotinus, its greatest exponent, is a spiritual monism—a philosophical doctrine holding that there is only one reality—according to which the universe exists as a series of emanations or degenerations from absolute unity. From the transcendent One arises self-conscious mind or spirit; from mind comes soul or life; and soul is the intermediary between the spheres of spirit and of sense. Matter is the lowest and last product of the supreme unity; and since the One is also the real and the good, the potentiality of evil is identified with unformed matter as the point of maximum departure from the One. Evil itself is thus the least real of all things, being simply the privation or absence of good. Neoplatonic mysticism relies on the principle that the inward is superior to the outward: to reach the good, which is the real, one must "return into" oneself; for it is the spirit at the heart of man's inmost self that links him to the ultimate unity.

In the seventh book of the *Confessions*, Augustine tells how in such an act of introspection he found God—the "changeless light," at once immanent and transcendent, which is the source of every intuitive recognition of truth and goodness. This discovery of God was more than the conclusion of a process of reasoning: it was a mystical experience, a vision or touch that came and went. But it left behind it the answer to Augustine's unsatisfied questionings. God is light, and evil is darkness, as the Manichaeans said. But neither is a material substance: the changeless light of God is pure spiritual being, and the evil is nonentity, as darkness is but the absence of light.

Conversion to Christianity. Augustine's mystical experience, his awareness of God, had been momentary and fleeting. He believed that this could be only because he had not made for himself the necessary total identification of supreme value with spirit; he was still himself entangled with the flesh. In fact, Neoplatonism had reinforced the Manichaean principle that the way of return to God must be through escape from the body; and for Augustine this meant primarily and immediately escape from the ties of sexuality. The immortal story of his conversion in the eighth book of the *Confessions* tells of his coming to learn of the heroic achievements of Christian asceticism in East and West, of the self-contempt induced in him by the contrast of his own weakness, and of the final breakdown of resistance in a Milan garden, when, at the sound of a child's voice calling *"tolle, lege: tolle, lege"* ("take up and read"), he opened the New Testament Letters and read in Letter of Paul to the Romans the words, ". . . put on the Lord Jesus Christ, and make no provision for the flesh, to gratify its desires" (Rom. 13:14).

Moral struggle

This was in the late summer of the year 386. Vacation was near, and Augustine resigned his teaching chair and went with some young pupils, his son Adeodatus, and his mother Monica to a reading party at a country house lent by a friend. Out of their literary study and philosophical discussions there came the earliest of Augustine's surviving works—the dialogues, which display so little of the storm and stress of a religious conversion and so little concern with specifically Christian themes that critics have been led to question the accuracy of the *Confessions* story written many years later. It is true that Augustine's struggle against the domination of his sexual nature can be regarded as the final phase in that fluctuating pursuit of the "philosophic life" first presented to him by Cicero's *Hortensius*. But there is no sufficient reason for doubting that he was a Catholic Christian in intention when he received Baptism at the hands of Ambrose in the spring of 387. It is certain that three or four years later, when he wrote his treatise *De vera religione* (*Of True Religion*), he was still thinking of Christianity in Neoplatonic terms. In this treatise, the divine Word (Logos) in Christ is the mind or spirit of Plotinus, illuminating the reason, through whom the human soul has access to the transcendent Godhead. Christ's human life is man's example of the ascetic victory over the pains and pleasures of the flesh; Christian morals serve only to

purify the soul for the life of contemplation; and Christian faith is the necessary acceptance of the church's authority in this preliminary stage of training.

Bishop and Christian philosopher. Shortly after his Baptism, Augustine left Milan, with his mother and a small party of friends, to return to Africa. At Rome's port city of Ostia, his mother died; and Augustine recorded his last talk with her, in which son led mother, through a discourse formed on the pattern of the Neoplatonic "ascent" from this world to the other, to share with him a momentary experience of the life eternal. Home again at Tagaste, the friends formed a little community devoted to the religious life of contemplation and study. But its peace was soon broken when, on a visit to Hippo in 391, Augustine was forced to accept ordination as assistant priest to its old bishop, Valerius. Five years later Valerius died, and Augustine entered the episcopate in which he was to labour until his death. The bishop in Roman Africa was not only the pastor of a parish, the busy teacher and preacher, but the presiding judge in a much-frequented court of summary jurisdiction in civil cases. Augustine never enjoyed robust health, and the vast extent of his literary output was made possible only by the constant services of stenographers and by an extraordinary capacity for the extempore formulation of ordered thought, of which at least 400 sermons remain as proof. He was not a systematic theologian. Much of his writing was in response to the appeals that his growing reputation in the Christian world brought to him for the solution of the most diverse problems. Over 200 of his letters have been preserved, many of them having the scale of minor treatises. He was tireless in controversy with heretics—Manichaeans, Donatists, and Pelagians. But his deepest thought, the real Augustinianism, is to be found in his scripture commentaries and homilies, especially his expositions of the Psalms and his writings on the Gospel and First Letter of John. The characteristic pattern he imposed upon Christian theology was not the outcome of controversy.

Pastor, teacher, and judge in Hippo

The decisive turn was given to his thinking by his ordination to the priesthood, which dragged him against his will from the *vita comtemplativa* into the world and at the same time diverted his studies from philosophy to Scripture. The realities of pastoral experience among the very imperfectly Christianized people of an African seaport, together with the rapid impregnation of his mind with the categories of biblical religion, made it impossible for him to overlook the differences between Neoplatonism and Pauline Christianity. The knowledge of God and of the soul always remained from the time of his Baptism the one and only knowledge that he desired; and Plotinus had not been mistaken in bidding him look within himself if he would find God, for the Bible also tells of a likeness to God imprinted on the soul. But although for the Neoplatonist the soul's likeness to God is that of a, so to speak, reduced divinity, for the Christian it is that of a temporal and mutable image of the "eternal and changeless." Augustine was assured that it is the task of a Christian philosophy, guided by the scriptural revelation, to seek to know God through his image in the soul; and this was the path he followed in his great treatise *De Trinitate* (*On the Trinity*). He insisted that a true knowledge of the soul's nature can be based only on the immediate awareness of self-consciousness; and the soul's awareness of itself is of a trinity in unity that reflects "as in a glass darkly" the being of its Maker. He claimed that knowledge of one's own being, of one's own thinking, of one's own willing is not open to doubt; there is an ego that exists, knows, and wills. But in none of these aspects is the ego self-sufficient or independent: it cannot maintain its own being, produce its own knowledge, or satisfy its own desires. Augustine believed that he had learned from the Platonists to find in God "the author of all existences, the illuminator of all truth, the bestower of all beatitude" (*De civitate Dei* viii, 4). But his theories of the universe, of knowledge, and of ethics were his own. The following three paragraphs summarize these theories.

Biblical and philosophical theologian

Theory of the universe. Creation in Plotinus is motiveless and purposeless, the automatic by-product of the divine self-contemplation; in Augustine its source is "the will of a good God that good things should be" (*De civitate Dei* xi, 21). The outgoing energy of creative love forms the basic principle of his entire theology. Since nothing can come into being or continue in it but by this divine will to create, all that exists is good "in so far as it has being"; and because there are evidently degrees of goodness, there must also be degrees of being. But even the formless matter that is nearest to "not being" is essentially good because God made it; the origin of evil is not to be sought in material existence. Augustine persistently refused to unload upon the material conditions of human life the responsibility for human wickedness.

Theory of knowledge. Following Plato, Augustine argued that the ability to make true judgments never can be inserted into the mind from outside. The human teacher never can do more than help his pupil to see for himself what he already knew without being aware of it. Augustine's favourite examples of these intuitive judgments are the propositions of mathematics and the appreciation of moral values; they are not the construction of the individual mind, because when properly formulated they are accepted by all minds alike. The individual thinker does not make the truth, he finds it; and he is able to do so because Christ, the revealing Word of God, is the *magister interior*, the "inward teacher," who enables him to see the truth for himself when he listens to him.

Ethics. Augustine accepts the basic assumption of ancient ethical theory that conduct is properly directed to the achievement of *eudaimonia*—the happiness or well-being that is taken to be the one universal desire of humanity. Augustine's cosmos is an ordered structure in which the degrees of being are at the same time degrees of value. This universal order requires the subordination of what is lower in the scale of being to what is higher: body is to be subject to spirit, and spirit to God. Man must know his place in the order of the universe and, knowing it, must voluntarily accept it; that is, he must set upon himself and upon everything else the relative value that is properly due. Augustine's word for the ethical valuation that influences conduct is *amor* ("love"). *Amor* is the moral dynamic that impels man to action. If it is rightly directed man will never set a higher value on what is lower in the scale. All lesser goods are to be "used" as means or aids toward the higher; only the highest is to be "enjoyed" as the ultimate end on which the heart is set. The supreme good in whose fruition alone man reaches his perfection is for Augustine the God whose nature is *agapē*, love in the New Testament sense of the word. If, then, man's love, his *amor*, can rise to the enjoyment of God, it will become a participation in the divine *agapē*, love itself. God will have given himself to men, and by sharing in his love men will love one another as he loves them, drawing from him the power to give themselves to others.

Struggle with the Donatist schism. The energies of Augustine, both pastoral and literary, were for the first 15 years of his episcopate distracted by the wearisome struggle to end the schism in the African Church that had persisted for nearly a century. The Donatists, a Christian sect (named after Donatus, one of its leaders) the members of which outnumbered the Catholics in the country districts and in many towns, claimed to be the only true church on the ground that their ministry was the only one the succession of which had not been stained by apostasy in the great persecution of the years 303–313, which had begun under the emperor Diocletian. Imperial attempts to suppress the schism had stimulated the martyr spirit that had always marked African Christianity and gained Donatism the support of strong elements in the native population whose grievances were social and economic rather than ecclesiastical. The schism maintained itself by fanatical violence, and Augustine's persevering attempts to settle the questions at issue by peaceful discussion were fruitless. In the end, the imperial government became convinced that the Donatists were a danger to the security of Africa. The Donatist bishops were compelled to meet their Catholic rivals at a formal conference held

under an official arbitrator at Carthage in 411, the foregone conclusion of which was a Catholic victory.

Answer to the Donatists

Donatists and Catholics were agreed that the power of the Holy Spirit is conveyed to the believer through the sacraments, which are administered by the church through the clergy. The Donatists alleged, however, that the sacraments require for their validity a ministry undefiled by serious sin; for the Spirit departs from the sinner, who cannot therefore "confer what he does not possess." Augustine replied that the sacraments convey the Spirit in virtue of Christ's ordinance alone and that this validity cannot be affected by the worthiness or unworthiness of the human minister. The unity of the church depends on the Spirit's supreme gift of charity, of which schism is the denial. Unfortunately, Augustine, who had for long stood out against the use of any means but persuasion to bring the schism to an end, eventually was induced to approve the enforcement of legal penalties upon the schismatics, in the interest, as he believed, of the many whose fear of Donatist violence had kept them from returning to the church. His famous saying, "Love, and do what thou wilt," was in fact a defense of compulsion in the service of charity.

Struggle with the Pelagian heresy. As the Donatist controversy was ending, the Pelagians were already beginning to threaten the traditional doctrines of sin and redemption in the Western Church. Pelagius had set himself to resist the slackening of Christian moral standards. Against those who pleaded human frailty in excuse for their failings, he insisted that God has made every man alike free to choose and to perform the good; that it is the essence of sin to be a voluntary act that God's law forbids and that the sinner was free to avoid; and that, were not this freedom real, there could be no justice in God's punishments and rewards. This reduction of Christianity to a bleak moralism could not avoid conflict with the plain implications of the church's sacramental and liturgical practice. Baptism had always been "for the remission of sins," and infants were held to need it because they inherit the guilt of Adam's transgression, which, as St. Paul taught, brought death upon the whole race of men. The doctrine of original sin was firmly established in the Western Church before Augustine's time; and when it was openly rejected by Pelagius' disciple Celestius, there was no escape for Pelagianism from being branded as a heresy. The prevarications of Pelagius were able to persuade Pope Zosimus (417–418) to reverse the condemnation pronounced by his predecessor, Innocent I. But in the spring of 418 the African bishops obtained from the emperor Honorius an edict banishing the heretics; and Zosimus was obliged to come into line.

Augustine was the soul of the Church's resistance. He had seen Pelagianism at once as not merely a denial of the virtue of Christian Baptism but also as a fatal misconception of the relationship between God and man. For to assert that man can achieve righteousness by his own effort is to contradict the fundamental truth that God is the giver of all good.

Original sin and its propagation

Before the controversy began, Augustine had worked out his own rationalizations of the doctrines of original sin and divine grace—rationalizations that the church was to prove unwilling to accept fully. He accepted the traditional belief in the fact and in the penal consequences of Adam's transgression, defining the fact as man's refusal to accept his place in the created order, and the consequences as a dislocation of the order of man's own nature—the revolt of flesh against spirit. He argued that not only are all men involved in Adam's guilt and punishment but also that this involvement takes effect through the dependence of human procreation on the sexual passion, in which the spirit's inability to control flesh is evident. It was this linking of original sin with human sexuality that exposed Augustine in his old age to the most damaging criticisms of the Pelagian bishop Julian, who boldly asserted the moral neutrality of the instincts that belong to man's created nature and charged Augustine with relapsing into Manichaeism in his argument that an impulse that a man is bound to fight and conquer must therefore be evil.

For Augustine the fall of man means that in all men the true order of love has been violated. Departing from the love of God above him, man has followed the love of self and become subject to what is below him. Man has fallen by the act of his own will. He cannot by a similar exercise of will reverse the consequences of that fall. The subjection of spirit to flesh is a slavery from which the perverted will has no power to deliver itself, just because it cannot will the deliverance. What is needed is a kind of reversal of gravity—the substitution of an uplifting for a down-dragging love. And Augustine believed that this could happen only by that gracious descent of the divine love to dwell within the sinner: the gospel of the incarnation and of Pentecost.

Pelagius, on the other hand, argued that all men have been created free to do what is right when they see it, and that Christians have received the needed moral enlightenment in Christ's teaching and example. Augustine knew the unreality of the Pelagian conception of freedom as an innate and absolute power of choice, unaffected by circumstances. He pointed to the inescapable conditioning of all moral activity by the situation of the agent—outside whose control are in general not only the presentation of an object but also the kind of feeling that the presentation excites. Moreover, the act of will is dependent on feeling as well as on cognition. In Augustine's words:

> Men will not do what is right, either because the right is hidden from them or because they find no delight in it. But that what was hidden may become clear, what delighted not may become sweet—this belongs to the grace of God" (*De peccatorum meritis et remissione*).

Augustine insisted that without this delight in righteousness there can be no true freedom in well-doing, but only a servile obedience to law. The love of God, which is the motive of the Christian life, must be free. Yet love of God, as St. Paul said, enters man's heart by the gift of the Holy Spirit; and Augustine found it increasingly difficult to leave room in his doctrine of grace for a genuinely free response on man's part to the Spirit's gift. The unexamined assumption that everything in human life must be ascribed either to God's or to man's working compelled him to hold that God alone is the cause of every human movement toward good. In the first year of his episcopate, the study of St. Paul's argument in Rom. 9–11 had convinced him that no event in time can alter the eternal setting of God's will toward any human soul: his elect are chosen before the foundations of the world. God knows—not before, but apart from, the time process—how each individual in the course of time will respond to the particular form in which grace is offered to him; and the elect alone receive the grace that will win their acceptance.

The City of God

The rigour of this doctrine did not soften in face of the Pelagian challenge. In *De civitate Dei* (*The City of God*), the masterpiece on which Augustine was working throughout the Pelagian controversy, he drew a picture, as majestic as it is appalling, of the "beginnings, course and destined ends" of the two invisible societies of the elect and the damned. The work seems to have been in his mind before the capture of Rome by the Visigoths in 410 had shaken the empire; but it took the form of a Christian apologetic against the pagan claim that the disaster was consequence and punishment of Rome's apostasy from its ancestral religion. Augustine's two cities are not to be identified with the Christian Church and the pagan or secular state. They are symbolic embodiments of the two spiritual powers that have contended for allegiance in God's creation ever since the fall of the angels—faith and unbelief, "the love of self extending to contempt for God, and the love of God extending to contempt of self." Neither power is embodied in its purity in any earthly institution; in this world the heavenly and earthly cities are inextricably intermingled. If there is a philosophy of history in the *De civitate Dei*, it is the religious philosophy of predestination.

Augustine found it difficult in his old age to reassure some of his own disciples, to whom his doctrine seemed to make moral effort futile and praise and blame alike

groundless. But he would retract nothing. His last completed treatises drew out the logic of predestination to its most ruthless conclusions. Though his doctrine in its final form was never accepted by the church, it reappeared virtually unmodified in the writings of both St. Thomas Aquinas and John Calvin, the most acute thinkers, respectively, of Scholasticism and Reform. It may indeed be regarded as product of the too audacious attempt of the time-bound human mind to contemplate existence with the eye of the eternal God.

The influence of Augustine. The end of Roman civilization in Africa was near and the Vandal armies were besieging Hippo when Augustine died there on August 28, 430. Not many years later, Vincent of Lérins defined Catholic orthodoxy in the famous phrase, *Quod ubique quod semper quod ab omnibus creditum est* ("What is everywhere, what is always, what is by all people believed"). He dared not call Augustine a heretic in so many words, but it was against the extravagances that he rightly detected in Augustinian doctrine that his definition was aimed. That these extravagances have been a noxious legacy to theology because of their author's authority cannot be denied. But that should not prevent the grateful acknowledgment of the debt that Christian thinking has owed through the centuries to Augustine's influence, which has spanned and may one day reconcile the divisions of Western Christendom. The secret of that influence is to be found not so much in the brilliance and profundity of his intellect, the magic of his style, or the validity of his constructions as in the unique power of his religious genius. St. Anselm of Canterbury, St. Bernard of Clairvaux, the makers of *The Book of Common Prayer*, St. Francis de Sales, Blaise Pascal, Jacques-Bénigne Bossuet, Joseph Butler, Jacques Maritain, Reinhold Niebuhr, and Paul Tillich—all these have in their different ways drawn inspiration from one in whom they have been compelled to recognize "the heart of the matter." *Verus philosophus est amator Dei* ("The true philosopher is the lover of God"). In those words from the *De civitate Dei*, Augustine has left at once the best portrait of himself and the fullest justification of his life's work.

St. Augustine has been revered as a doctor of the church since the early Middle Ages. His feast is celebrated on August 28.

MAJOR WORKS

TEXTS AND TRANSLATIONS: Modern critical editions of St. Augustine's works in the original Latin are in process of publication in the *Corpus Scriptorum Ecclesiasticorum Latinorum* and in the *Corpus Christianorum;* but the only available edition complete except for the Sermons is still that of the Benedictines of Saint-Maur (1670–1700), reprinted in Migne's *Patrologia Latina.* There are no complete English translations of all St. Augustine's works. The largest separate collection is in the series "Nicene and Post-Nicene Fathers of the Christian Church" (N.P.N.F.). Translations of most of the Major Works listed below can be found either in this collection or in one or other of the following more recent series: "Ancient Christian Writers" (A.C.W.); "The Fathers of the Church" (F.C.); "The Library of Christian Classics" (L.C.C.).

GENERAL: *Confessiones* (*c.* 400; *The Confessions,* L.C.C.); *De doctrina Christiana* (397–428; *Christian Instruction,* F.C.); *De Trinitate* (400–416; *On the Trinity,* N.P.N.F.); *De civitate Dei* (413–426; *The City of God,* F.C.); *Enchiridion ad Laurentium de fide, spe, et caritate* (421; *Enchiridion to Laurentius on Faith, Hope, and Love,* L.C.C.); *Sermones* (from 391; *Selected Sermons,* ed. by *Quincy Howe,* 1966); *Epistolae* (from 386; *Letters,* F.C.).

EXEGETICAL: *De Genesi ad litteram* (401–415), a commentary on the first three chapters of Genesis; *De sermone Domini in monte* (393–394; *Commentary on the Lord's Sermon on the Mount,* F.C.); *Enarrationes in Psalmos* (391–420; *Expositions on the Book of Psalms,* 1847–57; A.C.W. incomplete); *Tractatus in Joannis Evangelium* (407–418; *Homilies on the Gospel of John,* N.P.N.F.); *Tractatus in Epistolam Joannis ad Parthos* (*c.* 415; *Homilies on St. John's Epistle,* L.C.C.).

CONTROVERSIAL: (ANTI-MANICHAEAN): *De vera religione* (*c.* 390; *Of True Religion,* L.C.C.); *De libero arbitrio* (389–395; *On Free Will,* L.C.C.). (ANTI-DONATIST): *De Baptismo, contra Donatistas* (400–401; *On Baptism, Against the Donatists,* N.P.N.F.); *Contra litteras Petiliani* (400–403; *Answers to Letters of Petilian,* N.P.N.F.). (ANTI-PELAGIAN): *De spiritu et littera* (412; *The Spirit and the Letter,* L.C.C.); *De natura et gratia* (415; *On Nature and Grace,* N.P.N.F.); *De gratia Christi et de peccato originali* (418; *On the Grace of Christ, and on Original Sin,* N.P.N.F.); *De gratia et libero arbitrio* (426 or 427; *On Grace and Free Will,* N.P.N.F.).

BIBLIOGRAPHY. A fairly recent comprehensive bibliography of works dealing with St. Augustine is CARL ANDRESEN (ed.), *Bibliographia Augustiniana* (1962); T.J. VAN BAVEL and F. VAN DER ZANDE, *Répertoire bibliographique de Saint Augustin* (1963), covers material that appeared between 1950 and 1960. Annual bibliographies are provided in *L'Année philologique* (1924–), ed. by J. MAROUZEAU, in the *Revue des études augustiniennes* (quarterly), and in *Recherches augustiniennes* (1958–).

Biography: The best modern biography, both scholarly and readable, is PETER BROWN, *Augustine of Hippo* (1967). GERALD BONNER, *St. Augustine of Hippo: Life and Controversies* (1963), is also valuable. Of literary interest is REBECCA WEST, *St. Augustine* (1933). See also K. ADAM, *Die geistige Entwicklung des heiligen Augustinus* (1931; Eng. trans., *St. Augustine: The Odyssey of His Soul,* 1932); and HUGH POPE, *Saint Augustine of Hippo,* 2nd ed. (1949). The problems concerning the chronology and nature of Augustine's conversion, especially as related in his *Confessions,* are dealt with in P. AUBIN, *Le Problème de la "Conversion"* (1963); J.M. LE BLOND, *Les Conversions de Saint Augustin* (1950); A.M. LA BONNARDIERE, *Recherches de chronologie Augustinienne* (1965); P. COURCELLE, *Recherches sur les Confessions de Saint Augustin,* new ed. (1968); J.J. O'MEARA, *The Young Augustine* (1954); M. PELLEGRINO, *Les Confessions de Saint Augustin* (1961). Augustine's maturity is described in F. VAN DER MEER, *Augustine de Zielzorger* (1947; Eng. trans., *Augustine the Bishop,* 1961).

Thought: A general outline of Augustine's thought is provided in P. ALFARIC, *L'Évolution intellectuelle de Saint Augustin* (1918); H.I. MARROU, *Saint Augustine et l'augustinisme* (1956; Eng. trans., *St. Augustine and His Influence Through the Ages,* 1957), *Saint Augustin et la fin de la culture antique,* 4th ed. (1958); E. PORTALIE, *A Guide to the Thought of Saint Augustine* (1960). His philosophy is considered in J. BARION, *Plotin und Augustinus* (1935); ETIENNE GILSON, *Introduction à l'étude de Saint Augustin,* 2nd ed. rev. (1943; Eng. trans., *The Christian Philosophy of St. Augustine,* 1960). His political theory and view of history, especially as propounded in the *De civitate Dei,* is the subject of R.H. BARROW, *Introduction to St. Augustine: The City of God* (1950); J.H.S. BURLEIGH, *The City of God* (1949); H.A. DEANE, *The Political and Social Ideas of St. Augustine* (1963); G.L. KEYES, *Christian Faith and the Interpretation of History: A Study of St. Augustine's Philosophy of History* (1966).

Theology: For general accounts of Augustine's theology, see J. BURNABY, *Amor Dei: A Study of the Religion of St. Augustine* (1938, reprinted 1960); H. DE LUBAC, *Augustinisme et théologie moderne* (1965); E.A. TESELLE, *Augustine the Theologian* (1970).

Special topics: For Christology, see T.J. VAN BAVEL, *Recherches sur la Christologie de Saint Augustin* (1954); for the Eucharist, G. LECORDIER, *La Doctrine de l'eucharistie chez S. Augustin* (1930); for biblical exegesis, M. PONTET, *L'Exégèse de S. Augustin, prédicateur* (1946); for predestination and grace, H. RONDET, *Essais sur la théologie de la grâce* (1964).

(Jo.Bu.)

Augustus

Gaius Octavius, subsequently known as Gaius Julius Caesar and still later as Augustus or Caesar Augustus, was the first Roman emperor, following the republic, which had been finally destroyed by the dictatorship of Julius Caesar, his great-uncle and adoptive father. His autocratic regime is known as the principate because he was the *princeps,* the first citizen, at the head of that array of outwardly revived republican institutions that alone made his autocracy palatable. With unlimited patience, skill, and efficiency, he overhauled every aspect of Roman life and brought durable peace and prosperity to the Greco-Roman world.

Gaius Octavius was born on September 23, 63 BC, of a prosperous family that had long been settled at Velitrae (Velletri), southeast of Rome. His father, who died in 59 BC, had been the first of the family to become a Roman senator and was elected to the high annual office of the praetorship, which ranked second in the political hierarchy to the consulship. Gaius Octavius' mother Atia was the daughter of Julia, the sister of Julius Caesar; and it

was Caesar who launched the young Octavius in Roman public life. At the age of 12 he made his debut by delivering the funeral speech for his grandmother Julia. Three or four years later he received the coveted membership of the board of priests (*pontifices*). In 46 he accompanied Caesar, now dictator, in his triumphal procession after his victory in Africa over his opponents in the Civil War; and in the following year, in spite of ill health, he joined the dictator in Spain. He was at Apollonia (now in Albania), completing his academic and military studies when, in 44 BC, he learned that Julius Caesar had been murdered.

By courtesy of the trustees of the British Museum



Augustus, bronze sculpture from Meroe, Sudan, 1st century AD. In the British Museum.

Rise to power. Returning to Italy, he was told that in his will Caesar had adopted him as his son and had made him his chief personal heir. He was only 18 when, against the advice of his stepfather and others, he decided to take up this perilous inheritance and proceeded to Rome. Mark Antony (Marcus Antonius), Caesar's chief lieutenant, who had taken possession of his papers and assets and had expected that he himself would be the principal heir, refused to hand over any of Caesar's funds, forcing Octavius to pay the late dictator's bequests to the Roman populace from such resources as he could raise. Caesar's assassins, Brutus and Cassius, ignored him and withdrew to the east. Cicero, the famous orator who was one of Rome's principal elder statesmen, hoped to make use of him but underestimated his abilities.

Celebrating public games, instituted by Caesar, to ingratiate himself with the city populace, Octavius succeeded in winning considerable numbers of the dictator's troops to his own allegiance. The Senate, encouraged by Cicero, broke with Antony, called upon Octavius for aid (granting him the rank of senator in spite of his youth), and joined in the campaign of Mutina (Modena) against Antony, who was compelled to withdraw to Gaul. When the consuls who were in command of the Senate's forces lost their lives, Octavius' soldiers compelled the unwilling Senate to confer one of the vacant consulships on him. Under the name of Gaius Julius Caesar he next secured official recognition as Caesar's adoptive son. Although it would have been normal for him to add "Octavianus" (with reference to his original family name), he preferred not to do so. Today, however, he is habitually described as Octavian (until the later date when he assumed the designation of Augustus).

The Second Triumvirate

Octavian soon reached an agreement with Antony, as well as with another of Caesar's principal supporters, Lepidus, who had succeeded him as chief priest. On November 27, 43 BC, the three men were formally given a five-year dictatorial appointment as triumvirs for the reconstitution of the state (the Second Triumvirate—the first having been the informal compact between Pompey, Crassus, and Julius Caesar). The east was occupied by Brutus and Cassius, but the triumvirs divided the west among themselves. They also drew up a list of "proscribed" political enemies, and the consequent executions included 300 senators (one of whom was Antony's enemy Cicero) and 2,000 members of the class immediately below the senators, the equites or knights. Julius Caesar's recognition as a god of the Roman state in January 42 BC enhanced Octavian's prestige as son of a god.

He and Antony crossed the Adriatic and under Antony's leadership (Octavian being ill) won the two battles of Philippi against Brutus and Cassius, both of whom committed suicide. Antony, the senior partner, was allotted the east (and Gaul); and Octavian returned to Italy, where difficulties caused by the settlement of his veterans involved him in the Perusine War (decided in his favour at Perusia, the modern Perugia) against Antony's brother and wife. In order to appease another potential enemy, Sextus Pompeius (Pompey the Great's son), who had seized Sicily and the sea routes, Octavian married Sextus' relative Scribonia (though before long he divorced her for personal incompatibility). These ties of kinship did not deter Sextus, after the Perusine War, from making overtures to Antony; but Antony rejected them and reached a fresh understanding with Octavian at the treaty of Brundisium, under the terms of which Octavian was to have the whole west (except for Africa, which Lepidus was allowed to keep) and Italy, which, though supposedly neutral ground, was in fact controlled by Octavian. The east was again to go to Antony, and it was arranged that Antony, who had spent the previous winter with Queen Cleopatra in Egypt, should marry Octavian's sister Octavia. The peoples of the empire were overjoyed by the treaty, which seemed to promise an end to so many years of civil war. In 38 BC Octavian formed a significant new link with the aristocracy by his marriage to Livia Drusilla.

But a reconciliation with Sextus Pompeius proved abortive, and Octavian was soon plunged into serious warfare against him. When his first operations against Sextus' Sicilian bases proved disastrous, he felt obliged to make a new compact with Antony at Tarentum (Taranto) in 37 BC. Antony was to provide Octavian with ships, in return for troops Antony needed for his forthcoming war against the empire's eastern neighbour Parthia and its Median allies. Antony handed over the ships, but Octavian never sent the troops. The treaty also provided for renewal of the Second Triumvirate for five years, until the end of 33 BC.

Military successes. In the following year the balance of power began to change: whereas Antony's eastern expedition failed, Octavian's fleet, commanded by his former schoolmate, Marcus Agrippa, who, although unpopular with the influential nobles, was an admiral of genius, totally defeated Sextus Pompeius off Cape Naulochus (Venetico) in Sicily. At this point the third triumvir, Lepidus, seeking to contest Octavian's supremacy in the west by force, was disarmed by Octavian, deprived of his triumviral office, and forced into retirement. Ignoring Antony's right to settle his own veterans in Italy and recruit fresh troops, Octavian discharged many legionaries and founded settlements for them. His deliberate rivalry with Antony for the eventual mastership of the Roman world became increasingly apparent. Octavian's marriage two years earlier had begun to win over some of the nobles who had previously been Antony's supporters. Octavian also launched elaborate religious and patriotic publicity, centring on the classical god of order, Apollo, in contrast to Antony's more un-Roman patron, Dionysus (Bacchus). In addition, Octavian had started to prefix his name with the designation "Imperator," to suggest that he was the commander *par excellence;* and now, although he continued to use his triumviral powers, he omitted all reference to them from his coins, gradually concentrating on the plain, emotive name "Caesar Son of a God."

But if Octavian was to compete with Antony's military

seniority, successes in a foreign war were necessary; and so Octavian between 35 and 33 BC fought three successive campaigns in Illyricum and Dalmatia (parts of the modern Yugoslavia) in order to protect the northeastern approaches of Italy. With the help of Agrippa, he also lavished large sums on the adornment of Rome. When Octavian fomented public clamour against Antony's territorial gifts to Cleopatra, it was clear that a clash between the two men was imminent.

In 32 BC the triumvirate had officially ended, and Octavian, unlike Antony, professed no longer to be employing its powers. Amid a virulent exchange of propaganda, Antony divorced Octavia, whereupon her brother Octavian seized Antony's will and claimed to find in it damaging proofs of Cleopatra's power over him. Each leader induced the populations under his control to swear formal oaths of allegiance to his own cause. Then, in spite of grave discontent aroused by his exactions in Italy, Octavian declared war—not against Antony but against Cleopatra.

Defeat of Antony and Cleopatra

Accompanied by her, Antony had brought up his fleet and army to guard strongpoints along the coast of western Greece; but in 31 BC Octavian dispatched Agrippa very early in the year to capture Methone, at the country's southwestern tip. His enemies were taken by surprise; and after Octavian himself arrived—leaving his Etruscan friend and adviser Maecenas in charge of Italy—he and Agrippa soon shut Antony's fleet inside the Gulf of Ambracia (Arta). At the Battle of Actium Antony tried to extricate his ships in the hope of continuing the fight elsewhere. Though Cleopatra and then Antony succeeded in getting away, only a quarter of their fleet was able to follow them. She and Antony fled to Egypt and committed suicide when Octavian captured the country in the following year. Executing Cleopatra's son Ptolemy XV Caesar (Caesarion)—whose father she had claimed was Caesar—he annexed Egypt and retained it under his direct control.

The seizure of Cleopatra's treasure enabled him to pay off his veterans and made him finally master of the entire Greco-Roman world. From this point on, by a long and gradual series of tentative, patient measures, he established the Roman principate, a system of government that enabled him to maintain, in all essentials, absolute control. Gradually reducing his 60 legions to 28, he retained approximately 150,000 legionaries, mostly Italian, and supplemented them by about the same number of auxiliaries drawn from the provinces. A permanent bodyguard (the Praetorians), based on the bodyguards maintained by earlier generals, was stationed partly in Rome and partly in other Italian towns. A superb network of roads was created to maintain internal order and facilitate trade; and an efficient fleet was organized to police the Mediterranean. In 28 BC Octavian and Agrippa held a census of the civil population, the first of three during the reign. They also reduced the Senate from about 1,000 to 800 (later 600) compliant members; and Octavian was appointed its president.

Master of the Greco-Roman world

Government and administration. Remembering, however, that Caesar had been assassinated because of his resort to naked power, Octavian realized that the governing class would welcome him as the terminator of civil war only if he concealed his autocracy beneath provisions avowedly harking back to republican traditions. From 31 until 23 BC the constitutional basis of his power remained a continuous succession of consulships, but in January 27 BC he ostensibly "transferred the State to the free disposal of the Senate and people," earning the misleading, though outwardly plausible, tribute that he had restored the republic. At the same time he was granted a ten-year tenure of an area of government (*provincia*) comprising Spain, Gaul, and Syria, the three regions containing the bulk of the army. The remaining provinces were to be governed by proconsuls appointed by the Senate in the old republican fashion. Octavian, however, believed that his supreme prestige—crystallized in the meaningful term *auctoritas*—safeguarded him against any defiance by these personages; and he was indeed able, more or less indirectly, to influence their appointments, just as he was able (on the rare occasions when he regarded it as desirable) to influence the appointments to the consulships and other metropolitan offices that continued to exist in "republican" fashion.

Autocracy in republican guise

Four days after these measures, his name Caesar, acquired through adoption in Julius' will, was supplemented by "Augustus," an appellation with an antique religious ring, believed to be linked etymologically with *auctoritas* and with the ancient practice of augury. The word *augustus* was often contrasted with *humanus;* its adoption as the title representing the new order cleverly indicated, in an extraconstitutional fashion, his superiority over the rest of mankind. With the aid of writers such as Virgil, Livy, and Horace, all of whom in their different ways shared the same ideas, he showed his patriotic veneration of the old Italian faith by reviving many of its ceremonials and repairing numerous temples.

Military operations continued in many frontier areas. In 25, recalcitrant Alpine tribes were reduced, and Galatia (central Asia Minor) was annexed. Mauretania, on the other hand, was transferred from Roman provincial status to that of a client-kingdom, for such dependent monarchies, as in the later republic, bore a considerable part of the burden of imperial defence. Augustus himself visited Gaul and directed part of a campaign in Spain until his health gave out; in 23 he fell ill again and seemed on the point of death. Feeling, amid reports of conspiracies, that new constitutional steps were necessary, he proceeded to terminate his series of consulships in favour of a power (*imperium majus*) which was separated altogether from office and its practical inconveniences. This power raised him above the proconsuls; it was never referred to on the official coinage or in Augustus' political testament but was intended to be exercised mainly in emergencies and on personal visits. He was also awarded the power of a tribune (*tribunicia potestas*) for life. Earlier, he had accepted certain privileges of a tribune. The full power he now assumed carried with it practical advantages, notably the right to convene the Senate. But, more particularly, the office of a tribune, because of the ancient character of the annually elected tribunes of the people as defenders of the *plebs*, surrounded him with a "democratic" aura, one which, perhaps, was needed all the more because Augustus himself —while admittedly supporting the interests of poorer people by a great extension of the right of judicial appeal —tended to back the established classes as the keystone of his system.

Agrippa, too, was granted superiority over proconsuls, presumably in order to ensure that the armies would be in safe hands in case one of Augustus' recurrent illnesses proved fatal. The next to die, however, was the emperor's young nephew Marcellus, who had been married to his daughter Julia and might eventually have been envisaged as his successor. In the same year, 23 BC, Agrippa was sent out to the east as deputy *princeps;* two years later he became Julia's second husband. Meanwhile Augustus himself travelled in Sicily, Greece, and Asia (22–19). Important reorganizations were put into effect wherever he went; and immense satisfaction was caused by an agreement in 20 BC with Parthia, under which the Parthians recognized Rome's protectorate over Armenia and returned the legionary standards captured from Crassus 33 years earlier. In 19 Agrippa completed the subjugation of Spain. In this year there was some adjustment of Octavian's powers to allow him to exercise them more freely in Italy, and the two following years witnessed social legislation attempting to encourage marriage, regulate penalties for adultery, and reduce extravagance. In 17 there were resplendent celebrations of ancient ritual, known as the Secular Games, to purify the Roman people of their past sins and provide full religious inauguration of the new age.

Although the principate was not an office which could be automatically handed on, Augustus seemed to be indicating his views regarding his ultimate successor when he adopted the two sons of his daughter Julia, boys aged three and one who were henceforward known as Gaius Caesar and Lucius Caesar. Their father Agrippa, whose

powers had been renewed along with his master's, returned to the east. But now Augustus also gave important employment to his stepsons—his wife Livia's sons by her former marriage—Tiberius and Drusus the elder. Proceeding across the Alps, they annexed Noricum and Raetia, comprising large parts of what are now Switzerland, Austria, and Bavaria, and extended the imperial frontier from Italy to the upper Danube (16–15).

Reform of the administration

It was probably during these years that an executive, or drafting, committee (*consilium*) of the Senate was established in order to help Augustus to prepare senatorial business. His administrative burden was also lightened by the expansion of his own staff (knights, who could also now rise to a number of key posts, and freedmen) to form the beginnings of a civil service, which had never existed before but was destined to become an essential feature of the imperial system. Gradually, too, a completely reformed administrative structure of Rome, Italy, and the whole empire was evolved. The financial system that made this possible was evidently far more effective than anything the empire had ever seen until then. The system was based on the central treasury (*aerarium*), but the details of its relationship with the treasuries of the provinces, and particularly the *provincia* of Augustus, are still imperfectly understood, partly because, although the emperor proudly recorded his gifts to the central treasury, he did not report what funds passed in the opposite direction.

The taxation providing these resources apparently included two main direct taxes: a poll tax (*tributum capitis*), paid in some provinces by all adults and in others by adult males only, and a land tax (*tributum soli*). There were also indirect taxes, which (as in the past) were farmed out to contractors because their yield was unpredictable and the embryonic civil service lacked the resources to handle them. The republican customs dues continued; but its rates were low enough not to hamper trade, which, in the peaceful conditions created by Augustus, flourished in wholly unprecedented fashion. Industries did not exist on a very large scale, but commerce was greatly stimulated by a sweeping reform and expansion of the Roman coinage. Gold and silver pieces, their designs reflecting many facets of imperial publicity, were issued in great quantities at a number of widely distributed mints; and from about 19 (or perhaps 23) BC onward the absence of bronze token coinage, which had been sparse for many decades, was remedied by the creation of abundant mintages in two bright new metals, yellow brass and red copper. In the West, the principal mint for these pieces, besides Rome, was Lugdunum (Lyon), whose coins displayed a view of the Altar of Rome and Augustus that formed a model for other provincial capitals. The Roman citizen colonies of the West, many of them established by Augustus to settle his veterans, supplemented this output by their own local coinages, and in the East, particularly Asia Minor and Syria, numerous Greek cities were also allowed to issue small change.

Expansion of the empire. The death in 12 BC of Lepidus, who had lived on in retirement for 24 years, enabled Augustus finally to succeed him as the official head of the Roman religion, the chief priest (*pontifex maximus*). In the same year, however, another death came as a severe blow to him, for Agrippa, too, died. Augustus compelled his widow Julia to marry Tiberius against the wishes of both of them. During the next three years, however, Tiberius was away in the field, reducing Pannonia (Yugoslavia and Hungary) up to the middle Danube, while his brother Drusus crossed the Rhine frontier and invaded Germany as far as the Elbe, where he died in 9 BC. In the following year Augustus lost another of his intimates, Maecenas, who had been the adviser of his early days and was an outstanding patron of letters.

Tiberius, who replaced Drusus in Germany, was elevated in 6 BC to a share in his stepfather's tribunician power. But shortly afterward he went into retirement on the island of Rhodes. This was attributed to jealousy of his stepnephew Gaius Caesar, who was introduced to public life with a great fanfare in the following year; and the same compliments were paid to his brother Lucius in 2 BC, the year in which Augustus received his climactic title "father of the country" (*pater patriae*). Gaius was sent to the East and Lucius to the West. Both, however, soon died, Lucius in AD 2 and Gaius in 4. Tiberius returned home in 2, and in 4 Augustus realized that he had to make him his heir. He adopted Tiberius as his son, who in turn was required to adopt Germanicus, the son of his brother Drusus. The powers conferred upon Tiberius made him almost Augustus' own equal in everything except prestige.

Tiberius named heir

Tiberius' next task was to consolidate the invasion and provincial organization of Germany (4–5); and now that the Elbe was the frontier instead of the Danube, Augustus instructed him to establish a shorter frontier line incorporating Bohemia, which had become the nucleus of a German (Marcomannic) empire ruled by King Maroboduus. An invasion of Bohemia, therefore, was planned, and had already been launched, from two directions, when news came in 6 that Pannonia and Illyricum had revolted. It took three years for the rebellion to be put down; and this had only just been completed when Arminius raised the Germans against their Roman governor Varus and destroyed him and his three legions. As Augustus could not readily replace the troops, the annexation of western Germany and Bohemia was postponed indefinitely; Tiberius and Germanicus were sent to consolidate the Rhine frontier.

Although Augustus was now feeling his age, these years in association with Tiberius were marked by administrative innovations: the annexation of Judaea in AD 6 (its client king Herod the Great had died ten years previously); the establishment at Rome (in the same year) of a fire brigade with police duties—supplemented seven years later by a regular police force (*cohortes urbanae*); the creation of a military treasury (*aerarium militare*) to defray soldiers' retirement bounties from taxes; and the conversion of the hitherto occasional appointment of prefect of the city (*praefectus urbi*) into a permanent office (AD 13). When, in the same year, the powers of Augustus were renewed for ten years—such renewals had been granted at intervals throughout the reign—Tiberius was made his equal in every constitutional respect. In April, Augustus deposited his will at the House of the Vestals in Rome. It included a summary of the military and financial resources of the empire (*breviarium totius imperii*) and his ingenious political testament known as the "Res Gestae Divi Augusti" ("Acts of the Divine Augustus"). The best preserved copy of the latter document is still to be seen on the walls of the Temple of Rome and Augustus at Ankara, Turkey. In 14 Tiberius was due to leave for Illyricum but was recalled by the news that Augustus was gravely ill. He died on August 19, and on September 17 the Senate enrolled him among the gods of the Roman state. By that time Tiberius had succeeded him as the second Roman emperor, though the formalities involved in the succession proved embarrassing both to himself and to the Senate because the "principate" of Augustus had not, constitutionally speaking, been heritable or continuous. Like other emperors, Tiberius assumed the designation "Augustus" as an additional title of his own. Agrippa Postumus, who had been named his co-heir but was later banished, was put to death. The order to kill him may already have been given by Augustus, but this is not certain.

Personality and achievement. Augustus was one of the great administrative geniuses of history. The gigantic work of reorganization that he carried out in every field of Roman life and throughout the entire empire not only transformed the decaying republic into a new, monarchic regime with many centuries of life ahead of it but created a durable Roman peace, based on easy communications and flourishing trade. It was this Pax Romana that ensured the survival and eventual transmission of the classical heritage, Greek and Roman alike, and provided the means for the diffusion of Judaism and Christianity (Jesus Christ was born during Augustus' reign). Although his regime was an autocracy, Augustus, being a tactful and imaginative master of propaganda of many kinds, knew how to cloak that autocracy in traditionalist

forms that would satisfy a warworn generation—perhaps, most of all, the upper bourgeoisie immediately below the leading nobility, since it was they who benefitted from the new order more than anyone. He was also able to win the approbation, through the patronage of Maecenas, of some of the greatest writers the world has ever known, including Virgil, Horace, and Livy.

Their enthusiasm was partly due to Augustus' conviction that the Roman peace must be under occidental, Italian control. This was in contrast to the views of Antony and Cleopatra, who had envisaged some sort of Greco-Roman partnership such as only began to prevail three or four centuries later. Augustus' narrower view, although modified by an informed admiration of Greek civilization, was based on his small-town Italian origins. These were also partly responsible for his patriotic, antiquarian attachment to the ancient religion and for his puritanical social policy.

Augustus was a cultured man, the author of a number of works (all lost): a pamphlet against Brutus, an exhortation to philosophy, an account of his own early life, a biography of Drusus, poems, and epigrams. The conventional view of his character distinguishes between his cruelty in early years and his mildness in later life. But there was not so much need for cruelty later on, and when it was needed (notably in the suppression of alleged plots), he was still ready to apply it. It is probable that nothing short of this degree of political ruthlessness could have achieved such enormous results. His domestic life, however, was simple and homespun. Within his family, the successive deaths of those he had earmarked as his successors or helpers caused him much sadness and disappointment. His devotion to his wife Livia Drusilla remained constant, though, like other Romans, he was unfaithful. His surviving letters show kindliness to his relations. Yet he exiled his daughter Julia for offending against his public moral attitudes, and he exiled her daughter by Agrippa for the same reason; he also exiled the son of Agrippa and Julia, Agrippa Postumus, though the suspicion that he later had him killed is unproved. As for Augustus' male relatives who were his helpers, he was loyal to them but drove them as hard as he drove himself. He needed them because the burden was so heavy, and he especially needed them in the military sphere because he was not a great commander. In Agrippa and Tiberius and a number of others he had men who supplied this deficiency, and although, on his deathbed, he is said to have advised against the further expansion of the empire, he himself, with their assistance, had expanded its frontiers in many directions.

His physical condition was subject to a host of ills and weaknesses, many of them recurrent. Indeed, in his early life, particularly, it was only his indomitable will that enabled him to survive—a strange preliminary to an unprecedented and unequalled life's work. His appearance is described by the biographer Suetonius.

> He was unusually handsome and exceedingly graceful at all periods of his life, though he cared nothing for personal adornment. His expression, whether in conversation or when he was silent, was calm and mild . . . He had clear, bright eyes, in which he liked to have it thought that there was a kind of divine power, and it greatly pleased him, whenever he looked keenly at anyone, if he let his face fall as if before the radiance of the sun. His teeth were wide apart, small and ill-kept; his hair was slightly curly and inclining to golden; his eybrows met . . . His complexion was between dark and fair. He was short of stature, but this was concealed by the fine proportion and symmetry of his figure, and was noticeable only by comparison with some taller person standing beside him.

Augustus' countenance proved a godsend to the Greeks and Hellenized easterners who were the best sculptors of the time, for they elevated his features into a moving, never to be forgotten imperial type, which Napoleon's artists, among others, keenly emulated. The contemporary portrait busts of Augustus, echoed on his coins, formed part of a significant renaissance of the arts in which Italic and Hellenic styles were discreetly and brilliantly blended. Still extant at Rome are the severe yet delicate reliefs of the Ara Pacis ("Altar of Peace"), depicting a religious procession in which the national leaders are taking part; there are also scenes from the Roman mythology. The altar was dedicated by the Senate and people of Rome in 13 BC to commemorate the pacification of Gaul and Spain.

Cultural heritage

The architectural masterpieces of the time were also numerous; and something of their monumental grandeur and classical purity can be seen today in the remains of the Theatre of Marcellus at Rome and of the massive Forum of Augustus, flanked by colonnades and culminating in the Temple of Mars the Avenger—the Avenger of Julius Caesar. Outside Rome, too, there are abundant memorials of the Augustan age; on either side of the Alps, for example, there are monuments to celebrate the submission and loyalty of the local tribes, an elegant arch at Segusio (Susa) and a square stone trophy, topped by a cylindrical drum, at La Turbie. From Livia's mansion on the outskirts of Rome, at Prima Porta, comes a reminder that not all the art of the day was formal and grand. For one of the rooms is adorned with wall paintings representing an enchanted garden; beyond a trellis are orchards and flower beds, in which birds and insects perch among the foliage. Augustus himself had no interest in personal luxury. Yet if ever he or his associates had any spare time, such were the rooms in which they spent it.

BIBLIOGRAPHY. The principal ancient literary sources are SUETONIUS, *Life of Augustus;* and DIO CASSIUS, books 52–56 (both translated in Loeb editions). *Inscriptions: Res Gestae Divi Augusti* (*The Achievements of the Divine Augustus*), ed. by P.A. BRUNT and J.M. MOORE (1967); V. EHRENBERG and A.H.M. JONES (comps.), *Documents Illustrating the Reigns of Augustus and Tiberius*, 2nd ed. (1955). *Coins:* C.H.V. SUTHERLAND, *Coinage in Roman Imperial Policy, 31 B.C.–A.D. 68* (1951, reprinted 1971); M. GRANT, *From Imperium to Auctoritas* (1946, reprinted 1969). *Art:* J.M.C. TOYNBEE, *The Art of the Romans* (1965); A. BOETHIUS and J.B. WARD-PERKINS, *Etruscan and Roman Architecture* (1970).

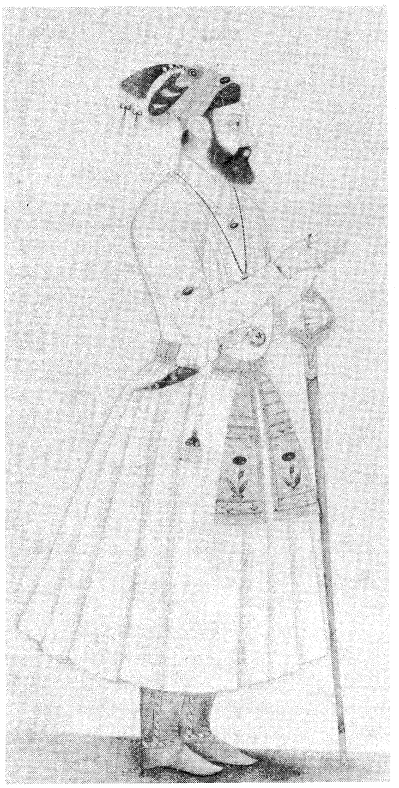
Modern sources: R. SYME, *The Roman Revolution*, new ed. (1960), a basically correct new point of view; A.H.M. JONES, *Augustus* (1971), a concise summing-up; M. HAMMOND, *The Augustan Principate*, new ed. (1968), sources on the constitution; *Cambridge Ancient History*, vol. 10, *The Augustan Empire 44 B.C.–A.D. 70* (1934, reprinted 1966), the fullest general survey; D.C. EARL, *The Age of Augustus* (1968), with many illustrations; J.M. CARTER, *The Battle of Actium: The Rise and Triumph of Augustus Caesar* (1970); H.H. SCULLARD, *From the Gracchi to Nero*, 3rd ed. (1970), references brought up to date; W. SCHMITTHENNER (ed.), *Augustus* (1969), collected articles (in German).

(M.Gr.)

Aurangzeb

Aurangzeb, the last of the great Mughal emperors of India, was a zealous Muslim during whose reign (1659–1707) the empire reached its vastest extent, though at the cost of serious depletion of economic resources and administrative stability. His kingly title was ʿĀlamgīr ("World Holder"), but in Europe he became best known by his princely title, Aurangzeb ("Ornament of the Throne"), because of the war of succession that brought him to the throne. Aurangzeb is important in Indian history because his policies toward the Hindus, the Deccan kingdoms, and the Marāthās of western India gave rise to tensions that were to lead to the dissolution of the empire in the 18th century. His conception of his role as that of an orthodox ruler of an Islāmic state rather than as that of the Muslim ruler of an Indian empire was vital to these issues.

Aurangzeb was born on October 24, 1618 (old style), the third son of the emperor Shāh Jahān and Mumtāz Maḥal (for whom the Tāj Mahal was built). Given the name Muḥī-ud-Dīn Muḥammad, he grew up as a serious-minded and devout youth, wedded to the Muslim orthodoxy of the day and free from the royal Mughal traits of sensuality and drunkenness. He early showed signs of military and administrative ability; and these qualities, combined with a taste for power, brought him into rivalry with his eldest brother, the brilliant and volatile Dārā Shukōh, who was designated by their father as his successor to the throne. From 1636 he held a number



Aurangzeb, Mughal miniature, 17th century. In the Metropolitan Museum of Art, New York.

By courtesy of the Metropolitan Museum of Art, New York, bequest of George D. Pratt, 1945

of important appointments, in all of which he distinguished himself. He commanded troops against the Uzbeks and the Persians with distinction (1646–47) and, as viceroy of the Deccan provinces in two terms (1636–44, 1654–58), reduced the two Muslim Deccan kingdoms to near-subjection.

When Shāh Jahān fell seriously ill in 1657, the tension between the two brothers made a war of succession seem inevitable. By the time of Shāh Jahān's unexpected recovery, matters had gone too far for either son to retreat. In the struggle for power (1657–59), Aurangzeb showed tactical and strategic military skill, great powers of dissimulation, and ruthless determination. Decisively defeating Dārā at Samugarh in May 1658, he confined his father in his own palace at Agra. In consolidating his power he caused one brother's death and had two other brothers, a son, and a nephew executed. The war became a legend and found its way to Europe.

Aurangzeb's reign falls into two almost equal parts. In the first, which lasted until about 1680, he was a capable Muslim monarch of a mixed Hindu–Muslim empire and as such was generally disliked for his ruthlessness but feared and respected for his vigour and skill. During this period he was much occupied with safeguarding the northwest from Persians and Central Asian Turks and less so with the Marāthā chief Śivajī, who twice plundered the great port of Surat (1664, 1670). He applied his great-grandfather Akbar's recipe for conquest: defeat one's enemies, reconcile them, and place them in imperial service. Thus Śivajī was defeated, called to Āgra for reconciliation (1666), and given an imperial rank. The plan broke down, however; then Śivajī fled to the Deccan and died, in 1680, as the ruler of an independent Marāthā kingdom.

Second phase of his rule

After about 1680, Aurangzeb's reign underwent a change of both attitude and policy. The pious ruler of an Islāmic state replaced the seasoned statesman of a mixed kingdom; Hindus became subordinates, not colleagues, and the Marāthās, like the southern Muslim kingdoms, were marked for annexation rather than containment. The overt sign of the first change was the reimposition of the *jizya*, or poll tax, on non-Muslims in 1679 (a tax that had been abolished by Akbar). This in turn was followed by a Rajput revolt in 1680–81, supported by Aurangzeb's third son, Akbar. Hindus still served the empire, but no longer with enthusiasm. The Deccan kingdoms of Bijāpur and Golconda were conquered in 1686–87, but the insecurity that followed precipitated a long and incipient economic crisis, which in turn was deepened by the Marāthā war. Śivajī's son Sambhājī was captured and executed in 1689 and his kingdom broken up. The Marāthās, however, then adopted guerrilla tactics, spreading all over South India amid a sympathetic population. The rest of Aurangzeb's life was spent in laborious and fruitless sieges of forts in the Marāthā hill country.

Aurangzeb's absence in the south prevented him from maintaining his former firm hold on the north. The administration weakened, and the process was hastened by pressure on the land by Mughal grantees who were paid by assignments on the land revenue. Agrarian discontent often took the form of religious movements, as in the case of the Satnamis and the Sikhs in the Punjab. In 1675 Aurangzeb arrested and executed the Sikh Gurū (spiritual leader) Tegh Bahādur, who had refused to embrace Islām; and the succeeding Gurū was in open rebellion for the rest of the reign. This was the real beginning of the still-existing Sikh–Muslim feud. Other agrarian revolts, such as those of the Jāṭs, were largely secular.

In general, Aurangzeb ruled as a militant orthodox Sunnī Muslim, who put through increasingly puritanical ordinances that were vigorously enforced by *muḥtasib*s, or censors of morals. The Muslim confession of faith, for instance, was removed from all coins lest it be defiled by unbelievers; courtiers were forbidden to salute in the Hindu fashion. In addition, Hindu idols, temples, and shrines were often destroyed.

Aurangzeb maintained the empire for nearly half a century and in fact extended it in the south as far as Tanjore and Trichinopoly. Behind this imposing facade, however, were serious weaknesses. The Marāthā campaign continually drained the imperial resources. The militancy of the Sikhs and the Jāṭs boded ill for the empire in the north. The new Islāmic policy alienated Hindu sentiment and undermined Rājput support. The financial pressure on the land strained the whole administrative framework. When Aurangzeb died on February 20, 1707 (O.S.), after a reign of nearly 49 years, he left an empire not moribund but confronted with a number of menacing problems. The failure of his son's successors to cope with them led to the collapse of the empire in the mid-18th century.

BIBLIOGRAPHY. J.N. SARKAR, *History of Aurangzib*, 5 vol. (1924–30), the standard work by a leading Indian historian; *Anecdotes of Aurangzib*, 3rd ed. (1949), contains translated excerpts about Aurangzeb from contemporary works in Persian; S.M. EDWARDES and H.L.O. GARRETT, *Mughal Rule in India* (1930), a reliable background book of 17th-century India; K.R. QANUNGO, *Dara Shekoh*, 2nd ed. (1952), a well-documented study of Aurangzeb's chief rival for the throne.

(T.G.P.S.)

Auroras

The aurora, a familiar and often beautiful display of light in the upper air of both the Arctic and Antarctic regions, is caused by light emitted from the upper atmosphere in a form resembling the light of an electrical discharge. It was once erroneously thought to be sunlight reflected from polar snow and ice, or refracted light, like that of the rainbow. The name aurora is derived from the Latin name for the mythological personification of dawn. In the Northern Hemisphere it is called the Aurora Borealis, or the northern lights, and in the Southern Hemisphere, the Aurora Australis, or the southern lights; together they are called the Aurora Polaris or Polar Aurora.

The aurora commonly appears as a distinct, greenish, curtain-shaped light. Another type of aurora appears as an extensive dark-red glow in the poleward sky. The light is caused by the glow of atoms in the thin upper atmosphere as they are hit by fast-moving electrons and

protons, here called auroral electrons and protons, respectively.

It was once thought that auroral particles were thrown into interplanetary space from sunspots and solar flares, and that some of them reached the polar upper atmosphere after being deflected by the Earth's geomagnetic field. Although there is no doubt that the ultimate source of the energy of auroral particles originates in the Sun, it is now recognized that the origin and acceleration processes of auroral particles are very complex. The problem is closely related to the electromagnetic environment in space around the Earth.

The Earth's magnetic field is confined to a cylindrical cavity with a blunt nose, formed by the solar wind, a hot gas streaming Earthward from the sun. This cavity, called the magnetosphere, contains the Van Allen radiation belts. The magnetosphere acts like a gigantic cathode ray or television tube, generating and acting upon electron beams that in effect produce the image (the aurora) on the gigantic screen that is the polar upper atmosphere. Because the source of energy of the aurora comes from the Sun, auroral activity is closely associated with solar activity. In particular, a stormy solar wind that is generated by solar flares activates this gigantic cathode-ray tube and causes intense auroral display.

During the last ten years, a large number of artificial satellites have been made to traverse the magnetosphere to survey and map the distribution of auroral particles and the magnetic field, and rockets have been shot into the polar upper atmosphere to gain data about the characteristics of the auroral particle precipitation. An extensive international network of auroral and magnetic observatories has operated in the polar wilderness during and after the 1957–58 International Geophysical Year. These efforts have greatly enhanced comprehension of auroral phenomena in terms of magnetospheric processes.

This article deals with the cause, occurrence, and general distribution of the auroras. For further information on the charged particles in the Earth's atmosphere and their generation, see IONOSPHERE; VAN ALLEN RADIATION BELTS; see also EARTH, MAGNETIC FIELD OF for relevant information on the precise location, intensity, and effect of the Earth's magnetic field.

THE OCCURRENCE AND DISTRIBUTION OF THE AURORA

The two auroral ovals

The aurora is a common feature of the planet Earth and appears along two oval belts called the auroral ovals. There is one oval in each hemisphere, surrounding the geomagnetic poles; these are not coincident with the Earth's geographic poles. The northern geomagnetic pole is located near the northwestern tip of Greenland and the southern pole near Vostok, a Soviet station in the Antarctic. Each oval has an approximate radius of 2,000 kilometres (1,200 miles). It is eccentric with respect to the geomagnetic pole; that is, its centre is shifted towards the dark hemisphere a few degrees from the geomagnetic pole. Thus, the midnight portion is at about geomagnetic latitude 67° and the midday portion at about geomagnetic latitude 76°.

Each oval is more or less fixed in space with respect to the sun, and the Earth rotates beneath it once a day. The locus of the midnight portion of the Earth is called the auroral zone. It lies along the curve connecting the northern tip of Scandinavia, Iceland, the southern tip of Greenland, the southern part of Hudson Bay, Central Alaska, and the Siberian coast. Figure 1 shows the location of the oval at 10 o'clock Universal Time (UT) or Greenwich Mean Time (GMT). An observer in the auroral zone will come directly under the oval once a day during local midnight hours; daily occurrence of the aurora for such an observer will therefore have a single peak around midnight. On the other hand, an observer located between geomagnetic latitude 67° and 74° comes under the oval twice a day, once in the evening hours and once in the morning hours, and the daily occurrence of the aurora in this case will have a double peak. An observer located at geomagnetic latitude 76° will see the aurora most frequently at noon, if his location is in darkness. In the Northern Hemisphere, however, the midday aurora can be observed visually only over the Arctic Ocean around the December solstice (the shortest day of the year in the Northern Hemisphere); it can be observed for a few months at the South Pole (Amundsen-Scott Base) during southern winter months.

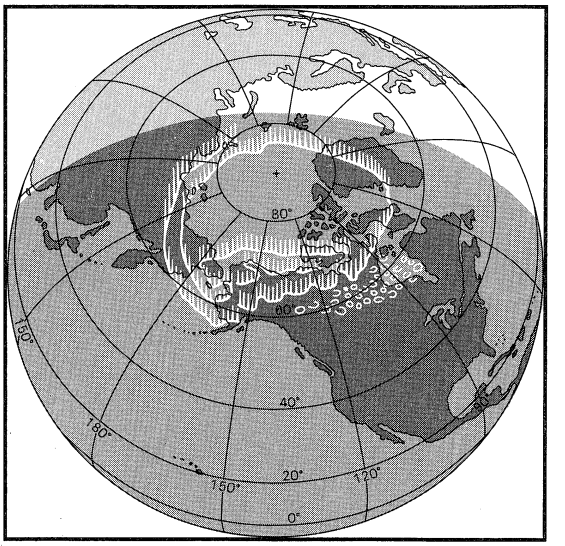


Figure 1: Distribution of the auroras in the Northern Hemisphere at 10:00 AM Greenwich Mean Time.

Changes in size of the ovals

The radius of the auroral oval varies with the intensity of solar activity. When there is little solar activity, the oval shrinks towards the geomagnetic pole; the midnight portion shifts polewards to about geomagnetic latitude 71°. The oval in Figure 1 approximately represents the size during moderate solar activity.

Conversely, during periods of great solar activity, about two days after an intense solar flare, the radius of the auroral oval increases considerably. On the evening of February 10, 1958, for example, the oval descended down as far as geographic latitude 40° in the United States (see Figure 1). The period of such an unusual expansion of the oval coincides with the period of an intense geomagnetic storm indicating the arrival of solar gas (consisting mainly of protons and electrons), which is ejected from the Sun during solar flares. The time lapse of two days after the flare is a result of the travel time of the solar gas from the Sun to the Earth at its average speed of about 880 kilometres per second.

In the polar cap region encircled by the auroral oval there sometimes appear very faint arcs called polar cap auroras, a unique feature of which is a tendency to align across the polar cap from the dayside to the nightside, approximately along the Sun–Earth line. These arcs differ spectroscopically from the auroral oval proper.

FORMS OF THE AURORA

One of the fundamental forms of the aurora has the appearance of a curtain and is greenish-white in colour. The lower border, generally fairly sharply defined, is located at a height of about 100 kilometres (60 miles). The upper border is much less clearly defined and extends to heights of several hundred kilometres; it may be as high as 1,000 kilometres for the aurora that appears in the twilight sky.

The curtain-like form lies along the auroral oval, extending roughly in the east–west direction with a length of more than several thousand kilometres, whereas its north–south extent (or the thickness of the curtain) is only about a few hundred metres. When this form is seen at a certain distance, particularly in the poleward sky, it appears as an arch rising from the horizon and is called an arc. If the brightness is nearly the same along its horizontal extent, it is called a homogeneous arc and appears to be motionless or relatively quiet. Multiple homogeneous arcs are often seen, and it is not uncommon for as

Arcs, bands, and patches

many as ten to stretch across the sky in the auroral zone (Figure 2). When the aurora becomes active, the arc develops five pleats that appear as vertical striations called rays; an arc with rays is called a rayed arc. In a more active form, called the band, the arc becomes wavy or folded. An active band with ray structure is called a rayed band, and a rayed band with a well-developed fold is often referred to as a drapery or horseshoe-type aurora. An active rayed band often moves with a speed of about 100 metres (300 feet) per second; a crimson-red colour often is added near the lower border of the greenish-white curtain. When a rayed band is seen from relatively nearby, the rays appear to converge towards a small region, resulting in an apparent fan-shaped form called a corona. Not a distinct form of the aurora, the corona results from the perspective effect of a great length of rays.

V.P. Hessler

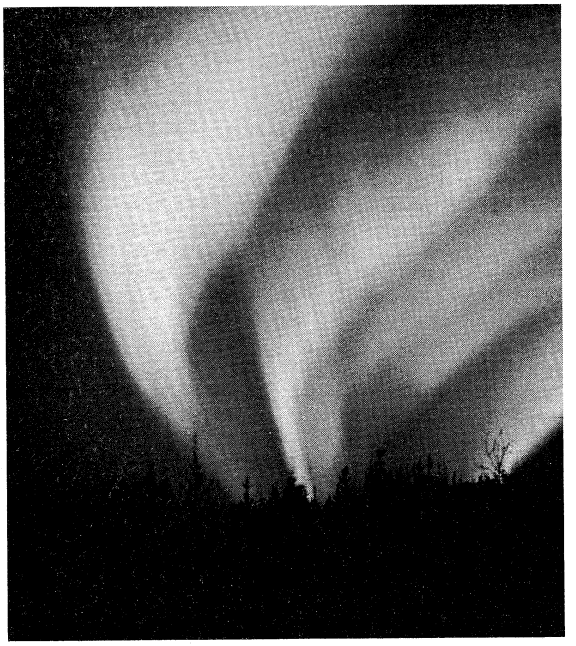


Figure 2: *Aurora borealis.*
Multiple auroral arcs photographed in Alaska.

After an intense auroral display, rayed bands often appear to break up into a number of isolated rays scattered over the whole sky. When the rays are not clearly seen, this broken-up form looks like a group of cumulus clouds; it is most commonly seen after midnight and is referred to as auroral patches. The brightness of most changes irregularly in a phenomenon called pulsation.

Auroral arcs, bands, and patches are quite common and occur practically daily in the auroral zone. Their frequency of occurrence decreases rapidly toward the Equator, and in the northern United States these forms may be seen an average of only five times a year, indicating that expansion of the oval is a rare event.

Another important form of the aurora is the veil, an extensive glow over a large region of the sky. The veil is commonly dark red and is seen well away from the oval, toward the Equator, during great geomagnetic storms. Indeed, many historic auroras seen as far away from the polar area as Samoa (May 13, 1921) and Mexico (September 13, 1957; February 11, 1958) were of this type. The frequency of occurrence of the veil in the southern United States is, on the average, once a year.

SPECTRAL CHARACTERISTICS OF THE AURORA

The curtain-shaped aurora is caused primarily by an intense sheet beam of energetic electrons (with energies of about 1,000 to 10,000 electron volts) that spiral down into the polar upper atmosphere. There the electrons collide with such particles as molecular nitrogen and oxygen and atomic oxygen, leading to such processes as ionization, excitation, or dissociation, or all three. Ionized nitrogen molecules emit radiation (emission lines) in several series of wavelengths, called bands. The first negative band extends from 3914 angstroms (one angstrom equals 10^{-7} millimetres) to 4700 angstroms. Some other emissions are in the infrared range. When the ionized nitrogen molecules recombine with electrons, some of the resulting molecules are in excited states and also emit several bands. The first positive band extends from 5800 to 6800 angstroms. There are some bands in the ultraviolet range that have been detected by rocket-borne instruments (see also ATMOSPHERE; IONOSPHERE).

Each ionization process by a high-energy electron produces a low-energy electron, called a secondary electron. The reaction itself is often sufficiently energetic to ionize or excite other atoms and molecules. The most common light of the aurora, the greenish-yellow emission at 5577 angstroms, is initiated in part by the excitation of oxygen atoms by these secondary (low energy) electrons.

Oxygen molecules also are ionized by incoming energetic electrons. The first negative band (5800 to 6800 angstroms) of O_2^+ is partly responsible for the crimson-red colour. When an ionized oxygen molecule recombines with an electron, it usually splits into two oxygen atoms, one of which is in an excited state; some of the excited oxygen atoms thus produced are another source of the aurora's greenish-yellow colour. After emitting radiation at 5577 angstroms (greenish-yellow light), the oxygen atom must emit additional radiation at 6300 angstroms before it returns to the ground (original, unexcited) state. This emission (dark-red colour) can be seen in the upper part of the curtain-like form when it is very intense.

The radiation at 6300 angstroms is mainly responsible for the extensive dark-red glow (the veil) in middle and low latitudes during great magnetic storms. The excitation mechanism for this particular type of aurora is, however, not well understood, and energetic electrons (approximately 10,000 electron volts) are not thought likely to be responsible for it. In mid-latitude regions during a great magnetic storm, there also occurs a subvisual belt of 6300 angstroms emission called the mid-latitude red arc. Its excitation mechanism is not well understood.

Collisions and emission shifts

Protons with energies of 200,000 electron volts also spiral down into the polar upper atmosphere. During descent, they collide with oxygen atoms, which are the major constituent of the atmosphere at heights between 500 and 1,000 kilometres (300 and 600 miles). The collision results in a charge exchange in which the proton becomes an energetic, neutral, hydrogen atom and the oxygen atom becomes ionized. Some of the neutral hydrogen atoms thus produced are in excited states, emitting fundamental radiation series that can be observed on the ground by optical instruments, such as a spectrometer and photometer, with appropriate filters. Being neutral and energetic, the atoms move in a straight line unhindered by the geomagnetic field, so that the observed emission lines are shifted from their normal positions.

The ultraviolet radiation from the neutral, excited hydrogen atoms is absorbed before reaching the ground, so that rocket-borne instruments are needed to detect it. When the hydrogen atom collides with an ionized oxygen atom, it exchanges charge and creates an energetic proton and a neutral oxygen atom. The energetic proton may exchange charge again by colliding with an oxygen atom, repeating this process more than 1,000 times before the proton is brought to rest. Eventually, the proton recombines with an electron to become a neutral atom.

The aurora containing this emitted radiation from the excited, neutral hydrogen atoms is called the proton aurora. It appears as a diffuse belt, a few hundred kilometres in width, that is displaced toward the Equator from the auroral oval. Because hydrogen atoms, in both neutral and ionized states, are energetic enough to ionize nitrogen and oxygen molecules during their descent, many of the emissions contained in curtain-shape auroras are also present in the glow; indeed, the glow seen by unaided eyes is mostly from such emissions, rather than from the hydrogen emissions.

A glow of similar spectral characteristics also appears in the polar cap several hours after an intense flare. It is called the polar-cap glow and is caused by the bombardment of solar cosmic rays, which consist mainly of very energetic protons (more than 1,000,000 electron volts).

THE AURORAL SUBSTORM AND AURORAL STORM

If one looked down on the Earth to observe the behaviour (or activity) of the aurora in the auroral oval, he could see homogeneous arcs, or relatively quiet forms, all along the auroral oval during a quiet period. Such a quiet period is often disrupted suddenly. The first indication of disruption occurs near the boundary between the auroral oval and the proton aurora in the midnight sector. There, a homogeneous arc becomes bright in a matter of minutes and begins to move toward the pole with a speed of several hundred metres per second. This motion produces a large "bulge" in the oval in the midnight sector. Along the boundary of the bulge closest to the pole lies a very bright rayed band with a crimson-red lower border. It is not uncommon to observe the bulge reaching geomagnetic latitude 75°.

Auroral activation

Subsequently, a surging motion of the aurora is generated on the evening side of the bulge. The surge is propagated westward along the oval with a speed of a few kilometres per second, and it may reach as far as the early afternoon sector of the oval. In the morning part of the oval, homogeneous arcs appear to disintegrate into isolated rays or patches. The resulting patches as a group drift eastward with a speed of about several hundred metres per second.

In this way, auroras are activated all along the auroral oval in about 30 to 60 minutes. Figure 1 shows the distribution of such activated auroras. After the boundary of the bulge closest to the pole reaches the highest latitude, it begins to shrink back toward the Equator. Rayed bands in the bulge start to move toward the Equator with a speed of about 50 metres (150 feet) per second; many of them fade away during the motion. Gradually, in one to two hours or so, homogeneous or other relatively quiet forms are restored at about the location where they were seen just before the onset of activity. The surges and drifting patches may still be seen at this time in the evening and morning twilight skies, respectively. Eventually, however, a quiet condition is restored all along the oval. The whole activity typically lasts for about three hours, although the period can be less than one hour.

Sequential activity

This sequence of activities, from a quiet to an active, and then back to a quiet condition, is called the auroral substorm, and it has two characteristic phases: the expansive phase and the recovery phase. The expansive phase is manifested most clearly by the expanding bulge of the auroral oval and the recovery phase by the shrinking bulge. The term breakup, often used in scientific literature, expresses only vaguely the overall auroral activity in the dark side of the oval during the auroral substorm.

As has been pointed out, the radius of the oval increases during intense geomagnetic storms. The increase takes about ten hours, during which as many as ten auroral substorms may occur, so that the oval undergoes a repeated expansion and subsequent contraction of the width, as well as an increase of radius. During this ten-hour period a geomagnetic storm develops, but after it the geomagnetic storm begins to subside, and the occurrence of substorms becomes much less frequent. Thus, it is possible to define the term auroral storm so as to describe overall auroral activity during the period of a geomagnetic storm. Clearly, the auroral storm consists of several to a dozen auroral substorms.

Both the geomagnetic storm and the auroral storm are different manifestations of a single stormy phenomenon in the magnetosphere, called the magnetospheric storm. There are a number of other manifestations of the magnetospheric storm; furthermore, the magnetospheric storm consists of about a dozen magnetospheric substorms. A large amount of energy is liberated in the magnetosphere during a magnetospheric substorm. A significant part streams into the polar region in the form of the kinetic energy of auroral particles.

RELATED PHENOMENA

The auroral substorm is not the only manifestation of the magnetospheric substorm. Greenish-yellow, curtain-shape auroras for the most part are caused by sheetlike electron beams with energy ranges from one to 10,000 electron volts. During magnetospheric storms, more-energetic electrons are also produced in the magnetosphere and precipitate in the vicinity of visible auroras. Electrons with energies of about 20,000 electron volts are responsible for exciting the crimson-red colour near the lower border of the curtain. Because they are more energetic, they can penetrate deeper into the upper atmosphere than the 10,000-electron-volt electrons, and they ionize oxygen molecules, which greatly increase in number below heights of 100 kilometres. Electrons with still greater energy penetrate even deeper.

X-rays and whistlers

Energetic electrons also generate strong X-rays (radiation produced by sudden retardation of the particle) when they collide with upper atmospheric particles. Since 1960 a great number of balloon-borne X-ray detectors have been sent up to observe the X-ray flux at the 30-kilometre (20-mile) level. Studies of the X-rays have been useful in observing fine time variations (as short as a few milliseconds) of the electron fluxes. It also has been found that intense bursts of X-rays occur during the auroral substorm along the morning half of the auroral zone, as well as in the region of active auroras. These energetic electrons are known to interact strongly with very low frequency radio waves called whistlers, which are generated by lightning and are then propagated through the magnetosphere. Some of the electrons are scattered out of the Van Allen radiation belts by such waves and are precipitated into the upper atmosphere. Under certain conditions, a cloud of the energetic electrons can generate very low frequency waves by itself. Indeed, it has been shown that this mechanism is so efficient as to limit the maximum flux of electrons to be trapped in the Van Allen belt.

Such energetic electrons also produce an appreciable amount of ionization in the lower ionosphere where short (radio) waves are most effectively absorbed. Thus, intense auroral activity may seriously disrupt transpolar short-wave communication. This phenomenon has been used to monitor the precipitation of the energetic electrons by a device called a riometer, which continuously monitors the intensity of cosmic radio waves (short-wavelength waves from interstellar sources) with frequencies of about 30 megahertz. When a layer of ionization is produced by the precipitation of the energetic electrons, the ionosphere becomes quite opaque to the radio waves, so that the intensity received by a riometer decreases.

Effect of auroras on communications transmission

On the other hand, such an ionization may be capable of reflecting (or scattering) radio waves of much higher frequencies. An unusual long-distance radio communication or television transmission may occur during great auroral displays. Using such a reflection of very high frequency waves, radars have been extensively used to study the aurora. A strong echo returns from the aurora when the radio beam strikes the auroral curtain perpendicularly.

During the auroral substorm, some auroras move with a speed exceeding that of sound in the upper atmosphere. It recently has been demonstrated that such a supersonic motion of the aurora is associated with infrasonic (below the audibility range of the human ear) shock waves.

Another important characteristic of the magnetospheric substorm is the worldwide extent of magnetic disturbance. Because it is most intense in the polar region, this particular magnetic disturbance is called the polar magnetic substorm. The distribution of the magnitude and orientation of magnetic disturbances over the Earth's surface has been studied by the analyzing of records from a large number of magnetic observatories. The main feature of the distribution is as follows: very intense magnetic disturbances, oriented toward the Equa-

Magnetic disturbances

tor, occur under the midnight part of the auroral oval, whereas less intense disturbances are oriented toward the poles along the auroral zone in the afternoon and evening sectors; in middle and low latitudes the orientation is toward the Equator in the afternoon and evening sectors and toward the poles in the midnight and morning sectors; and eastward and westward in the evening and morning sectors, respectively, in middle latitudes. It has long been known that the intense magnetic disturbances under the auroral oval are produced by a strongly concentrated (about 1,000,000 amperes) electric current flowing westward along the oval at about the level where the auroral curtain is brightest (110 kilometres [70 miles]); this current is called the auroral electrojet. The origin and distribution of the electrojet is not fully understood. Satellite observations are of great help in understanding the cause of magnetospheric substorms. A number of studies have yielded knowledge of the existence of drastic changes of the distribution of electrons and the magnetic field in the Van Allen belt and in the "tail" region of the magnetosphere during the substorm. It also has been reported that an extensive "sheet" of low energy, ionized gases, called plasma sheet, in the tail region of the magnetosphere undergoes a considerable change. Indirect observations show that a large-scale convective motion of the plasma in the sheet is generated during the substorm and that it plays a fundamental role in auroral processes. Theoretical and observational evidence shows that the direction of interplanetary magnetic fields is related to the occurrence of the substorm.

Magnetospheric processes

The magnetospheric substorm is also associated with the growth of an intense Van Allen radiation belt that consists mainly of protons with energies of about 20,000 electron volts and contains an electric current of about 1,000,000 amperes; they are partially responsible for the magnetic disturbance oriented toward the Equator in middle and low latitudes in the afternoon and evening sectors. The magnitude is much less than that produced by the auroral electrojet, simply because the belt is much farther from the Earth's surface than the electrojet. The radiation belt does not encircle the Earth completely.

When intense magnetospheric substorms occur very frequently, the radiation belt grows abnormally intense because a large number of protons are accumulated in the radiation belt. Such a belt, known as the equatorial ring current, causes a large decrease in the strength of the Earth's magnetic field in low latitudes, which is one of the major features of the geomagnetic storm field. It has been suggested that the proton aurora is caused by protons that leak out from the ring current and impinge into the polar upper atmosphere, just toward the Equator from the auroral oval.

Analogy with television

The magnetosphere might be considered to be a gigantic television tube (Figure 3). The polar upper atmosphere would correspond to the screen of a television tube that has a diameter of about 4,000 kilometres (2,500 miles) instead of 16 or 18 inches (41 or 46 centimetres), so that a person would be able to watch only about 1/50 of the screen at a time. The aurora would then correspond to an image on the screen in the television tube. A television screen is coated by fluorescent material that emits light when it is hit by an electron beam. When an electron beam hits the polar upper atmosphere, one can observe auroras. In a television tube, the electron beam is acted upon (modulated) by a pair of electric plates and an electromagnet that create and produce a moving image. The movement of the aurora during the auroral storm is also caused by the modulation of the beam by changing electric and magnetic fields in the magnetosphere during the magnetospheric substorm.

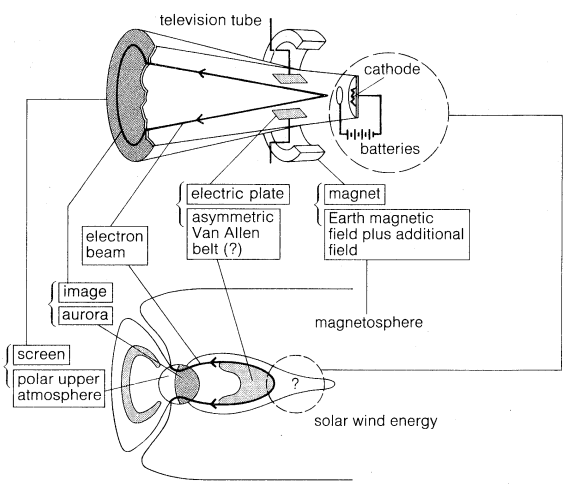


Figure 3: Analogy between the mechanisms of auroral displays and a television tube.

Auroral activity is, indeed, always associated with geomagnetic disturbances. For example, an intense ring current is known to deform considerably the Earth's dipole-like field in the outer magnetosphere. It has been found that the intensity of the ring current is closely related to the expansion toward the Equator of the auroral oval during an intense geomagnetic storm; that is to say, the ring-current field modulates the auroral-particle beam in such a way that it is shifted toward the Equator.

Recently there have been great efforts to measure, directly or indirectly, the electric fields in the magnetosphere and the aurora. Various types of electric probes have been devised to measure this field, but the most extensively used method has been to observe the drift velocity of a barium cloud released by a rocket or satellite and then ionized by solar radiation. An electric field of the order of 100 millivolts per metre that is directed toward the Equator has been detected in the vicinity of the auroral electrojet.

Several fundamental questions still puzzle scientists concerning the comparison of the aurora to a television tube. It is unclear what processes play the roles of the tube's cathode and anode in the magnetosphere, or what processes play the roles of the electric plate and electromagnet during the auroral substorm. It is not known how the energy carried by the solar wind is transformed into the energy of the aurora. It has been demonstrated that the southward directed interplanetary magnetic field is an essential ingredient in causing auroral substorms, so that the energy transfer appears to occur through interactions between the interplanetary and geomagnetic fields. A chain of processes leading to a substorm has not yet been unfolded, however. Current research efforts may unveil the secret of the aurora. It is not impossible that one day "magnetospheric weather" will be observed by watching auroral activity on the gigantic screen—the polar upper atmosphere.

BIBLIOGRAPHY. The literature on the aurora is vast and because of extremely rapid progress in this particular field during the last decade, no updated text dealing with all aspects of auroral studies is available. S.I. AKASOFU, S. CHAPMAN, and A.B. MEINEL, "The Aurora," in *Handbuch der Physik*, vol. 49/1 (1966); S.I. AKASOFU, *Polar and Magnetospheric Substorms* (1968); S. CHAPMAN, *Solar Plasma, Geomagnetism and Aurora* (1964); and S. MATSUSHITA and W.H. CAMPBELL (eds.), *Physics of Geomagnetic Phenomena*, 2 vol. (1967), provide reviews of morphological studies of the aurora and associated phenomena in the polar upper atmosphere. J.W. CHAMBERLAIN, *Physics of the Aurora and Airglow* (1961), is the most comprehensive text dealing with auroral and airglow spectra; new studies after its publication must, however, be supplemented by recent literature, such as B.M. MCCORMAC (ed.), *Aurora and Airglow* (1967), and *Atmospheric Emissions* (1969). W.N. HESS, *The Radiation Belt and Magnetosphere* (1968), is a comprehensive text on the electromagnetic environment surrounding the Earth which has an important relation to auroral phenomena. S. CHAPMAN and J. BARTELS, *Geomagnetism*, 2 vol. (1940); C. STORMER, *The Polar Aurora* (1955); and L. HARANG, *The Aurorae* (1951), are classic treatises on auroral phenomena.

(S.-I.A.)

Austen, Jane

While the birth of the English novel is to be seen in the first half of the 18th century in the work of Daniel Defoe, Samuel Richardson, and Henry Fielding, it is with Jane Austen that the novel takes on its distinctively modern

character in the realistic treatment of unremarkable people in the unremarkable situations of everyday life. In her six novels, published between 1811 and 1817—*Sense and Sensibility*, *Pride and Prejudice*, *Mansfield Park*, *Emma*, *Northanger Abbey*, and *Persuasion*—she created the comedy of manners of middle-class life in the England of her time, revealing the possibilities of "domestic" literature. Her repeated fable of a young woman's voyage to self-discovery on the passage through love to marriage focuses upon easily recognizable aspects of life. It is this concentration upon character and personality (second only to Shakespeare's, some critics say) and upon the tensions between her heroines and their society that relates her novels more closely to the modern world than to the traditions of the 18th century. The "daughter" of Dr. Samuel Johnson she is also the "parent" of Henry James. It is this modernity, together with the wit, realism, and timelessness of her prose style; her shrewd, amused sympathy; and the satisfaction to be found in stories so skillfully told, in novels so beautifully constructed, that helps to explain her continuing appeal for readers of all kinds. In her hands, the novel of entertainment becomes a work of art.

By courtesy of the National Portrait Gallery, London



Jane Austen, pencil and watercolour portrait by C. Austen, *c.* 1810. In the National Portrait Gallery, London.

Steventon: early life and education

Family and upbringing. Jane Austen was born on December 16, 1775, in the Hampshire village of Steventon, where her father, the Rev. George Austen, was rector. She was the second daughter and seventh child in a family of eight: six boys and two girls. Her closest companion was her elder sister, Cassandra, who also remained a spinster: the young clergyman to whom she was engaged died in 1797. Of the relationship between the sisters, the niece who had known them longest was to write: "They alone fully understood what each had suffered and felt and thought."

Her formal education began in about 1782, when the sisters were sent to be tutored by a Mrs. Cawley at Oxford; and, in 1783 or 1784, they moved to the Abbey School, Reading, where they remained until about 1787. Thereafter, their education continued at home. This was no deprivation, as the household at the rectory was unusually gifted; and it is difficult to imagine circumstances in which Jane Austen's genius could have been better developed.

Her father was a scholar who, before entering the church, had been a fellow of St. John's College, Oxford; and he encouraged the love of learning in his children. His wife, Cassandra (*née* Leigh), was a woman of ready wit, famed for her impromptu verses and stories. Together they created a home decidedly, although not forbiddingly, bookish. Reading and writing were enjoyed as family activities, and their reading was by no means confined to "polite" literature or to the accepted English classics. Samuel Richardson and Henry Fielding were favourite novelists; and even Mr. Austen kept up with the latest fiction, including the Gothic romances or thrillers by such popular writers as Mrs. Ann Radcliffe, whose supernatural horrors Jane Austen was to make fun of in *Northanger Abbey*. The great family amusement was acting. A Steventon dramatic company was recruited from the Austens and their neighbours; and the rectory barn was converted into a little theatre for productions in the summer holidays, while at Christmas, plays were performed in the house. Again, the repertoire was not restricted and included even the broader kind of 18th-century comedy.

So the fact that Jane Austen's early years were spent in an actively Christian household in a remote country village does not mean that her life was intellectually or socially limited. The family circle was large and changing and constantly supplied with news. Two of her four elder brothers, James and Henry, went to Oxford and there edited a literary periodical, *The Loiterer* (running to 60 issues between 1789 and 1790). It was undoubtedly planned and discussed in their sister's presence when they were at home. The younger brothers, Francis (Frank) and Charles, joined the navy and wrote home about their adventures during the Napoleonic Wars. The Steventon household included young men whom Mr. Austen was preparing for the university. And there was a wide circle of relatives and friends to visit (sometimes for months at a stretch) and to be visited by, particularly on Mrs. Austen's side, for she was a member of a large and well-connected family. George Austen belonged to the poorer branch of an ancient Kentish family and had, through family patronage, become clergyman in charge of two small parishes—Steventon and its neighbour, Deane—which brought in a modest but adequate income.

Thus, although Jane Austen's early life was unmomentous and at a distance from the social and political upheavals of the time, a lively and affectionate family circle provided a stimulating context for her writing. Moreover, her experience was carried far beyond Steventon rectory by an extensive network of relationships by blood and friendship. It was this world—of the minor landed gentry and the country clergy, in the village, the neighbourhood, and the country town, with occasional visits to Bath and to London—that she was to use in the settings, characters, and subject matter of her novels. She wrote in a letter of 1814, "3 or 4 Families in a Country Village is the very thing to work on."

Literary development. Her earliest known writings date from about 1787, and between then and 1795 she wrote a large body of material that has survived in fair copy in three manuscript notebooks: *Volume the First*, *Volume the Second*, and *Volume the Third*. In all, these contain 21 items: plays, verses, short novels, and other prose. They are the product of an analytical mind engaged in parody of existing literary forms, notably sentimental fiction (remorselessly burlesqued in *Love and Friendship*, in *Volume the Second*). In this early writing can be traced her development toward a more humane interest in the portrayal of realistic characters in realistic situations; and in many of these pieces the antecedents of the mature novels can be recognized.

Jane Austen's literary development also reflects the outlines of her emotional and intellectual growth between the ages of 12 and 17; and her passage to a more serious view of life from the exuberant high spirits and extravagances of her earliest writings is announced finally in *Lady Susan*, a short novel-in-letters written about 1793–94. This, the portrait of a woman bent on the exercise of her own powerful mind and personality to the point of social self-destruction, is, in effect, a study of frustration and of woman's fate in a society that has no use for woman's stronger, more "masculine," talents. How far can this be read as the young Jane Austen's own self-admonitory fable? In a letter written in 1815, the essayist and dramatist Mary Russell Mitford, who had never known Jane Austen personally, repeats a description of her as "a poker of whom everyone is afraid." Her

Lady Susan

source was a member of a family on bad terms with the Austens. But if a malicious statement can contain a measure of truth, the reference was probably to Jane Austen's silent, watchful presence—the young novelist making her mental notes of this one's stupidity and that one's folly. The evidence of *Lady Susan* suggests that to encounter Jane Austen as a girl of 17 could have been a formidable experience. Somewhere within this notion may stand a solution to the problem of her spinsterhood.

Relations with men

The question can only be put forward speculatively, for the record of Jane Austen's life is tantalizingly incomplete on the subject of her relationships with men. From January 1796 onward some of her letters survive, to tell of her enjoyment of local parties and dances in Hampshire, of visits to London, Bath, Southampton, Kent, and to seaside resorts in Devon and Dorset. A gossipy recollection of Miss Mitford's mother (who had left the Steventon neighbourhood when Jane Austen was seven) describes her as "the prettiest, silliest, most affected, husband-hunting butterfly [she] ever remembers." (Only a remnant of "prettiness" can be glimpsed in the one authentic portrait, Cassandra's sketch of about 1810: here the dark eyes, dark ringlets, and round cheeks attend an expression of slightly soured resignation.)

A number of the intimate family letters that have been preserved, especially two to her niece, Fanny Knight, of November 1814, provide remarkable confirmation that the severe, yet sensible and humane, morality of the novels was wholly at one with the practicing morality of her own life and with the morality she advocated for others. Fanny, then aged 19, was uncertain about her affection for a 21-year-old suitor. Jane Austen's letters of advice are a model—analytical rather than persuasive—requiring the girl, gently and sympathetically, to search her own heart, to discover for herself whether what she feels is really love and of what nature and quality. But she is emphatic on one point: "Anything is to be preferred or endured rather than marrying without Affection"; "nothing can be compared to the misery of being bound *without* Love." She sketches the one man "in a Thousand" who could be thought "perfection": "Where Grace & Spirit are united to Worth, where the Manners are equal to the Heart & Understanding. . . ." Perhaps it is these letters, rather than *Lady Susan* or the tittle-tattle of gossips, that tell us the truth about Jane Austen's single life and the moral urgency of the novels, in their preoccupation with the appraisal of human relationships, the precise quality of feeling between one person and another.

Jane's first recorded romantic association was a flirtation early in 1796 with Tom Lefroy, a handsome young Irishman, nephew of the rector of a village near Steventon. In 1798 or 1799 she may have refused Samuel Blackall, a fellow of Emmanuel College, Cambridge, who was then staying with the Lefroys. In November 1802 it seems likely that she agreed to marry Harris Bigg-Wither, the 21-year-old heir of a Hampshire family: but next morning she changed her mind. Then there are a number of mutually contradictory stories connecting her with someone (sometimes a naval officer, sometimes an army officer, sometimes a clergyman) with whom she fell in love but who died very soon after.

Since Jane Austen's novels are so deeply concerned with love and marriage, there is some point in attempting to establish the facts of these relationships. Unfortunately, the evidence is unsatisfactory and incomplete. Cassandra was a jealous guardian of her sister's private life, and after Jane's death she (and other members of the family) censored the surviving letters, destroying many and cutting up others. Stories in the family memoirs and documents are confused; and Jane Austen's own references in her letters are always ironic and evasive. The novels, however, provide indisputable evidence that their author understood the experience of love and of love disappointed.

This observation relates most obviously to her last novel, *Persuasion*. Yet it has some relevance, too, to the earlier novels, on which she was working during the period of the recorded romances. The earliest, *Sense and Sensibility*, was begun about 1795 as a novel-in-letters called "Elinor and Marianne," after its heroines. Between October 1796 and August 1797 she completed the first version of *Pride and Prejudice*, then called "First Impressions." In November 1797 her father wrote to inquire from the London publisher Thomas Cadell about the possibilities for its publication, but there was no answer. *Northanger Abbey*, the last of the early novels, was written about 1798 or 1799, probably as "Susan."

Up to this time the tenor of life at Steventon rectory had been propitious for Jane Austen's growth as a novelist. The family provided an appreciative audience. There was all the variety of the neighbourhood society, with a much-loved home to return to after long visits to friends, relatives, and married brothers (Edward, the Austens' third son, adopted as heir by wealthy cousins, opened his home at Godmersham, Kent, to his sisters and later provided them with 11 motherless nieces and nephews to console and entertain; and Henry, living in London, often claimed their company). This stable pattern ended in 1801, when George Austen, then aged 70, handed on his parish duties to his eldest son, James, and retired to Bath with his wife and daughters. For eight years Jane Austen had to put up with a succession of temporary lodgings or visits to relatives, in Bath, London, Clifton, Warwickshire, and, finally, Southampton, where the Austens lived from 1805 to 1809.

Meanwhile, in 1803, the manuscript of "Susan" had been sold to the publisher Richard Crosby for £10. He took it for immediate publication, but although it was advertised, unaccountably it never appeared. In 1804 she began *The Watsons* but soon abandoned it. It is an unhappy work. Its social picture is one of unrelieved bleakness; its central character is the most assailed of her heroines in distress. Its satire is sharp to the point of cruelty. It signals a failing of generosity, a loss of creative power, perhaps the consequence of disappointment in love and other sorrows. In December 1804 her dearest friend, Mrs. Anne Lefroy, died suddenly; on January 21, 1805, her father died in Bath.

Critical acclaim. Eventually, in 1809, Edward provided his mother and sisters with a large cottage in the village of Chawton, his Hampshire property, not far from Steventon, where they were joined by a close friend who had lived with them in Southampton. The prospect of settling at Chawton had already given Jane Austen a renewed sense of purpose. In April she wrote to Crosby to sound his intentions about "Susan," and, once installed at the cottage, she began to prepare *Sense and Sensibility* and *Pride and Prejudice* for publication. Two years later Thomas Egerton agreed to publish *Sense and Sensibility*, with the author's guarantee against loss. It came out, anonymously, in November 1811. Both of the leading reviews, the *Critical Review* and the *Quarterly Review*, welcomed its blend of instruction and amusement. Meanwhile, in February 1811, she had begun *Mansfield Park*, finished in the summer of 1813 and published in 1814. By then she was an established (though anonymous) author; Egerton had published *Pride and Prejudice* in January 1813; in November there were second editions of *Pride and Prejudice* and *Sense and Sensibility*. Between January 1814 and March 1815 she wrote *Emma*, which appeared in December 1815. In February 1816 there was a second edition of *Mansfield Park*, published, like *Emma*, by Byron's publisher, John Murray ("a rogue, of course, but a civil one," Jane Austen commented). *Persuasion* (written August 1815–August 1816) was published posthumously, with *Northanger Abbey*, in December 1817.

The years after 1811 seem to have been the most rewarding of her life. She had the satisfaction of seeing her work in print and well reviewed and of knowing that the novels were widely read. They were so much enjoyed by the Prince Regent (later George IV) that he had a set in each of his residences; and *Emma*, at a discreet royal command, was "respectfully dedicated" to him. The reviewers praised the novels for their moral entertainment, admired the character drawing, and welcomed the homely realism as a refreshing change from the romantic melodrama then in vogue. Although she was at pains to pre-

serve her anonymity and avoided literary circles, Jane Austen was nonetheless concerned about the reception of the novels, not least because they earned her money as well as praise. At Chawton, she remained a housekeeper and an affectionate and dutiful attendant upon her mother. To an ever-growing band of nephews and nieces she was the favourite "Aunt Jane," a prized confidante and adviser on literature and affairs of the heart.

For the last 18 months of her life, she was busy writing. Early in 1816 she set down the burlesque *Plan of a Novel, According to Hints from Various Quarters* (first published in 1871). Until August 1816 she was occupied with *Persuasion*. She looked again at the manuscript of "Susan" (*Northanger Abbey;* to which she was now referring as "Miss Catherine"). Then, in January 1817, she began her last work, *Sanditon* (so named by the family), writing and revising more than 24,000 words in less than eight weeks. It was finally put aside on March 18. This may have been a race against time, for her health had been in decline since early 1816. She supposed that she was suffering from bile. But the symptoms (reported, wryly and unself-pityingly, in her letters) make possible a modern clinical assessment that she was suffering from Addison's disease of the suprarenal capsules. Her condition fluctuated. A final burst of energy seems to have gone into *Sanditon*, with its robust and self-mocking satire on health resorts and invalidism. In April she made her will. In May she was taken to Winchester to be under the care of an expert surgeon. On the morning of July 18, at 4:30 AM, she died. Six days later she was buried in Winchester Cathedral.

Her authorship was announced to the world at large by her brother Henry, who supervised the publication of *Northanger Abbey* and *Persuasion* and contributed a "Biographical Notice of the Author," paying tribute to his sister's qualities of mind and character and recording her final words: in answer to a question about her last wants, she replied, with characteristic decorum and economy, "I want nothing but death."

Reputation

There was no recognition at the time that regency England had lost its keenest observer and sharpest analyst; no understanding that a miniaturist (as she maintained that she was and as she was then seen), a "merely domestic" novelist, could be seriously concerned with the nature of society and the quality of its culture; no grasp of the fact that Jane Austen was a historian of the emergence of regency society into the modern world. During her lifetime there had been a solitary response in any way adequate to the nature of her achievement: Sir Walter Scott's review of *Emma* in the *Quarterly Review* for March 1816, where he hailed this "nameless author" as a masterful exponent of "the modern novel" in the new Realist tradition. Commenting on this to her publisher, Jane Austen was cool. She had, she said, "no reason . . . to complain of the review, except at the (anonymous) critic's failure to mention *Mansfield Park*." After her death, there was for long only one significant essay, the review of *Northanger Abbey* and *Persuasion* in the *Quarterly* for January 1821 by Richard Whately, the eccentric but far-sighted theologian, logician, and political economist, later archbishop of Dublin. Together, Scott's and Whately's essays provided the foundation for serious criticism of Jane Austen: their insights were appropriated by critics throughout the 19th century. Modern critics remain fascinated by the commanding structure and organization of the novels, by the triumphs of technique that enable the writer to lay bare the tragicomedy of existence in stories of which the events and settings are apparently so ordinary and so circumscribed.

MAJOR WORKS

NOVELS: *Sense and Sensibility* (1811); *Pride and Prejudice* (1813); *Mansfield Park* (1814); *Emma* (published December 1815, dated 1816); *Northanger Abbey* and *Persuasion* (published together, posthumously, December 1817, dated 1818).

UNFINISHED WORKS AND JUVENILIA: The early novel-in-letters, *Lady Susan* and *The Watsons* (unfinished), were first published in the 1871 edition of J.E. Austen-Leigh's *Memoir of Jane Austen. Lady Susan* was republished from the original autograph by R.W. Chapman (1925), and *The Watsons*, also by R.W. Chapman (1927). *Sanditon* (left unfinished; extracts published in the 1871 *Memoir*) was first published in full from the original autograph by R.W. Chapman (1925). Three manuscript notebooks, *Volume the First, Volume the Second*, and *Volume the Third*, containing juvenilia written *c.* 1787–95, were first published in full by R.W. Chapman (*Volume the First*, 1933; *Volume the Third*, 1951) and B.C. Southam (*Volume the Second*, 1963). *Volume the Second* contains *Love and Friendship* (first published 1922).

BIBLIOGRAPHY

Bibliographies: GEOFFREY L. KEYNES, *Jane Austen: A Bibliography* (1929), places emphasis on the early editions. R.W. CHAPMAN, *Jane Austen: A Critical Bibliography*, 2nd ed. (1969), lists critical works as well as original and early editions. B.C. SOUTHAM, "Jane Austen," in the *New Cambridge Bibliography of English Literature*, vol. 3 (1969), lists collections, editions, letters, and critical and biographical books and articles, 1812–1967.

Manuscripts: Only eight literary manuscripts are known to have survived: two are privately owned, two are in the Pierpont Morgan Library, New York; two are in the British Museum, London; one is in the Bodleian Library, Oxford; and one in the library of King's College, Cambridge. B.C. SOUTHAM, *Jane Austen's Literary Manuscripts* (1964), provides a full bibliographical, historical, and critical account of the manuscript material. Of many thousands of letters written by Jane Austen, only 154 are known to have survived; the principal collections are in the Pierpont Morgan Library and the British Museum. The Jane Austen Memorial Trust maintains Chawton Cottage, Hampshire (her last home), as a place of literary pilgrimage; and holds the only collection of Jane Austen's personal possessions, and of other items belonging to the period and to the Austen family. The Jane Austen Society (Alton, Hampshire, England) is an association of those interested in her and her work.

Editions: The Oxford Illustrated Jane Austen, ed. by R.W. CHAPMAN, 6 vol. (1923–54; vol. 6 rev. by B.C. SOUTHAM, 1968), is the definitive, standard edition of the works. Vol. 5 contains *Persuasion* and *Northanger Abbey;* vol. 6, the minor works. The Everyman's Library edition of the 6 novels, by MARY LASCELLES, 5 vol. (1962–64; vol. 5 containing *Persuasion* and *Northanger Abbey*), provides a sound working text. The only complete edition of Jane Austen's surviving *Letters* is that by R.W. CHAPMAN, 2 vol. (1932; 2nd ed., with corrections and additional indexes, 1952). The World's Classics *Selected Letters of Jane Austen* (1955), also ed. by Chapman, is useful.

Biography and criticism: HENRY AUSTEN's brief "Biographical Notice of the Author," written as a preface to *Northanger Abbey* and *Persuasion* (1818), was the first public acknowledgement of the identity of the author of the novels. J.E. AUSTEN-LEIGH, *A Memoir of Jane Austen* (1870; enlarged in 1871; later reprinted), written in old age by her favourite nephew, draws on family letters and personal recollections: the 1871 edition includes extracts from Jane Austen's letters. It should be supplemented by W. and R.A. AUSTEN LEIGH, *Jane Austen: Her Life and Letters* (1913), which provides a more complete account of her life and, although not a full edition of the letters, quotes almost all of those known. LORD BRABOURNE, *Letters of Jane Austen*, 2 vol. (1884), provides an edition of letters, mainly to her sister, Cassandra, with a biographical introduction based on recollections. CONSTANCE HILL, *Jane Austen: Her Homes and Her Friends* (1902), is an informal but knowledgeable account, drawing on family papers now unlocatable. R.W. CHAPMAN, *Jane Austen: Facts and Problems* (1948), is not a full biography, but provides the first scholarly and authoritative examination and evaluation of the evidence. MARGHANITA LASKI, *Jane Austen and Her World* (1969), in appearance a popular, illustrated biography, is the best introductory life. MARY LASCELLES, *Jane Austen and Her Art* (1939; corrected 1941, 1954), remains the best account of the novelist's art, and contains an introductory biographical chapter. Other modern critical works include W. LITZ, *Jane Austen: A Study of Her Artistic Development* (1965); M. MUDRICK, *Jane Austen: Irony As Defense and Discovery* (1952), a brilliant and contentious interpretation; N. SHERRY, *Jane Austen* (1969); and A.H. WRIGHT, *Jane Austen's Novels: A Study in Structure*, 2nd ed. (1962). B.C. SOUTHAM in *Jane Austen: The Critical Heritage* (1968), collects and discusses the contemporary reviews and criticism, 1811–70; his *Critical Essays on Jane Austen* (1968), is a collection of newly written modern critical approaches. I.P. WATT, *Jane Austen: A Collection of Critical Essays* (1963), collects the best essays from 1870 onward.

(B.C.So.)

Austin, John

The most distinguished English writer on the philosophy of law after Jeremy Bentham, John Austin exerted a pro-

found influence on English jurisprudence and legal education. During the 20th century much of his work has been subjected to valid criticism, yet the lucidity and precision of his analysis of law and legal terminology have made a permanent contribution to our understanding of the basic concepts of jurisprudence. Austin was born March 3, 1790, at Creeting Mill, Suffolk. He began to study law in 1812 after five years in the army and from 1818 to 1825 practiced unsuccessfully at the chancery bar. His powers of rigorous analysis and his uncompromising intellectual honesty deeply impressed his contemporaries, and in 1826 when University College, London, was founded, he was appointed its first professor of jurisprudence, a subject that had previously occupied an unimportant place in legal studies. He spent the next two years in Germany studying Roman law and the work of German experts on modern civil law whose ideas of classification and systematic analysis exerted an influence on him second only to that of Bentham. Both Austin and his wife, Sarah, were ardent Utilitarians, intimate friends of Bentham and of James and John Stuart Mill, and much concerned with legal reform. Austin's first lectures, in 1828, were attended by many distinguished men, but he failed to attract students and resigned his chair in 1832. In 1834, after delivering a shorter but equally unsuccessful version of his lectures, he abandoned the teaching of jurisprudence. He was appointed to the Criminal Law Commission in 1833 but, finding little support for his opinions, resigned in frustration after signing its first two reports. In 1836 he was appointed a commissioner on the affairs of Malta. The Austins then lived abroad, chiefly in Paris, until 1848 when they settled in Weybridge, Surrey, where Austin died in December 1859.

Best known work

Austin's best-known work, a version of part of his lectures, is *The Province of Jurisprudence Determined*, published in 1832. Here, in order to clarify the distinction between law and morality, which he considered to be blurred by doctrines of Natural Law, he elaborated his definition of law as a species of command. According to Austin, commands are expressions of desire that another shall do or forbear from some act and are accompanied by a threat of punishment (the "sanction") for disobedience. Commands are laws "simply and properly so-called" when they prescribe courses of conduct, not specific acts, and are "set" by the "sovereign" (*i.e.*, the person or persons to whom a society renders habitual obedience and who render no such obedience to others). This is the mark distinguishing "positive law" both from the fundamental principles of morality, which are the "law of God," and from "positive morality," or man-made rules of conduct, such as etiquette, conventional morality, and international law, which do not emanate from a sovereign. *The Province* also contains a version of Utilitarianism in which "utility" is regarded as the index of God's commands and the test of the moral quality of general rules of conduct rather than of particular actions.

Austin viewed the doctrines in *The Province* as "merely prefatory" to the study that he termed "general jurisprudence": the exposition and analysis of the fundamental notions forming the framework of all mature legal systems. He devoted the main part of his lectures (published in 1863) to an analysis of such "pervading notions" as those of right, duty, persons, status, delict, and sources of law. Austin distinguished this general, or analytical, jurisprudence from the criticism of legal institutions, which he called the "science of legislation"; he thought both were important parts of legal education.

Bouts of nervous illness and self-distrust prevented Austin from fully utilizing his great powers; his life, as his widow wrote, was one of "unbroken disappointment and failure," in ironic contrast with his posthumous fame and influence. A long succession of English writers have echoed or elaborated his doctrines or, when opposing them, have accepted his conception of the analysis of legal concepts as the central concern of jurisprudence. In the United States jurists such as J.C. Gray and Oliver Wendell Holmes welcomed his bold distinction between law and morality as a major clarification.

The reaction to Austin's work at the turn of the century was severe. His command theory was condemned as a misidentification of all law with the product of legislation and a distortion of many types of legal rule. The severance of a purely analytical jurisprudence from moral criticism of law was criticized as sterile verbalism obscuring the social function of law and the judicial process. Some critics consider that Austin's doctrine of sovereignty confuses the ideas of legal authority and political power; others hold "legal positivism" responsible for subservience to state tyranny or absolutism.

Some of these criticisms are well founded, but even so Austin's work is of permanent value. The rigour and clarity of his analysis have demonstrated the complexity of many important legal and political concepts and the perennial need for just such an analytical study as he proposed; and repeated efforts to show precisely where his simple distinctions between law and morality are wrong have increased the understanding of both.

BIBLIOGRAPHY. JOHN STUART MILL, "Austin on Jurisprudence," in *Dissertations and Discussions*, vol. 4, pp. 157–226 (1874); R.A. EASTWOOD and G.W. KEETON, *The Austinian Theories of Law and Sovereignty* (1929); E.M. CAMPBELL, *John Austin and Jurisprudence in Nineteenth Century England* (1959).

(H.L.A.H.)

Australia

The most striking characteristics of the continent of Australia, a vast, 3,000,000-square-mile (8,000,000-square-kilometre) landmass in the Southern Hemisphere, are its isolation, its low relief, and the aridity of much of its surface. Its isolation from other continents explains much of the strangeness of Australian plant and animal life; its low relief results from the long and extensive erosive action of the forces of wind, rain, and the heat of the sun during the great periods of geological time when the continental mass was elevated well above sea level.

This article describes the geological history of the development of Australia as a continent; its current physical geography; its plant and animal life; and the natural resources available for exploitation by man. The article AUSTRALIA, COMMONWEALTH OF, describes the contemporary aspects of the life of man on the continent; the people and their relationship to the landscape; the national economy; and political and cultural life. The related article AUSTRALIA, HISTORY OF, describes the historical associations of man with the continent.

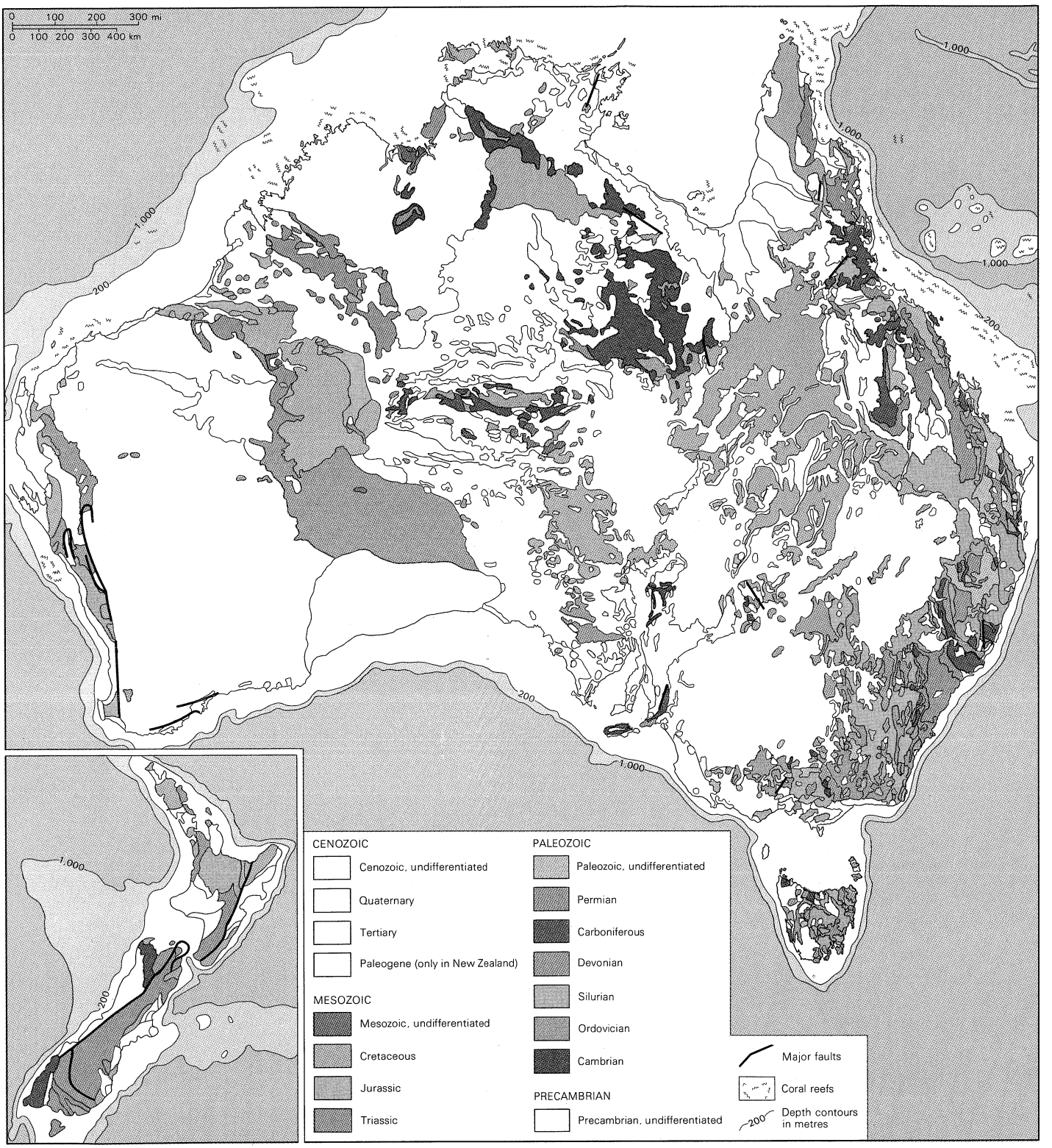
GEOLOGICAL HISTORY

A popular misconception about the observed peculiarities of Australia is that it is the oldest of continents. The statement is, in fact, unjustified: the cores of all of the continents are of approximately the same age, and rocks of all geological ages, of sedimentary as well as volcanic origin, are found in Australia. The misconception about its age stems partly from its relative paucity of surface relief, the result of relative stability and the absence of recent strong movements of the Earth's crust, rather than the old age of its rocks. The geologically young mountain-building forces, such as those that formed the Himalayas in Asia, the Alps in Europe, the Rocky Mountains in North America, and the Andes in South America, have bypassed Australia.

Comparison with other continents

Australia, like Eurasia and Africa, is strikingly different from the landmass of North America, which consists basically of an old core, the Canadian Shield, surrounded by mountain zones formed from belts of folded rocks of varying ages, all younger than the central core. Australia is strikingly asymmetric, with a folded belt of rocks bordering its east coast only and extending westward to no more than one quarter of the continent's width.

Like the other continental masses, Australia is built of three types of structural units. In the west is the West Australian Craton, a comparatively stable and resistant block consisting mainly of ancient rocks that have not been folded during at least the last 1,000,000,000 years. In the east is an ancient fold mountain belt, greatly eroded but relatively uplifted again. Between these two

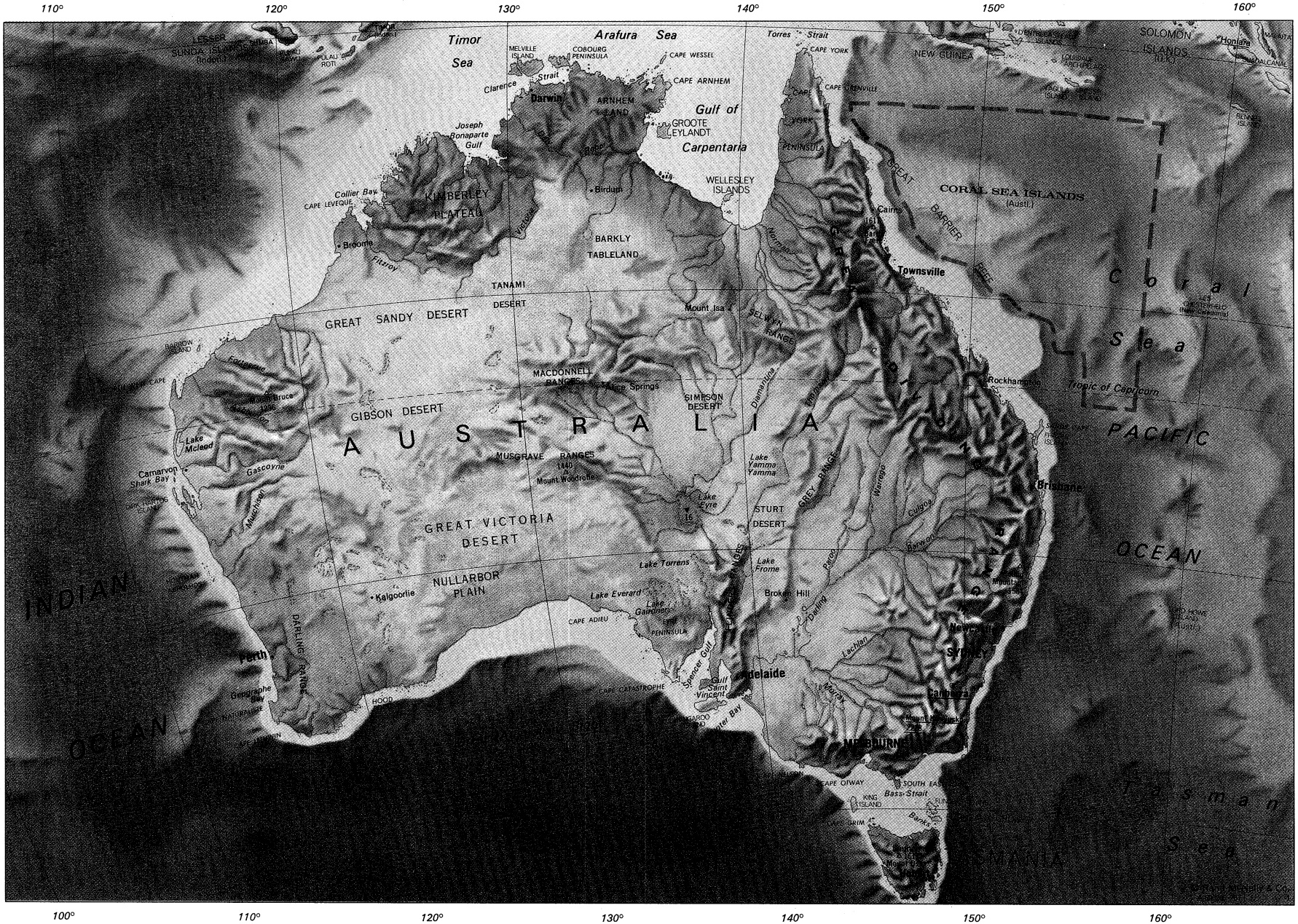


Geological structure of Australia and (inset) New Zealand.
By courtesy of the Department of National Development, Canberra

lies the platform-like area occupied by the central plains, a region where crystalline or folded rocks are overlain by relatively thin sequences of flat-lying or only gently deformed sediments. Each of these major units is complex. The West Australian Craton, for example, includes, in addition to extensive outcrops of crystalline rocks in the core area itself, many blocks of sedimentary strata deposited in troughs and basins within this structural unit. The platform consists of three major, and numerous minor, basin structures. The Eastern Uplands include upfaulted and depressed structural blocks, many large, exposed, formerly molten intrusions (batholiths) of granite, and extensive lava flows of the Cenozoic Era.

The continental outline

The outline of the Australian continent, unlike those of South America and Africa, and the Indian subcontinent, is not roughly triangular with an apex pointing southward, though some resemblance to this distinctive configuration of the other continents and subcontinents can be seen in the shape and position of Tasmania. This island is part of the continent if the base of the shallow peripheral submarine area known as the continental shelf is taken as its outer boundary. The continental shelf and the



AUSTRALIA Size of symbol indicates relative size of town • ◉

Elevations and depths in metres

slope separating it from the ocean depths are moderately wide in the south and southwest and spectacularly wide in the northwest from Exmouth Gulf to the Timor Sea, where the distance of the slope base from the shore exceeds 300 miles, and the line marking depths of over 650 feet is over 250 miles from the shore. To the north of Australia, the continental shelf extends to the island of New Guinea. The shelf is narrow along the east coast, where the Great Barrier Reef (*q.v.*) marks its edge. A trough, with depths of about 3,000 feet, separates the Great Barrier Reef from the largely submerged Queensland Plateau, where only a few small islands and reef tops reach sea level. Off Brisbane and Sydney, the continental shelf is only some 30 miles wide, and the slope plunges steeply to the floor of the Tasman Sea.

Douglass Baglin



A portion of the Great Barrier Reef, lying off the coast of northeast Australia and forming the largest coral structure in the world.

The discovery of vast and varied mineral deposits in Australia stimulated intensive studies of its geological history and structure. At the same time, new developments in general geological theory, including the concepts of continental drift and ocean floor spreading, have led to renewed speculation regarding possible former land connections between Australia and other continents, particularly Antarctica. As a result, the answer to the problem of how the history of the Australian continent fits into the picture of the development of the Earth's crust as a whole will have to come from a deeper understanding not only of the geology of the continent but also of that of the ocean surrounding it.

The oldest rocks: the Precambrian. The oldest rocks of the continent occur in the southwest, where they form part of the Western Australian Shield, an area underlain by rocks that have not been folded since early Precambrian (Archaean) time. These oldest rocks are now thought to be about 3,000,000,000 years old. The Yilgarn nucleus of the Western Australian Shield extends over 230,000 square miles (600,000 square kilometres) and consists mostly of granite, with belts of altered sedimentary rocks (metasediments) and greenstones, the latter including ancient rocks rich in economic mineral deposits, as well as basic intrusions of formerly molten rocks. Similar to metal-bearing greenstone rock belts occurring in the nuclei of the ancient shield areas of other continents, these are being studied in the hope that the origin of these rock belts and their ore bodies will be explained in terms of the state of the Earth's crust and mantle at the beginning of geological time. The conditions under which the continental crust was formed and then reworked are not yet known, but geologists believe that some of the peculiarities of ancient rocks, their structure, and their composition can be explained only by assuming that the crust of the Earth was originally much thinner than it is today.

Formation of the continental crust

The Yilgarn nucleus does not reach the west and south coasts of Western Australia, where younger Precambrian rocks occur. Those on the west coast are separated by the Darling Fault, which forms the western margin of the shield near Perth, and by the Perth Basin. The Frazer Fault apparently marks the boundary between rocks of the main shield area that were metamorphosed by heat and pressure more than 2,500,000,000 years ago and others to the southeast that were similarly affected less than 2,000,000,000 years ago. Each of these major fault lines is about 600 miles long and is considered as a major lineament in the continental structure. The older rocks appear again in the Pilbara nucleus in the northwest, where some granites are more than 3,000,000,000 years old and where altered volcanic, greenstone, and sedimentary rocks also occur.

Large platforms and basins separate the Yilgarn and Pilbara nuclei of the Western Australian Shield from other occurrences of ancient rocks in the northwest and north of the continent and in central Australia. It is not yet known whether a major part of these rocks are as old as those that have been dated as Archaean in the west. Therefore, the earliest geographic configuration of the Australian continent cannot be reliably reconstructed; it could have extended northward and eastward far beyond the limits of the present Western Australian Shield.

Old (Archaean) rocks on the margins of the Kimberley Block form its basement; they occur also in the Katherine–Darwin area of the Northern Territory, in the Rum Jungle Complex. No definite datings of rocks in the east indicate ages as great as those of the oldest shield rocks, although part of the Arunta Complex, north of Alice Springs in the centre of Australia may be approximately 2,250,000,000 years old.

The post-Archaean geological history of Australia can be seen in a clearer framework; the general terms Early, Middle, and Late Proterozoic, or the regional terms Nullaginian, Carpentarian, and Adelaidean, are used to identify intervals in the time scale of this period and are documented by sedimentary rock sequences.

The Nullaginian System comprises platform cover south of the Pilbara nucleus, including the vast banded iron ore formations of the Hamersley Ranges (notably the Mt. Bruce supergroup, which also contains volcanic rocks dated at 1,950,000,000 to 2,200,000,000 years ago) and the rocks in the east Kimberley area. The Carpentarian System comprises a belt of sedimentary rocks—the McArthur Basin, southwest of the Gulf of Carpentaria—that may be 1,500,000,000 to 1,800,000,000 years old; younger granites in the Katherine–Darwin region; and the sediments and intrusive granites of the Mount Isa–Cloncurry region. The rocks of these regions are rich in mineral deposits. In South Australia, the basement rocks (gneisses) of Eyre Peninsula in the west and the Mt. Painter Complex in the northeast, and extending to Broken Hill in New South Wales (Willyama Block), may be of the same age as the rocks of the Carpentarian System, although the iron formations of the Middleback ranges have been generally considered to be as old as the early Proterozoic iron formations found in the Pilbara region.

Post-Archaean regional rock units

The Adelaidean System is made up of rocks of Late Precambrian age. Consolidation of the Musgrave Block in central Australia and intrusion of various molten rocks took place during the time interval 1,000,000,000 to 1,400,000,000 years ago, forming an extension of the Western Australian Shield. At the same time, sedimentary rocks were laid down in the Adelaide Geosyncline, a vast downwarping of the Earth's surface lying to the southeast. Sedimentation continued into mid-Cambrian time in a shallow marine trough extending from the coast south of Adelaide at least 400 miles northward, through the Mt. Lofty and Flinders ranges, and measuring 400 kilometres from west to east, from Lake Torrens to beyond Broken Hill. These rocks, with a total maximum thickness of over 10 miles, formed in an intermittently

The Adelaide Geosyncline

sinking trough in which the accumulating sediments were subjected to moderately strong folding and faulting in Late Precambrian and Early Paleozoic time. This Adelaidean sedimentary sequence has characteristic features of outstanding importance for Earth history: most of the sediments, which have remained unaltered, contain definite evidence of a major glacial period; they also contain the oldest known rich and varied assemblage of fossil animals, all of which were essentially soft-bodied. Rocks of similar age also occur in the Peake and Denison ranges west of Lake Eyre, in the Amadeus Basin, and on the margins of the Kimberley Block. In the area there is also clear evidence of a Precambrian glaciation. Its earlier phase has been dated at about 740,000,000 years, while the later phase was little older than 660,000,000 years. This is the likely time range of glaciation in the Adelaide Geosyncline and probably also of related glaciations in other continents: it was probably one of the most extensive periods of cold climate in the Earth's history. No definite Precambrian animal remains are known from rocks that are older than the onset of this glaciation; on the other hand, primitive aquatic plant fossils have been found abundantly in earlier Precambrian strata.

Paleozoic history of eastern Australia. The Adelaide Geosyncline is often considered as a precursor of the more extensive fold belts of eastern Australia. These form another great downwarp known as the Tasman Geosyncline. While the rocks of the Adelaide Geosyncline were folded in Late Cambrian time, about 500,000,000 years ago, and elevated so that sedimentation ceased, the Tasman Geosyncline began to form in Cambrian time, the first period of the Paleozoic Era, which was to last for the next 300,000,000 years.

The Tasman Geosyncline

Cambrian strata. Cambrian strata, with marine fossils indicating their age, are known only from a few areas in eastern Australia: Tasmania, central Victoria, and northeast of Broken Hill in western New South Wales. It is possible that Cambrian rocks outcrop also on the coast of southern New South Wales. These few occurrences are insufficient to support any well-founded reconstruction of the geography of Australia in Cambrian time, but they do indicate a fundamental difference in conditions of sedimentation in the east where the rocks are similar to those deposited in deep geosynclinal troughs generally, and the large areas in southern, central, and northern Australia, including the Amadeus and Georgina basins, where the Cambrian rocks were mostly formed in shallow water. These differences are the expression of the relative stability of the western areas as compared with the greater geological mobility of eastern Australia during Paleozoic time, although any boundary between these two areas is likely to be gradational rather than a definite line.

A generalized western margin of the Tasman Geosyncline might be drawn from the north Queensland coast, at about longitude 144° E and latitude 14° S, around the Precambrian Georgetown area to Charters Towers, south-southwestward to near Wilcannia on the Darling River, and thence either due south or southwestward to Kangaroo Island. The discontinuous folded and granite-intruded Paleozoic rocks east of this somewhat hypothetical boundary are now separated by generally flat-lying younger rocks deposited in separate sedimentary basins. As there is no obvious continuity of folded zones and of belts of contemporaneous folding from the south coast to the north of eastern Australia, it has been suggested that a Lachlan Geosyncline in the southeast (mainly in Victoria and south central New South Wales) should be distinguished from the New England Geosyncline, the western boundary of which extends from Newcastle to south of Townsville. They are two component parts of the Tasman geosynclinal belt, similar in rock types, but different in age and geographic position.

The Lachlan Geosyncline

Ordovician rocks. The geography of eastern Australia in Ordovician time, from 430,000,000 to 500,000,000 years ago, is sufficiently well known to permit some general statements. The Lachlan Geosyncline was the site of sedimentation that took place in a series of troughs separated by rising upwarped ridges. No western shoreline can be defined, but the characteristics of Ordovician rocks in the Broken Hill region, and the evidence of bores in northeastern South Australia, indicate that these localities lay outside the geosynclinal belt in platform areas where subsidence was less intense. The geosynclinal nature of most of the rocks of Victoria, and southern and central New South Wales, is indicated by the great thicknesses of marine rocks deposited partly in deep quiet water (as in the case of black shales) and partly by mud-laden currents flowing from higher areas of the sea floor, producing the muddy sandstones known as graywackes. The rocks in the central part of the geosyncline were altered by heat and pressure, and volcanic rocks are widely distributed. In many parts of the Lachlan Geosyncline, folding movements (known as the Benambran orogeny, or mountain-building movement) occurred at the end of Ordovician and in early to mid-Silurian time. Accompanied by granite intrusions, these movements are thought to have caused the changes in composition and texture of the Ordovician rocks in the Snowy Mountains.

Silurian rocks. The Benambran mountain-building movements changed the geographic framework of the sedimentation that occurred during Silurian times (from 395,000,000 to 430,000,000 years ago). The eastern margin of the Lachlan Geosyncline now extended from the Melbourne area northward through Cobar to the Adavale Trough. In the Melbourne Trough there are thick Silurian graywackes and mudstones with graptolites and other fossils, and sedimentation extended into Tasmania, where shallow-water rocks were laid down. Farther east, in southern New South Wales, the Yass Shelf was an area of volcanic rocks, limestones, and shales that are now famous for their fossils. Silurian rocks are not well known in the New England Geosyncline and in southern Queensland, but farther north there is another peripheral shelf development, with Silurian limestones occurring along the western margin of the Tasman geosynclinal belt and the Chillagoe Shelf and its southward extension.

Devonian rocks. Deposition in the Lachlan Geosyncline generally continued from Silurian into Lower and Middle Devonian Periods, from 350,000,000 to 400,000,000 years ago, although the earth-movements of the Bowning orogeny occurred at the end of the Silurian and through early Devonian. Evidence of this can be seen as an unconformity, or break, in the rock layers of the sedimentary series of the Canberra–Yass area: molten rocks were intruded into the sedimentary rocks during this period. This activity again changed the geography of sedimentary areas in the Lachlan Geosyncline. In particular, a shallow-water platform with limestone deposition extended from eastern Victoria (Buchan Caves) to the Yass–Taemas area. Geosynclinal deposition during early and mid-Devonian time was widespread in the New England Geosyncline and its extension in southeastern Queensland, but only corals preserved in isolated limestones permit precise dating of the rock layers.

The Tabberabberan earth movements

The Tabberabberan period of earth movements, which occurred at the end of mid-Devonian time, had a profound effect on the geographic development of eastern Australia. Marine sedimentation in the Lachlan Geosyncline ended, and nonmarine rocks, in part red sandstone beds resembling similar rocks in northern Europe, were deposited. Discoveries of abundant remains of armoured fishes—notably *Bothriolepis*, which is also found in central Australia, North America, Eurasia, and Antarctica—have been made in western New South Wales. Remains of the scale-tree *Leptophloeum* are also common. Erosion of uplifted areas in central Australia, Victoria, and most of New South Wales must have produced the material for extensive Late Devonian and Early Carboniferous lake and river deposits. These include sandstones of the Grampian Mountains in western Victoria.

Carboniferous and Permian rocks. In the eastern troughs, from New England to eastern Queensland, geosynclinal deposition continued under the influence of volcanic activity. In the Carboniferous Period, about 300,000,000 years ago, nonmarine rock deposition occurred in the south and west, while marine sediments ac-

cumulated in the east. The central part of New England was uplifted. The earliest Paleozoic glaciations in eastern Australia are of Late Carboniferous age and have been considered as the result of mountain glaciation.

The youngest rocks in the New England Geosyncline that were deposited before final deformation are of Permian age and marine origin. In Middle and Late Permian time the great New England batholith was emplaced in the older rocks, which were heavily folded toward the west. A line from the mouth of the Hunter River near Newcastle in the south to the eastern margin of the Bowen Basin in the Rockhampton–Bowen area marks the western limit of the Hunter-Bowen orogeny, which was the last of the major mountain-building phases in the Tasman geosynclinal belt and which ended its existence as a belt of troughs of sedimentation.

The Hunter-Bowen earth movements

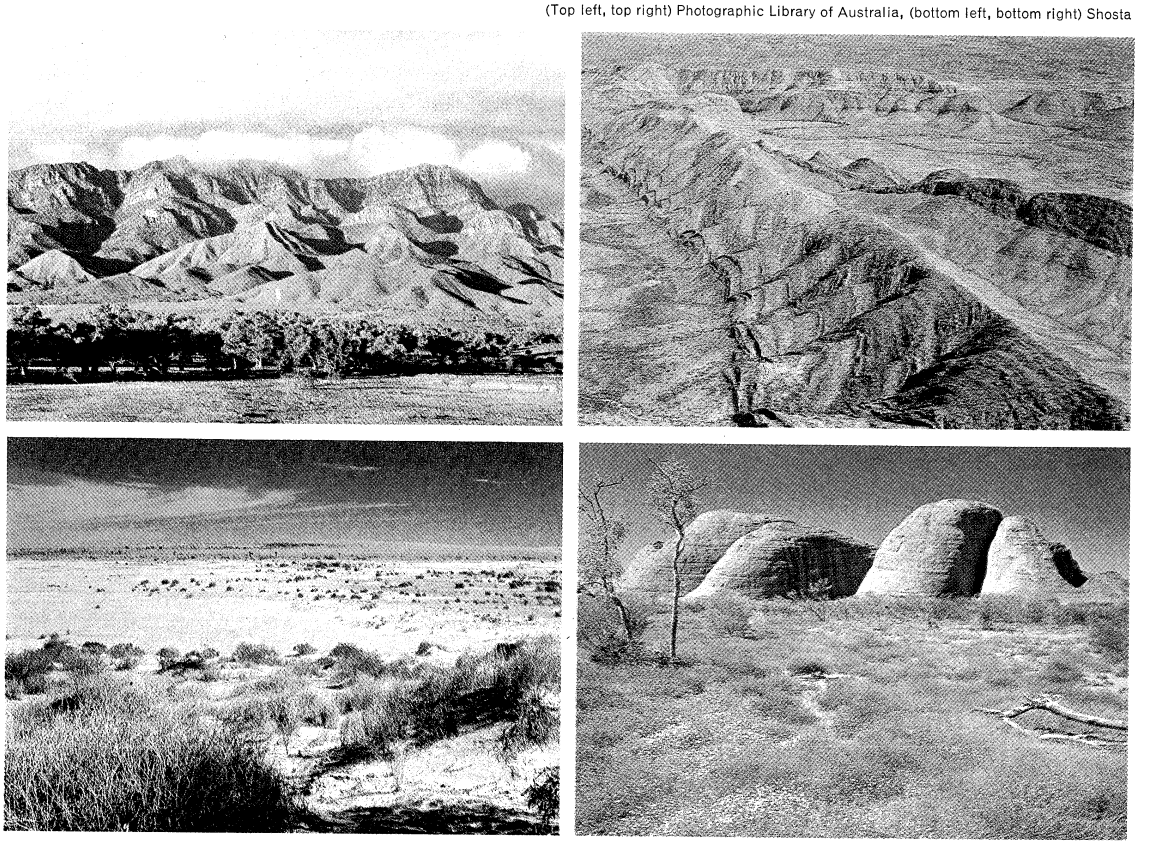
Sedimentary basins and the Mesozoic history of Australia. Quite distinct from the history of the successive, more or less elongated, highly mobile (in geological terms) troughs discussed above is the development of sedimentary basins over the Australian continent. Some of these developed at various times on the Precambrian basement of the western and central parts of the continent, while in other areas this type of sedimentation followed the consolidation of folded belts of Paleozoic rocks and interrupted their continuity. Others are superimposed on differently shaped pre-existing basins.

Like the Perth Basin, which adjoins it to the south, the Carnarvon Basin of Western Australia began to form in Silurian time and contains Late Paleozoic deposits, including Early Permian glacial sediments, followed by Lower Triassic and, in places, nearly 13,000 feet of Jurassic sediments, dating from about 140,000,000–180,000,000 years ago. During the Upper Cretaceous Period, about 100,000,000 years ago, sedimentation in the Carnarvon Basin changed from sands and shales to limestones and chalks, and this may be related to a change to warmer water conditions, which continued in the north until Miocene time, less than 25,000,000 years ago. The west coast of the continent was, to varying extents in the course of the last 400,000,000 years, a geologically mobile shelf area, with an alternation of marine and nonmarine deposits and of transgressions and regressions of the sea.

In northwestern Australia the shallower Canning Basin with the adjoining deeper Fitzroy Trough formed a large subsidence area between the Western Australian Shield and the Kimberley Block. The Joseph Bonaparte Gulf and Ord basins extend along its northern and eastern margins on the West Australian–Northern Territory border. These basins date back to Ordovician time. The Canning Basin may have been linked eastward, through the Officer and Amadeus basins of central Australia, with the Tasman geosynclinal belt in the east, and, again in Permian time it may have extended southeastward in a similar manner. Such temporary configurations are exceptional in ancient Australian geography. As a rule, the structural independence of the basins is supported by differences in their fossil remains. An outstanding development of Devonian reef complexes is known from the northern edge of the Canning–Fitzroy basins, while Permian glacial deposits occur on its southern margin and continue into South Australia, Victoria, and New South Wales.

The Amadeus Basin. The Amadeus Basin in central Australia is a large depression extending east–west for 450 miles and north–south for no more than 160 miles and filled with many varieties of ancient sedimentary rocks. It is situated between the so-called Musgrave structural block in the south and the Arunta structural complex in the north, and both have moved toward, and

(Top left, top right) Photographic Library of Australia, (bottom left, bottom right) Shosta



Australian landscapes.
(Top left) St. Mary Peak, Flinders Range, South Australia. Mineral traces are evident in upper crests. (Top right) Macdonnell Ranges, west of Alice Springs, Northern Territory. The low parallel ridges, cut in folded strata, run practically unbroken for 150 miles. (Bottom left) Simpson Desert, central Australia, one of the most desolate areas of the continent, contains neither water nor habitation. In the foreground are sand humps covered with spinifex, the scattered vegetation characteristic of the desert. (Bottom right) Mt. Olga, a huge red sandstone monolith, rises 1,500 feet above the plain, south of Lake Amadeus in the Northern Territory.

folded sediments of, the basin at different times. Some sediments were deposited during the folding of the Macdonnell Ranges near Alice Springs, which later erosion shaped into spectacular scenery. The isolated sandstone masses of Ayres Rock and Mt. Olga are now considered to be of Cambrian age.

The Canning Basin continued to exist as a depositional trough during Jurassic time, when only its northwestern margin received marine sediments, and through the early Cretaceous Period when a great marine transgression joined it again southeastward with the Eucla and Officer basins and eastward with the Great Artesian Basin.

The Eucla and Officer basins. The Eucla Basin is a large but shallow depression lying between the Western Australian Shield and the basement rocks of Eyre Peninsula. Permian and Cretaceous rocks are known to exist from the evidence of bores. Above them lie Tertiary limestones, which are no more than a few hundred feet thick, perfectly flat-lying, with their surface forming the arid Nullarbor Plain. They are riddled with caves, many of which remain unexplored. The Officer Basin, south of the Musgrave Block, is filled with late Precambrian and Cambro-Ordovician rocks.

The Great Artesian Basin. The Great Artesian Basin is a large (670,000 square miles) and complex, generally slowly subsiding area, in which, by Jurassic time, several previously separate basins (such as the Cooper Basin, with its rich gas-bearing Permian sands) had been incorporated. The Jurassic nonmarine sands of the Great Artesian Basin, which spread from the eastern highlands to the western margins of the basin, provide its most important sources of water. By mid-Cretaceous time the sea invaded the basin from the northeast and southwest, but the early Upper Cretaceous is represented by marine fossil-bearing deposits only near the northern coasts. The northernmost part of the Great Artesian Basin is separated from its main body by a basement ridge that hardly reaches the surface; north of it only Cretaceous and younger rocks are found. The southern boundary is formed by extensive basement outcrops in the Broken Hill–Wilcannia–Cobar areas, and the lowlands to the south are considered a separate unit, the Murray Basin. Permian and Mesozoic rocks play only a minor part in its filling. Its rocks are mostly of Tertiary age, marine in the west and mainly nonmarine in the east. A range of granitic hills in the south, the Padthaway Ridge, extending southeastward from Murray Bridge, forms a persistent boundary between the Murray and Otway basins. The Otway Basin, which extends into western Victoria and the continental shelf offshore, contains much greater thicknesses than the Murray Basin, of Cretaceous and Tertiary rocks in alternating marine and nonmarine development. Oil exploration in the Murray and Otway basins, and in the adjoining submarine Bass Basin between Victoria and Tasmania, had not been commercially successful by the early 1970s, but significant oil and gas reserves had been found offshore in the Gippsland Basin, in rocks of early Tertiary age.

The Murray and Otway basins

The Sydney Basin. The Sydney Basin was superimposed on a part of the folded rocks of the Tasman geosynclinal belt in Permian time and was filled with a considerable thickness of Permian and Triassic marine and nonmarine rocks. Overlapped by the Great Artesian Basin in the northwest, it continues southeastward under the continental shelf. The Permian begins with shallow-water marine sediments with erratic blocks indicating continued glaciation of the adjoining land, and volcanic material. The plant fossils are typical of the Permian of the southern continents. Coal deposition followed, and after another marine interlude, the Permian Period ended with the formation of further coal-bearing sediments, while volcanic activity and some earth movements continued. The Triassic sequence of the Sydney Basin consists of the Narrabeen Group of sandstones and shales, the Wianamatta Group, and the Hawkesbury Sandstone, which forms conspicuous cliffs around Sydney Harbour and is much used as a building stone.

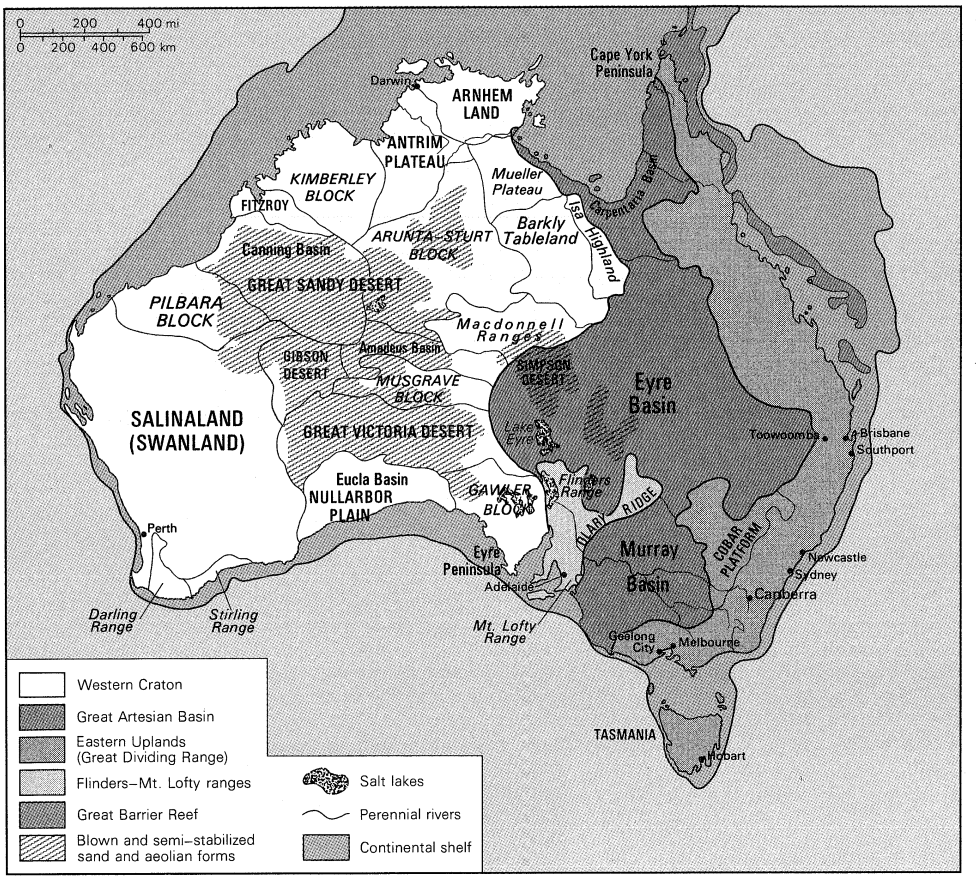
The Clarence–Moreton and Maryborough basins. On the border between New South Wales and Queensland, from south of Grafton to north of Brisbane, the Clarence–Moreton Basin developed in Triassic time as an offshoot of the Great Artesian Basin. It contains basic volcanic rocks and coal measures of several ages. The Maryborough Basin is the easternmost of the sedimentary basins. It is filled with nonmarine Triassic and coal-bearing Jurassic strata, followed by volcanic rocks, Lower Cretaceous marine sediments, and coal measures. These beds are the most strongly folded post-Paleozoic rocks in eastern Australia. It is now known that in other basins the Mesozoic and Tertiary rocks are slightly folded, but it is believed that the observed folds, such as those in the central part of the Great Artesian Basin, are due to movements of the underlying older rocks. The more intense folding on the eastern seaboard has been thought to be related to intrusions of Jurassic granites and a late structural phase in the development of eastern Australia.

Tasmania. The most important Mesozoic event in Tasmania was the large-scale (covering 2,000 cubic miles) intrusion of dark, layered volcanic rocks in Early to mid-Jurassic time. This process produced rocks similar to those in the Karroo region of South Africa, in Brazil, in Antarctica, and also in the Palisades of the New York area, in the Northern Hemisphere.

Cenozoic history of the Australian continent. During the Cenozoic Era, occupying the last 60,000,000 years of geological time, true mountain-building movements were confined to the island arcs in the north and east, but the Australian continent was not altogether stable. On the mainland, movements of the basic structural blocks produced some folding in the sedimentary cover of the basins, and uplifts and downwarps restricted or extended their boundaries. Faulting in the southeast and east led to outpouring of large fields of basalts in Victoria in mid-Cenozoic and Late Cenozoic time, and also to scattered but important volcanicity along the east coast. The mid-Miocene (20,000,000 years old) volcanoes of the Warrumbungle Mountains and the Tweed River areas in northern New South Wales and the spectacular Glass House Mountains in southeastern Queensland are the most important examples of this activity. In Victoria and parts of New South Wales, basaltic lava flows followed ancient valleys and buried their stream-deposited gravels. Known as "deep leads," they are locally gold bearing.

Late volcanic activity

The relative ages of the geological events that took place during Cenozoic time are still under study. There is no doubt about the late Cenozoic age of the uplift of the eastern highlands and also of the Flinders Ranges in South Australia, and the corresponding retreat of the sea from the Murray Basin. This process seems to have occurred intermittently throughout Tertiary time, perhaps culminating in a Pliocene–Pleistocene phase that has been named the Kosciusko Epoch, after the highest mountain on the continent (7,316 feet [2,230 metres]). During the Pleistocene ice age, an area of some 400 square miles around this mountain, and some surrounding valleys, were glaciated, as were large areas of the Tasmanian highlands. This weak but obvious expression of a worldwide climatic change is only one phase in a sequence of changes that affected Australia during Cenozoic time. Like the sequence of structural movements with which it is likely to be connected, the history of climatic changes is still disputed. The evidence of plant and animal fossils, together with determinations of the temperatures of the period, indicate that the climate in Early Tertiary (Eocene) time, some 50,000,000 years ago, was temperate in the south to tropical in the north. It was also humid, with rivers flowing to the south coast and coal-forming swamps existing in the areas of Eyre Peninsula and south Gippsland. Currents warmer than at present reached the south coast in early Miocene time (some 25,000,000 years ago), while during the late Miocene and Pliocene epochs there was a fall in temperature and a change in the plant life. This came to be dominated by eucalypts and acacias, replacing the southern beech (*Nothofagus*) that had been dominant in Early and mid-Tertiary time. Climatic change is also reflected by other evidence, for example by the large extent of lateritic weathering, causing the formation of red, iron-rich residual deposits as the result of



Physiographic regions of Australia.

Recent climatic changes

alternating wet and dry seasons. Whether the change to the present arid conditions of the inland regions was gradual and progressive is uncertain. Some soil scientists and biologists have postulated a period of greater aridity than the present, dated a few thousand years ago, but this is disputed by others. There is no doubt about changes of sea level as well as earth movements during the Pleistocene ice age: the eastern highlands were uplifted by some 2,600 feet, as was the western plateau, and there were corresponding changes of wind and ocean water-current patterns. These factors combined in an intricate manner with worldwide climatic changes to alter the climate during the last few million years. The last of the great hippopotamus-like marsupial herbivores (*Diprotodon*, *Nothotherium*) and giant flightless birds (*Genyornis*) died out when the great lakes in the centre of the continent (Lake Eyre [*q.v.*], Lake Callabonna, etc.) turned into salt pans only a few thousand years ago.

(M.F.G.)

PHYSICAL GEOGRAPHY

Relief. *Overall characteristics.* Australia is a land of great plains. The island continent is almost 3,000,000 square miles in area, but of this, only 6 percent is above 2,000 feet elevation. Its highest peak, Mt. Kosciusko, rises to only 7,316 feet. This situation stems in part from Australia's position at the edge of a zone of significant and recent earth movement and in part from the long periods of geological time during which Australia has been subject to weathering and erosion.

Patterns of faulting and folding in large measure control the distribution and attitude of rocks and thus play a significant part in determining the shape of the land surface. But the nature and intensity of the processes at work at and near the land surface also give rise to characteristic assemblages of forms. Australia is an arid continent; fully one-third of its area is occupied by desert, another third is steppe or semidesert, and only in the north, east, and southeast is rainfall adequate to support a vegetation that significantly protects the land surface.

The Western Craton. The ancient west Australian core area, known geologically as a shield, or craton, is subdivided, most obviously in the north and west, by long, straight (or only gently arcuate) fractures called lineaments. These fractures delineate prominent rectangular or rhomboidal blocks, some of which have been raised to form uplands, others of which have been depressed to form lowlands or topographic basins. The lineaments display strong northwest–southeast and northeast–southwest trends in the northern, northwestern, and southeastern parts of the shield, but east–west alignments are prominent in the centre, and major structural lines are more nearly meridional in the west and southwest. In all areas, however, trends other than those locally dominant can be discerned.

The role of lineaments

Within such structurally defined areas as the Kimberleys, the Isa highlands, and the Pilbara, the nature of the land surface varies according to the type and disposition of the rock outcrops. In the Kimberleys and the Mueller Plateau there are extensive outcrops of flat-lying massive sandstone that have been dissected to give rise to the striking isolated rock features known variously as plateau, mesa, and butte. Under these circumstances, local joints and bedding planes in the rocks, combined with the permeable nature of the bedrock, control the local landforms. Similar plateau forms dominate the Pilbara and Arnhem Land, though in the former region horizontally bedded or only gently warped massive ironstone formations, together with massive sandstones, give rise to prominent bluffs bordering the plateau assemblages; and in the latter karst landforms (greatly eroded by solution) are developed where limestone occurs at the surface. At the margins of the Kimberleys (in the Fitzroy region and in the Durack Range) and in the southern part of the Pilbara, in the Ophthalmia Range, dipping rock strata have been differentially eroded to form ridges and valleys. Such features are also extensively and well developed in the uplands of central Australia (the Macdonnell, James, and Kirchauff ranges), in the Isa highlands, and in the Stirling Range of the southwest. In all of these areas

it is the sandstones and quartzites that underlie the upstanding ridges, the intervening valleys being eroded in siltstones or shales; and in all these areas the pattern in plan of ridge and valley reflects the pattern of folding in the underlying rocks.

In the far southwest, the Darling Range forms an upfaulted block underlain mainly by granite but capped by laterite, a reddish, iron-rich product of weathering rock. The Gawler Block in the southeast, is complex. There are crystalline and sandstone uplands in the east, sandstone plateaus in the northeast, and, in the centre and north, the rounded Gawler Ranges built of Precambrian lava flows. Much of Eyre Peninsula is occupied by a rolling plain traversed by fixed sand dunes, but in the northwest numerous low isolated granite rocks of spectacular appearance, called inselbergs, stand above the plain. These epitomize the isolated ranges and hills widely developed in the northwest of South Australia, in the Musgrave, Everard, Birksgate, Mann, and Tompkinson ranges.

Isolated inselbergs

The lowlands between these raised blocks also display varied topography. The so-called Barkly Tableland is in reality a high plain of remarkable flatness, partly eroded in Cambrian sedimentary rocks and partly underlain by Tertiary swamp deposits. The Nullarbor Plain, is approximately coincident with the Eucla Basin. A vast area of the southwest of Western Australia is occupied by Salinaland, an extensive high plain traversed by elongate ribbons encrusted with salt, the desiccated and disrupted remnants of former river courses. The Gibson Desert consists in large part of a laterite-capped plain, but huge areas of the plains of central and northern Australia are occupied by active sand dunes, and large areas of southern South Australia and Western Australia are covered by fields of fixed dunes.

Actively developing and moving sand ridges occupy the Canning Basin (the Desert Basin), the Great Victoria Desert, the Amadeus depression and large areas of the Arunta–Sturt Complex. The dune fields extend to the east into the Great Artesian Basin, where the dunes constitute the well-known Simpson Desert. These dune deserts reflect the prevailing aridity of most of Australia, and the dune trend displays a huge swirl around the centre of the continent.

But even in these most arid areas—the area around Lake Eyre averages less than five inches of precipitation per year—rain falls from time to time, and the rivers run occasionally. Because of the scarcity of vegetational protection, and because of the common development of impermeable rock layers of various types, runoff in the arid lands tends to be rapid and to achieve dramatic and significant results. Hillslopes are scoured and washed bare of weathered debris; streams erode gullies and transport large volumes of sediment from the uplands to the plains; broad braided river channels are developed; and extensive alluvial plains are formed. It is the alluvium, carried to the lowlands by rivers and deposited on the plains, that is, in large measure, the source of the sand out of which the desert dunes are molded by the wind.

Effects of rapid water runoff

In the far southwest of the shield, and especially in the northern areas, rainfall is sufficient to support a considerable vegetation and is regular enough for streams to flow seasonally. Here the work of rivers in shaping the land surface is more obvious and widespread; the landscape consists essentially of valleys and intervening divides, the precise form of each depending on local structure. But in such areas the rate of landscape change is more rapid than in the arid zones.

Many of the landforms of the shield are inherited from the past, when different climatic conditions obtained. Remnants of laterite are widespread in many parts of Australia: the Darling Range, Salinaland, the Isa highlands and Mueller Plateau, the Darwin area, southern Eyre Peninsula. Clearly, at some time or times in the Tertiary Period, these areas had been reduced to low relief, and humid tropical climates prevailed, for laterite is at present forming only under such conditions in such areas as Southeast Asia and the Congo Basin. The disrupted former drainage system of Salinaland has already been referred to, and remnants of similar old stream networks occur in the Amadeus depression, on the Nullarbor Plain, and in the Great Victoria Desert. A large swamp formerly occupied the south of the Barkly Tableland; and Lake Woods, near Newcastle Waters, is now dry, with a bed of some 70 square miles in extent, but shorelines indicate that the lake formerly occupied some 1,100 square miles. Fossil remains also suggest wetter climates in the past in many parts of Australia, and subsequent deterioration toward aridity. But in the south the occurrence of dunes now fixed by vegetation shows that the climate there has recently become moister.

Finally, it may be mentioned that in several parts of the shield remnants of eroded surfaces, planed off and covered with hard, silicified crusts of weathered rock cut across local bedrock and are preserved either high in the relief or buried beneath later sedimentary deposits. They attest to changes in the disposition of the land surface (either base-level changes or regional warping or faulting) and also indicate that at times in the past surfaces of low relief similar to those that exist at present were widely developed. Reference has already been made to the distribution of the laterite surface. At the eastern margin of the shield there are remnants of a still older surface, of late or middle Mesozoic age, which has been warped by subsequent earth movements and now disappears beneath the sediments of the Great Artesian and similar basins. Evidence of the existence of this surface has been forthcoming from northwestern Queensland, central Australia, and South Australia.

The Flinders–Mt. Lofty ranges. Adjacent to the southeastern extremity of the shield, these uplands occupy the site of the Adelaide Geosyncline, or downwarp in the Earth's surface. These sediments were folded and faulted, principally in the early Paleozoic, though recurrently since. The Flinders Ranges are a much-eroded fold mountain belt characterized by ridge and valley forms in which sandstone ridges and bluffs are dominant. The Willochra Plain occupies an elongate intermontane basin excavated from a major upwarped structure and achieved through the erosion of some 20,000 feet of sediments. There are remnants of old land surfaces of low relief, and, in the north, extremely rugged relief developed on a much-shattered granite outcrop.

The Willochra Plain

To the south, the Mt. Lofty Ranges are a much dissected and complex horst, or ancient uplifted structural block. Bounded on both east and west by meridional or gently arcuate fault scarps, which developed initially in the Early Paleozoic but which have suffered recurrent movements since (and which indeed are still active), the ranges are surmounted in many areas by the remnants of a lateritic plain. In many other areas, such a hard capping of rock, if ever present, has been eliminated by stream erosion. Sandstones again form prominent ridges and residuals (isolated relief features), like Mt. Lofty itself; small granite outcrops give rise to boulder-strewn surfaces; and exposures of gneiss form slablike blocks known as tombstones, monk stones, or penitent rocks.

Between the Mt. Lofty and the Flinders ranges is the midnorth, a region of broad simple folds in which the sandstone ridges run for the most part north–south and in which the broad open valleys were in some instances occupied by lakes during the Tertiary Period, some 50,000,000 years ago. Similar upland areas of low relief, but with domes of crystalline rock standing above the general level, dominate the Olary Arc, which swings northeastward from the midnorth region.

The Great Artesian Basin. This platform area consists of three major basins, the Carpentaria Basin, the Eyre Basin, and the Murray Basin. Between the Carpentaria and Eyre basins the Euroka–Kynuna Plateau barely rises to reach the surface in such minute residual relief elements as Mt. Brown and Mt. Fort Bowen, in northwest Queensland. The Wilcannia threshold separates the Eyre and Murray basins, and the latter is separated from the Otway Basin and the Southern Ocean by the Padthaway Ridge. The two southern basins are entirely terrestrial, but the Carpentaria is partly inundated by the sea.

The Carpentaria plains, occupying the basin of the same name, form a narrow lowland corridor between the Isa

highlands and the Einasleigh uplands (part of the Eastern Uplands). They are drained by the Leichhardt, Flinders, and Gilbert rivers, and in the south take the form of broadly rolling plains underlain by heavy gray lime-enriched (pedocalcic) soils and known as the Rolling Downs. In the north, however, there are extensive flat depositional plains, some of them related to Pleistocene swamps, some associated with the present floodplains of the braided river systems. Standing above the plains, for example around Normanton, are considerable plateau and mesa remnants of the Tertiary laterite surface.

Similar rolling plains with laterite residuals standing above them occur in the Eyre Basin, particularly around the headwaters of the Diamantina, near Kynuna. But to the south, toward the more arid interior, the plains become flatter and are protected by a veneer of stones—the well-known stony desert with its mantle of gibber (hammada, serir, desert armour). In many parts of southwestern Queensland, northeastern South Australia, and northwestern New South Wales there are plateau and related relief remnants similar to those found in other parts of the lowlands, although these are capped and protected not by laterite but silcrete, another hard rock residue. This region is folded in places, and the subsequent dissection by erosive forces has brought about disintegration of the silcrete, which is of Middle Tertiary age and which formerly extended over vast areas of central Australia. This process provided much stony debris for the gibber plains so characteristic of much of central Australia and particularly of the Lake Eyre depression.

Lake Eyre

The catchment of Lake Eyre (*q.v.*) extends over some half million square miles of central and northern Australia. It occupies the lowest point of the Australian continent (46 feet [14 metres] below sea level) and many large river systems drain into it. These rivers drain the driest part of the continent. But no desert is rainless, and floodwaters cover the bed of Lake Eyre about twice each century, the waters deriving not only from central Australia but also from the higher rainfall areas drained by the headwaters of the Georgina, Diamantina, Thomson, Barcoo, and similar rivers. It is now clear that during the late Pleistocene period, more than 10,000 years ago, the rainfall of central Australia was heavier than it is now; as a consequence, these rivers, past and present, have brought vast quantities of sediment and salt to the interior drainage basin. As a result, there is ample source material for the Simpson Desert dunes, and many of the normally dry lake beds, including all the large ones, are encrusted with salt. Most of the very large salinas, or salt pans (Eyre, Frome, Torrens, Gregory, Blanche), are, at least in part, of structural origin, having been formed by downfaulted blocks. Torrens and Gregory are surfaced mainly by gypsum, but the remainder carry a crust of sodium chloride, common salt. Around the major salinas there are extensive alluvial plains.

Under the prevailing arid conditions, fine dust is winnowed from the surface sediments and can be carried high into the air in dust storms. Some is carried long distances, even reaching New Zealand from time to time. The sand of the alluvium is molded into dune ridges.

Sand dunes also occupy large areas of the Murray Basin. They are, by contrast, fixed (or "fossil") dunes, which developed at some time in the recent past and have since been stabilized by higher rainfall conditions. In the east of the basin, near the foothills of the Eastern Uplands, there is evidence of former higher rainfalls in the numerous abandoned river channels of the Riverina. But the western Murray plains are a stony as well as a climatic desert. The plains are underlain by Miocene limestones and, in many areas, by calcrete, a calcareous soil accumulation. There are instances of water-dissolved sinkholes and enclosed depressions, and lack of surface drainage characteristic of this type of topography. Only the Murray, which originates outside the area in a different environment, crosses the basin, flowing in a narrow trench in its lower reaches.

In the east of this region there are extensive alluvial plains associated with major tributaries of the Murray. One feature of interest is the diversion of the Murray, near Echuca, by a rising structural block bounded by fault zones and known as the Cadell Fault Block.

The Eastern Uplands. This complex series of high ridges, high plains, plateau, and basins extends from Cape York Peninsula in the north to Bass Strait in the south, with a southerly extension into Tasmania and one extending westward into western Victoria. The uplands are the eroded remnants of an ancient mountain range recently rejuvenated by block faulting. They occupy the site of the Tasman geosynclinal belt, the sediments of which were folded and faulted in late Paleozoic times. Granite batholiths were intruded into this region, and during the Cenozoic Era lavas appeared extensively in areas as far apart as north Queensland and Tasmania. Characteristic features associated with this process were lava fields, with stony rises, soil-filled depressions, and lava caves. Extinct cones and craters survive in southeastern Queensland, in the Monaro district of New South Wales, and in western Victoria.

Volcanic landforms

The landforms, in considerable measure, reflect these various geological events. Uplifted structural blocks, many of them trending north to south, are common in some areas, while straight river courses reflect the control exercised by fault zones. Ridge and valley forms, as found in the Grampians of Victoria, reflect the differential erosion of broken and folded rock strata. On the exposed granitic batholiths, massive domes or clusters of boulders are common. The lava plains and plateaus display stony rises, shallow alluvial depressions, and volcanic vents and plugs of various types and ages.

Other features reflect the erosional history of the region. Wide areas of the upland had been reduced to a uniform low relief by the time of the later Mesozoic Era, about 100,000,000 years ago, and many remnants of this ancient surface, exhumed by erosive action from beneath a Later Cretaceous cover, survive in the landscape, notably in north Queensland. The Middle Tertiary leaching of rocks by weathering in humid climates with the formation of iron-rich residuals (laterization) also affected the uplands, from northern Queensland to Tasmania.

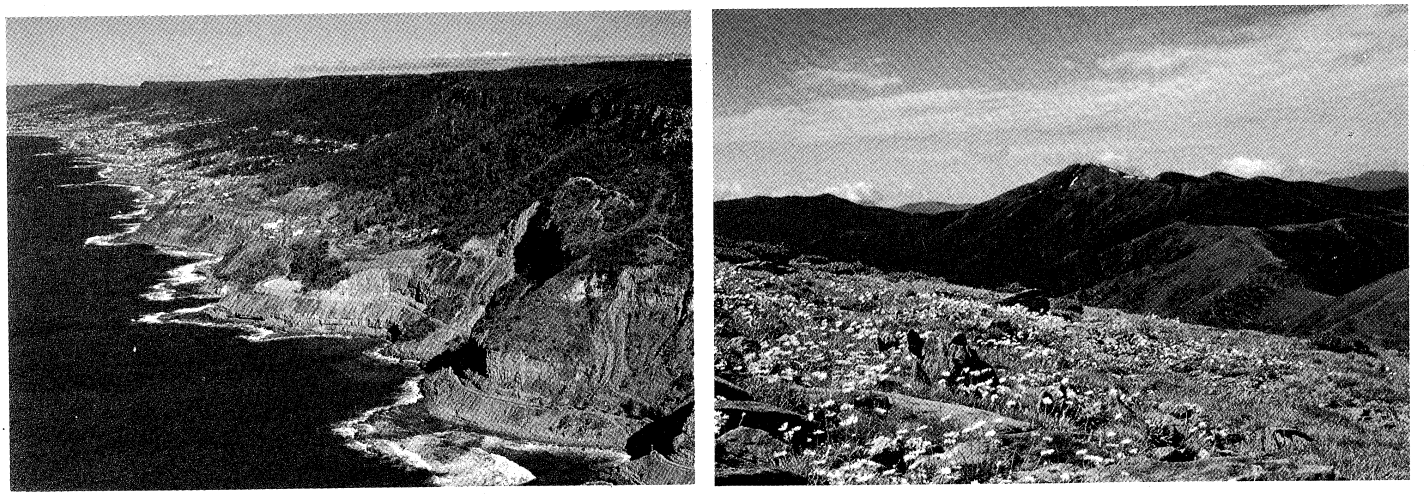
Lastly, during the Pleistocene Epoch, some 2,000,000 years ago, small glaciers developed in the Kosciusko area of New South Wales and the central plateau of Tasmania. Small, ice-scoured hollows and small moraines (ridges of glacial debris) attest to these events, while over rather wider areas frost shattering of rocks and resulting down-slope flowage of soil (solifluction) have helped shape the surface. No snow normally survives through summer in either of these areas now, but in winter the snowfields of the Kosciusko area alone are more extensive than those of all of Switzerland.

The Great Barrier Reef influence

The Great Barrier Reef (*q.v.*) is related in important respects to the Eastern Uplands. Lying off the Queensland coast, this great system of coral reefs and atolls owes its origin in part to Pleistocene changes in sea level but in most part to long-continued subsidence, related to faulting, of the offshore region. This slow subsidence has enabled a great thickness of coral to develop, and it is on this basement that the present reefs and coral atolls have grown in the clear warm waters of the Coral Sea.

The coming of man. Though neither aboriginal man nor the later European settlers have been long in Australia, they have already achieved widespread, and in most ways deleterious, effects on the landscape. European man in particular has been responsible for initially minor, but later significant and widespread, changes, notably considerable soil erosion. The clearing of vegetation for agricultural purposes, overgrazing, the introduction of exotic plants and animals, the making of tracks and roads, even the clearing of stones from paddocks—all have rendered the land surface more susceptible to soil erosion. Man has set in train his own great cycle of erosion, similar to that which beset many parts of western Europe in the 18th century and which has assailed many parts of the American West since late in the 19th century.

Climate. Australia is the arid continent. Over two-thirds of its landmass rainfall per annum averages less than 20 inches and over one-third of it is less than 10



Eastern Highlands (Great Dividing Range).
(Left) South of Sydney, where they shape the coastline near the towns of Bulli and Wollongong in left background; and (right) at Mt. Feathertop, inland Victoria.
(Left) G.R. Roberts, (right) John R. Brownlie—Photo Researchers

inches. Only just over 10 percent receives more than 40 inches per annum. As has been noted, in winter the snowfields of Tasmania and the Kosciusko area can be far more extensive than those of Switzerland, but on the whole Australia is a very hot country with a high incidence of heat waves, in consequence of which evaporation losses are high and the effectiveness of the rainfall received is reduced. In addition, the severity of climate, the predominance of the outdoors in the minds and lives of many, and the national importance of agricultural and pastoral pursuits, all make Australians perhaps more climate-conscious than most. In no country of comparable development do climate, the weather map, and the latest forecast loom so large in the lives and conversation of the common man.

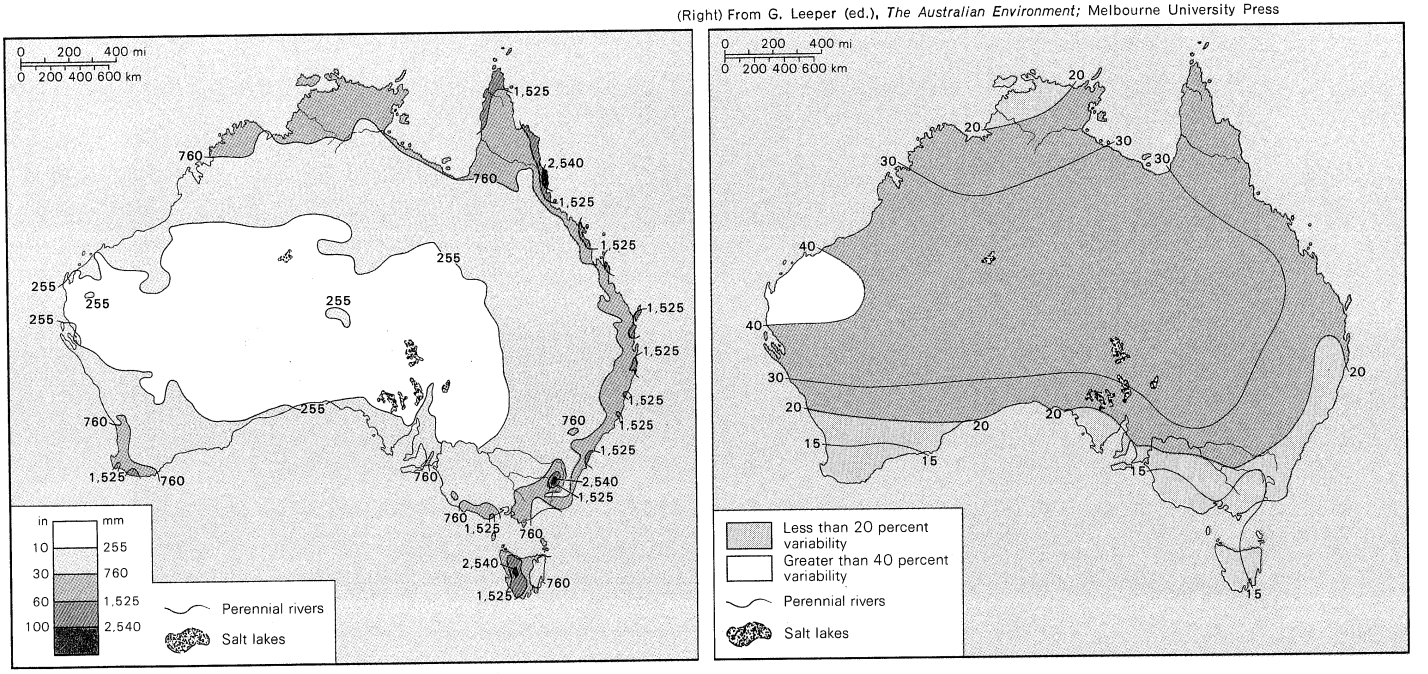
Major influences on climate

The principal features of Australia's climate stem from its position, shape, and size. Australia is a compact tropical and near-tropical continent located between latitudes 10°41′ S and 43°39′ S. No major arms or embayments of the sea penetrate far into the landmass. The only extensive uplands occur near the east coast, and even they are not, by world standards, very high.

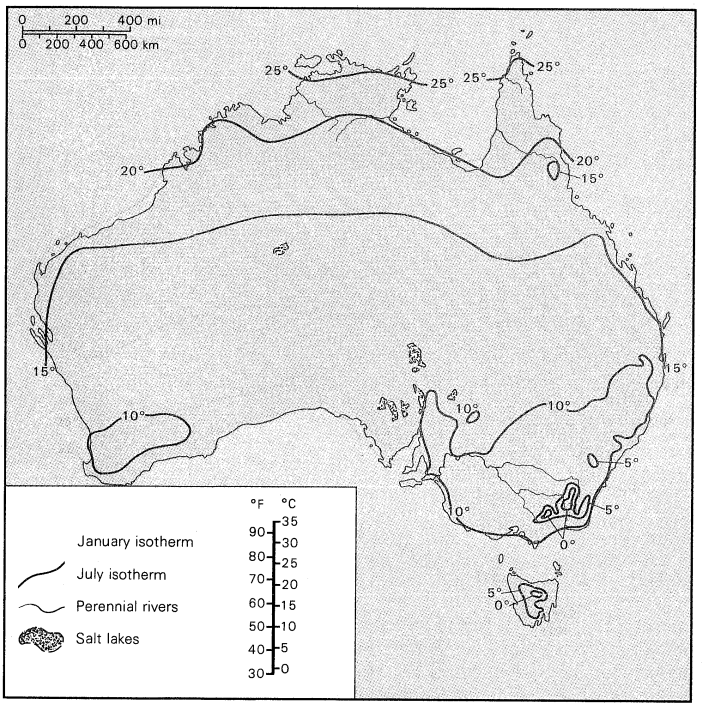
In summer, when the sun is directly overhead in northern Australia, temperatures are extremely high. The sea exerts little moderating influence, and the uplands are not sufficiently extensive or high to have more than local effects. Because of the lack of cloud over most of the interior, however, while maximum temperatures commonly soar over the 100° F mark, there is considerable radiation loss at night, and daily temperature ranges are quite high. But, on the whole, high temperatures dominate the Australian summers in all but Tasmania. Heat waves are common, and though the highest amounts of radiation are received in northern South Australia, the highest temperatures and most extensive heat waves are recorded in the northwest of Western Australia. Marble Bar, for instance, has recorded a maximum temperature of 100° F or more on 162 consecutive days. Temperatures in winter remain moderate except in the uplands of Tasmania and southeastern Australia, where snow is common. Night frosts are common in winter throughout southern Australia and in the interior.

Because of its latitudinal position, Australia comes under the influence of the southeast trade winds in the north and the westerlies in the south. Northern Australia is affected by a northerly monsoon, partly because of

(Right) From G. Leeper (ed.), *The Australian Environment;* Melbourne University Press



Rainfall patterns in Australia.
(Left) Average annual rainfall in Australia (measurement figures on the map are in millimetres). (Right) Percentage mean variability from annual mean rainfall.



Average temperatures, in degrees Celsius, for January and July in Australia.
By courtesy of the Department of National Development, Canberra

the latitude and the seasonal migration of planetary wind zones and partly because of the summer heating of the continent and the indrawing of surface winds. The monsoon brings summer rains to the northern coastal area and extends inland for variable distances. These summer rains are all the more important because most of northern Australia is in the sheltered rain shadow of the Eastern Uplands so far as the southeast trades, which are dominant in winter, are concerned. The trades, forced to rise by the uplands, bring heavy rains to the Pacific coast of Queensland and of northern New South Wales. These areas, which are also affected by tropical cyclones, receive the heaviest rains of any part of Australia, and, within this coastal fringe, the north Queensland area around Cairns is the wettest.

Southern Australia receives winter rains from depressions associated with the west-wind zone. Again there are local topographic controls, with uplands receiving more than the adjacent plains. Parts of the southern Mt. Lofty Ranges, in South Australia, average over 40 inches of rainfall per annum, but Adelaide, to the west, averages only 21 inches, while the Murray plains, in the rain shadows of the ranges, receive 15 inches or less rainfall annually.

The dry interior

In the great mass of the interior of Australia, rainfall averages less than 20 inches per annum, and over vast areas the total is less than 10 inches; the Lake Eyre region averages less than 5 inches. Rainfall in these areas is unreliable and capricious, with long droughts broken by damaging rains and floods. Over Australia as a whole, rainfall is indeed extremely variable. Only in the far north, around Darwin, in the southwest of Western Australia, in southern South Australia and Victoria, in Tasmania, and in eastern New South Wales is the recorded annual rainfall not more than 10 percent above or below the long-term average in different years.

Drainage. Permanently flowing rivers are found only in eastern Australia and in Tasmania. The major exception is the Murray, a stream that rises in the Kosciusko region in the Eastern Uplands and is fed by melting snows. As a result, it acquires a volume sufficient to survive the passage across the arid and semi-arid plains that bear its name and to reach the Southern Ocean southeast of Adelaide. All other rivers in Australia are seasonal or intermittent in their flow, and those of the arid interior are episodic.

Many areas—notably the Nullarbor Plain, underlain by limestone, and the sand ridge deserts—are without surface drainage. A map of Australia can be misleading;

By courtesy of (bottom right) Australian News and Information Bureau; photographs (top left, top right, bottom left) Photographic Library of Australia



Australian drainage patterns.
(Top left) Tweed River near Murwillumbah, New South Wales. (Top right) Salt lakes and "islands," Lake Amadeus, central Australia. (Bottom left) Fitzroy River system near Rockhampton, Queensland. (Bottom right) The Ord River, in the Kimberly district of Western Australia, is an important source of irrigation.

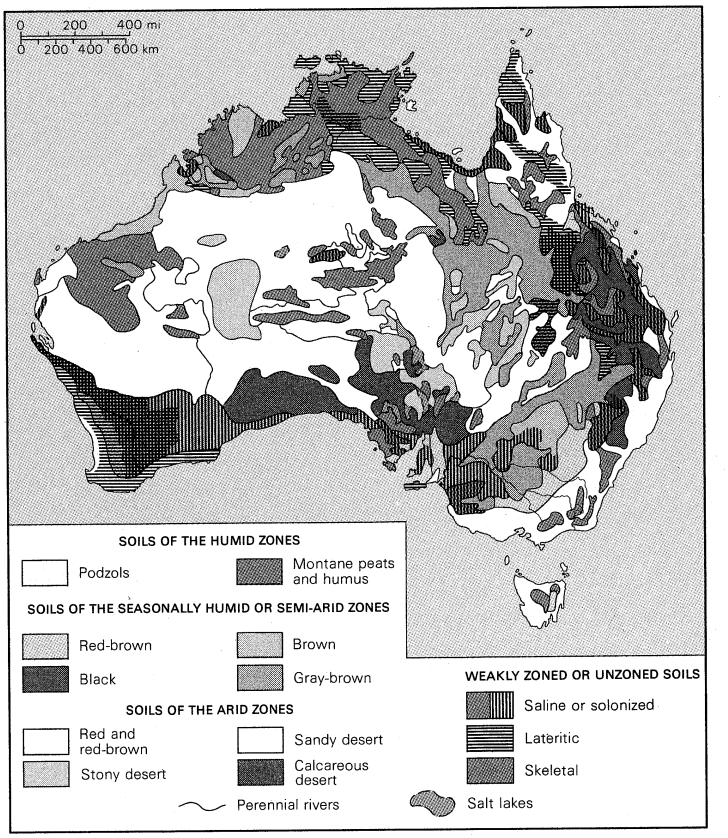
though many "lakes" are depicted in the interiors, the fact is that many of them are now salt lakes that contain no water for years on end (see also below, *Natural resources: water resources*).

Soils. In broad view, the continental pattern of soils is closely related to climatic factors; mineral or skeletal soils exist over much of arid Australia, with virtually no organic content and little development to any depth. They may consist merely of a rough mantle of weathered rock. Gypsum is present in many of the desert loams and arid red earths. The soils of the semi-arid regions (where annual rainfall is from 8 to 15 inches) are also alkaline, with gypsum and lime a common feature. The organic content of the soils is again low in the solonized (salt-enriched) brown soils and the gray and brown soils of heavy texture that are common in these areas.

In both the arid and semi-arid regions gilgai—patterns of swells and depressions due to the alternate swelling and contraction following wetting and drying of clay soils—have developed. They are especially well represented in areas of seasonal rainfall. In areas with 15 to 25 inches of annual rainfall, black earths, brown soils, and red-brown earths are the most common soils. In the wetter areas the leaching out of minerals is a prominent feature of the soils. Podsols—sandy, with much humus at the surface and acid throughout—are the characteristic soil types. In the alpine regions humus soils—surface peats over a mineral—are noteworthy.

Local soil variations

Superimposed on these broad, climatically determined, soil patterns are local variations due to topography, groundwater conditions, and parent materials. For example, red soils of one kind (krasnozems) are developed on the basalt outcrops so common in eastern Australia, and those of different composition (terra rossas and rendzinas) on calcareous bedrock. In addition, laterite and silcrete originated in remote geological times, when conditions were very different from those of today. Laterite is represented in every state, including Tasmania, though it is forming nowhere in Australia at the present time, while silcrete is restricted to arid Australia and parts of subhumid Western Australia, South Australia, and Queensland. (C.R.T.)



Types of soils in Australia.
By courtesy of the Department of National Development, Canberra

PLANT AND ANIMAL LIFE

Plants and vegetation. *Overall characteristics.* Less than 200 years ago, Australian vegetation was still in its primal stage; the older plants in its present woodlands and forests began their lives before the continent was invaded by civilized man. So close is the continent's prehistory that many a eucalypt still bears the great scar of a canoe or shield cut from its bark by Aborigines. Others can be found bearing blazes and inscriptions dating back to the first years of exploration and settlement.

In the short history of modern Australia, vast changes have been wrought in the continent's vegetation. Agricultural expansion stripped whole regions, substituting introduced crops, pastures, and plantations, while uncleared areas near the new settlements, ranging from the densest forest to the sparsest woodland, were cut through for timber. Enormous central and northern areas, too arid for agriculture or too remote for timber getting, were stocked with millions of sheep and cattle. In addition, many weeds were introduced, along with rabbits, other herbivorous vermin, and frequent bush fires. The native vegetation is still in that moment of destructive upheaval that marks the passing from aboriginal man to technological man and domestic flocks. By the 1970s, reserves and wilderness areas held some native vegetation as national heritage, although it remains doubtful if these are adequate or sufficiently protected.

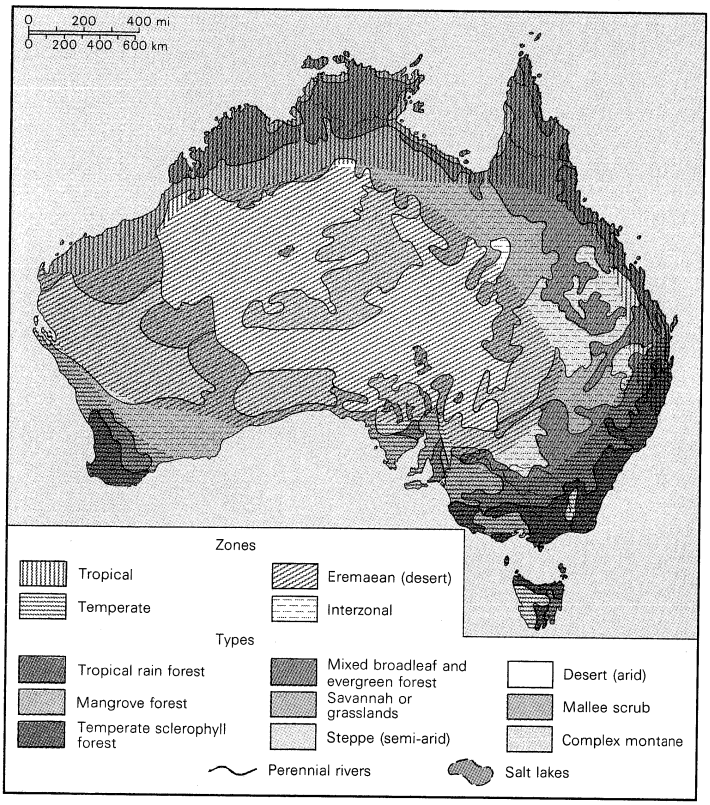
Plant life. Australian federal and state governments maintain institutions for the scientific collection and study of the kinds (or taxa) of plants. Cumulative knowledge of Australia's flora stems mainly from these endeavours and is partly available in handbooks (Floras) listing species, together with appropriate "keys" for their recognition. G. Bentham's *Flora Australiensis* (1863–78), based on 19th-century collections sent to Europe, remains—although much outdated—the only comprehensive survey of Australian flora. More up-to-date information, scattered in the botanical literature, is not easily accessible except to professionals.

Plant origins

Australia's phanerogamic (seed plant) flora is estimated at 15,000 to 20,000 species, believed to be a blend of elements from various original sources and the outcome of a long and complicated history. One contribution to the present plant patterns is thought to have originated, in the remote past, from some southern landmass which then linked all the present southern continents. There is evidence to support this view: Australia shares with South Africa, Madagascar, New Zealand, and South America such plants as the southern beech (*Nothofagus*) and characteristic genera in other families, which constitute an Antarctic element.

Australia also shares many more kinds of plants with its northern neighbours. Also, typically Australian genera such as *Callitris* (native pine), *Banksia*, and *Eucalyptus*, extend into the Indo-Malaysian area, some as far as the Celebes. This evidence has given rise to the idea of an Indo-Malaysian element in the Australian plant life, in a process that involved some two-way exchanges. These, and other links with plant life elsewhere, have prompted various hypotheses invoking the development of species, migrations, and extinctions throughout geological time, along with changes in the disposition of land, sea, and climate.

That characteristic part of Australia's plant life that is not much shared with other lands, together with those specialized characteristics that apparently originated in the continent long ago, form what has been designated as an Australian (or autochthonous) element. It includes many of the plants lending character to typical Australian vegetation scenery. It also shows a marked tendency to sclerophylly (formation of hard leaves) and wide-ranging adaptation in many genera, extending their species throughout the whole range of the continent's environmental habitats. Most obvious to the visitor are the gum trees (*Eucalyptus*), which are represented by over 600 species, ranging in size from diminutive mallees, smaller than a man, to forest giants matching in bulk



Zones and types of vegetation in Australia.

and height the world's biggest plants. Their habitat is similarly varied, ranging from rain forest to snowfield to hot desert fringe. *Acacia* has undergone similar adaptive radiation; its species range from mulga and myall—the dominant trees of vast areas—to small leafless blades at ground level. The banksias and hakeas (Proteaceae), the grass trees and blackboys (Xanthorrhoeaceae), and the kangaroo-paws (Haemodoraceae) are examples of the many other characteristically Australian plants.

Vegetation. Vegetation, as opposed to plant life, implies the structure and communal relations of the landscape's plant cover, whether it be forest, grassland, or marsh. There is no standard, or worldwide, classification system (such as exists for describing flora) for this aspect of the environment. Initial attempts to apply European and American classification ideas and methods to Australia were not particularly satisfactory, as a result of the peculiarities of the continent's vegetation and environment. For example, climatic control of local vegetation zones was often found insufficient to support vegetation analysis; on the contrary, soil patterns and geological history quite override climatic control in many localities. Similarly, structural descriptive schemes useful for Northern Hemisphere coniferous and deciduous vegetation proved quite inappropriate when confronted by the great variety of evergreen vegetation—notably mallees and scrubs—found in Australia. By the 1970s, mapping of Australian vegetation was based largely on factual descriptive features, and by this means such comprehensive and detailed accounts and maps as those contained in the *Atlas of Australian Resources* and *The Australian Environment* (see *Bibliography*) had been produced. Scientific vegetation analysis, as opposed to mapping, is also well advanced in Australia, contributing modern ideas equal to those anywhere else in the world.

Classification difficulties

Modern Australian plant life is distributed in three main zones—the Tropical, the Temperate, and the Eremian—which reflect overall climatic conditions. The first zone, arced east and west across the north margin of the continent and extending halfway down the eastern seaboard, has a mainly dry monsoonal climate, with very wet patches. The second, with a cool to warm temperate to subtropical climate and mostly winter or nonseasonal rainfall, is arced across the southern margin, embracing Tasmania and extending up the eastern seaboard to overlap slightly with the Tropical Zone. The third zone covers the whole of central Australia, through to the central west coast; its climatic characteristic is aridity.

The major structural units comprising this geographical distribution are rain forest, sclerophyll forest (dominated by hard-leaved plants such as eucalypts), woodland, scrub, savanna, and grassland forms, each with a range of subforms. The bulk of the Tropical Zone carries mixed deciduous woodland and sclerophyll low-tree savanna, with areas of tussock grassland, coastal mangrove complexes, and tropical rain forest. There is a high occurrence of exotic vegetation, particularly in the northeastern part of Cape York Peninsula, in Queensland, and a strong Indo-Malaysian influence occurs throughout the entire zone. The rain forests, with large trees with stem buttresses, and multiple vegetation layers with interlaced canopies, lianas and epiphytes growing parasitically on the trees, qualify technically as jungles.

The Temperate Zone

The Temperate Zone is characterized by dry and wet sclerophyll forests, temperate mixed woodlands, savannah woodlands, mallees, and scrubs, with areas of Alpine vegetational complexes, temperate rain forest, and sclerophyll heath. Native plants form a much higher proportion of the vegetation cover, much of which is typically and recognizably Australian. Within this zone the southwest corner of Western Australia is outstanding, both for the high proportion of Australian plants and for the richness of the plant life, while the vegetation of Tasmania is notable for its southern beech forests and for its links with New Zealand and South America. In marked contrast to the tropical rain forests, the predominant trees throughout the majority of the Temperate Zone communities are either *Eucalyptus* or *Acacia*. Much of the Temperate Zone vegetation has been cleared for agricultural purposes, leaving only the vegetation communities of infertile or inaccessible localities.

The vegetation of the Eremian Zone ranges from barely vegetated desert sand hills through a variety of semi-arid shrub savannas, shrub steppes, semi-arid tussock grassland, and sclerophyll hummock grasslands. Many shrubs have adapted themselves similarly to the arid conditions so that in their vegetative state representatives of numerous families look alike. *Acacia*, *Eremophila*, and *Casuarina* are examples of genera that tend to displace *Eucalyptus* as the dominant tree shrub. Much of this vegetation is badly degraded. (R.T.La.)

Animal life. Human activities had a modifying and generally destructive influence on Australian animal life, or fauna. The arrival, about 30,000 years ago, of the Aborigines was in this respect less significant than European occupation, which has had profound effects over a period of only two centuries. The Aborigines brought only the dingo, a placental carnivore that must have affected the native marsupials. The Aboriginal hunters and food gatherers lived in ecological balance with their environment, and their demands were not large enough nor their technology sufficiently developed to upset it. There is even less scientific evidence to link them with the extinction of the large Pleistocene mammals than there is in other continents. All this changed with the advent of the Europeans: their cats, rats, foxes, rabbits, cattle, and sheep and their technology modified and largely destroyed important habitats of the Australian fauna. The inhabited coastal region (and even part of the arid inland regions) of the continent now contain a very much modified fauna compared with the indigenous life of only 200 years ago, which cannot now be reconstructed.

Destructive effects of Europeans

The Australian fauna is markedly different from that of the nearest land areas, the islands of Indonesia. A 19th-century biologist, Alfred Russel Wallace, designated an Australian zoogeographic region in 1876 and drew the boundary separating it from the Oriental region of Southeast Asia between Bali and Lombok and between Borneo and Celebes. This division, which became known as Wallace's line, is still recognized in modern biogeography as the boundary of a transitional zone, across which animals spread according to their ability to cross

narrow seaways. These passages were much narrower during the Pleistocene glacial periods of the last 600,000 years, when so much oceanic water was frozen at the poles that the sea level fell to or beyond the edges of the continental shelf that now fringes Australia. At such times, the continent had land connections with New Guinea and Tasmania, while remaining separated from the Indonesian Archipelago during all of Cenozoic time (*i.e.*, for the last 50,000,000 years). The history and origin of the distinctive Australian fauna has led to much controversial speculation and searching for relationships among the southern continents.

A published count (Keast *et al.*; see *Bibliography*) indicates the existence of 108 placental mammal species, 119 marsupials (125 according to a later count), 2 monotremes (the spiny anteater and the platypus), 520 birds, about 380 reptiles, 122 frogs, and 180 freshwater fish. There are also more than 54,000 species of insects (Mackerras, 1970; see *Bibliography*) and about 750 species of mollusks. The placental mammals (apart from those introduced by man) belong to groups that can swim (seals), fly (bats), or drift on logs (rodents).

The marsupials are considered distinctive Australian animals because they are more abundant and more differentiated on the Australian continent than in America, where few, and in Europe, where none, has survived. They are not closely related to any fossil or living species found elsewhere. Their history in Australia has been traced back to fossil remains not more than 25,000,000 years old; that is, for less than one-half of the time of their presumed existence in isolation in Australia. There isolation from placental predators and competitors gave them time to differentiate. The kangaroos, herbivores of the open woodlands and grasslands (occupying the habitat of horses or antelopes elsewhere); the tree-dwelling cuscuses (*Phalanger*) and flying phalangers or gliders (*Petaurus*), resembling flying squirrels; the "native bear" or koala (*Phascolarctos*), which is specialized to live on eucalypt leaves; the burrowing wombat (*Vombatus*); the native cats (*Dasyurus*); the marsupial mice (*Sminthopsis* and *Antechinus*); the numbat or banded anteater (*Myrmecobius*); the dog-sized Tasmanian wolf (*Thylacinus*); and, finally, the marsupial mole (*Notoryctes*) are examples of the adaptive differentiation of these mammals.

The role of marsupials

The only surviving monotremes (egg-laying primitive mammals) are the duck-billed platypus (*Ornithorhynchus anatinus*) and the echidna or spiny anteater (*Tachyglossus aculeatus*). The echidna is common and widely distributed; the platypus, now completely protected by law, is common in streams and lakes in eastern Australia.

There is much concern among nature lovers and conservationists in Australia about the obvious and increasing losses suffered by the marsupial fauna. These are due both to increases in rangeland agriculture and to the largely uncontrolled exploitation of the kangaroos. According to one observer,

> The lands of arid Australia have been grazed by sheep, cattle and rabbits for periods varying from a few years in the most recently occupied areas to a little over a century in the oldest settled parts, a period short in comparison with most other arid lands of the world. In that short period we have caused far more degeneration of land resources than the aborigines caused in twenty to thirty millennia. Australia is fortunate indeed that the period has been so short. (From R.O. Slatyer and R.A. Perry [eds.], *Arid Lands of Australia;* Australian National University Press, 1969).

To this progressive destruction of habitat has to be added the effect of loosely controlled, or uncontrolled, shooting of kangaroos for meat and fur. The export value of kangaroo products doubled from 1960–65 to reach over A$3,000,000 and the kangaroo harvest in Queensland alone exceeded A$1,000,000 in 1965.

> The crop must be controlled or the resource will be brought so low that it will be lost and these wonderful animals be reduced, like the American bison, to a small remnant confined to reserves. (From W.D.L. Ride, *A Guide to the Native Mammals of Australia*, 1970.)

Discussions of the feasibility of commercial kangaroo farming was just starting in the 1970s, although popular opinion had forced the government to take measures for the strict protection of the koalas, which were in danger of extermination, some 50 years earlier. Many small marsupials are also threatened, and some have been eliminated in their original habitats by marauding cats and foxes. Conservationists hope that the increasing number of national parks and reserves will assist in the preservation of the native fauna, though supervision in the vast, almost uninhabited inland areas, where natural conditions remain least disturbed, is a serious problem.

Australian birds

Dangers similar to those threatening the marsupials also affect the birds of Australia, many of which are also native to the continent and, therefore, of great scientific interest. The best known typically Australian birds are the flightless emu (*Dromaeus*), the mallee fowl (*Leipoa*, which builds a mound-shaped nest for hatching its eggs and actively controls the mound's temperature), the spectacularly abundant cockatoos (*Cacatuinae*), the lyrebirds (*Menura*), fairy wrens (*Malurus*), honey-eaters (Meliphagidae), Australian magpies (Cracticidae), and bowerbirds (Ptilonorhynchidae).

Australia has one freshwater crocodile (*Crocodylus johnstoni*), which lives in the tropical north but is also represented in rock engravings made by Aborigines in South Australia. There are ten freshwater tortoises belonging to a family (Chelyidae) that is also known from South America. Lizards include geckos, skinks, legless lizards, and goannas, or monitor lizards (*Varanus*), which have relatives in Southeast Asia and Africa. There are many poisonous snakes, of which the taipan (*Oxyuranus scutellatus*), the tiger snake (*Notechis scutatus*), the death adder (*Acantophis antarcticus*), and the brown snake (*Pseudonaja textilis*) are most dangerous to man.

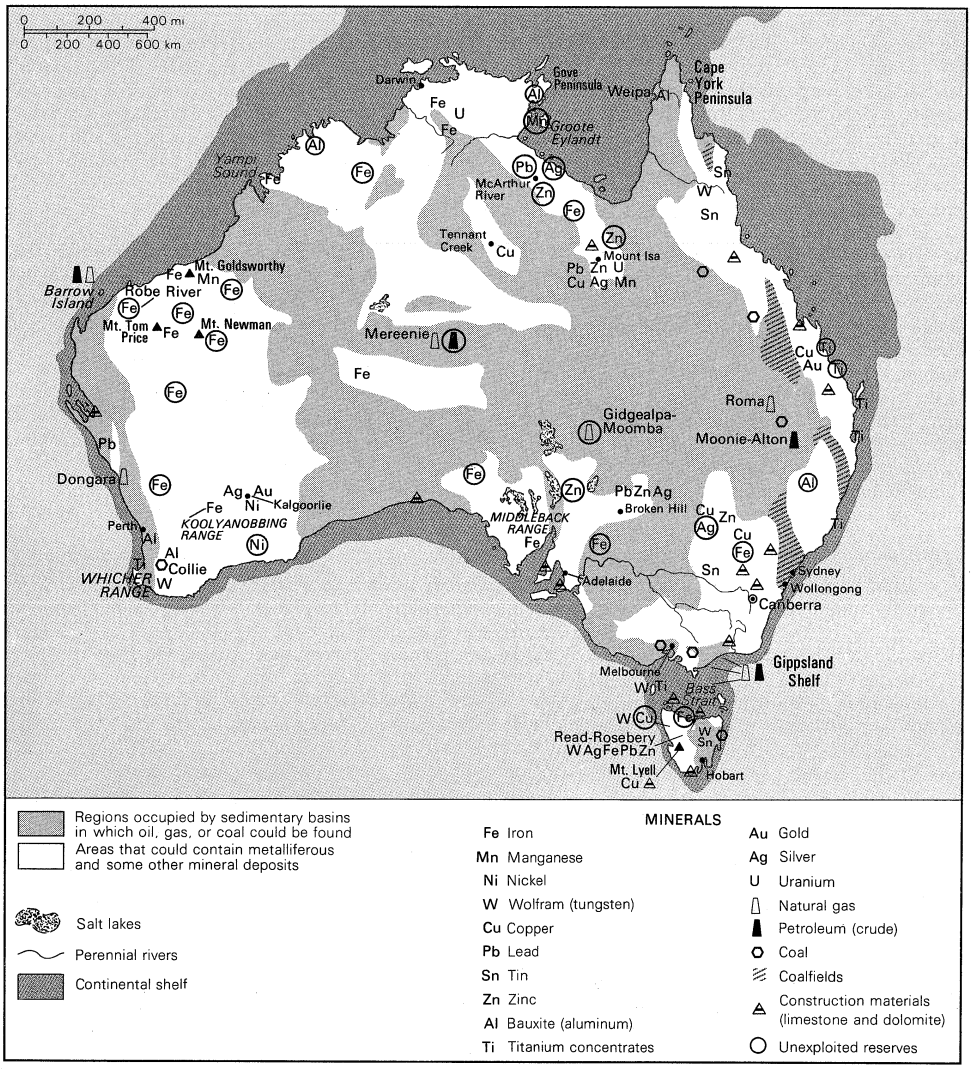
The isolation and predominant aridity of Australia makes its freshwater fish fauna an interesting object for study. The Queensland lungfish (*Neoceratodus*) has lived in Australia without change for millions of years and is very much like its European ancestors of 200,000,000 years ago; it is very different from African and South American lungfishes. European fishes introduced recently into some Australian rivers and streams seem to be a danger to the survival of native species.

Insects have been as successful in Australia as in other continents, showing many adaptations to hot and dry conditions, particularly by adjusting the timing of drought-resistant developmental stages. Leaf-eating insects, including locusts, may be plagues in pasture regions. Other insects attack timber, and the destructive, as well as the constructive, activities of the widespread termites are well known. Bloodsucking insects attack cattle and sheep, and some are disease carriers. The larvae of blowflies (Calliphoridae) attack the living tissues of sheep and continue to cause sheep farm losses of millions of dollars. (M.F.G.)

NATURAL RESOURCES

Mineral resources. The first settlers, who arrived in Australia at the beginning of the Industrial Revolution in Europe, soon reported the existence of coal near Sydney and elsewhere. Discoveries of alluvial gold west of Sydney, later in Victoria, and finally in Western Australia led to great movements of population across the continent and to extensive immigration. Prospectors discovered many rich mining areas throughout the continent, particularly in the west, the area around Adelaide, Victoria, Tasmania, western and central New South Wales, northern New England, coastal and western Queensland, and the Northern Territory. More recently, in 1949, Australia began a period of almost continuous mineral discoveries. This period continued into the 1970s, the rate of discoveries and, more importantly, of mineral production, rapidly increasing (see AUSTRALIA, COMMONWEALTH OF for a consideration of the role of minerals in the national economy).

Mineral fuels. The coal resources of eastern Australia were attracting much interest from overseas buyers by the 1970s, and new mining ventures ensued. Black coal is found in the Permian sediments—some 250,000,000 years old—in the Sydney Basin and in Queensland.



Mineral resources of Australia.
By courtesy of the Department of National Development, Canberra

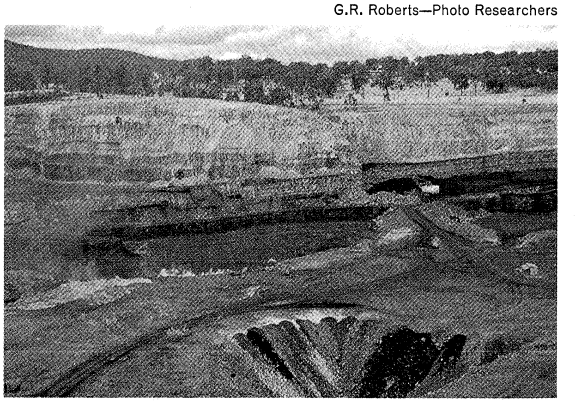
Western Australia has a small Permian coal basin at Collie, and South Australia has a Triassic coalfield—dating from 200,000,000 years ago—at Leigh Creek. Large brown-coal deposits are exploited for electricity production at Yallourn and Morwell in Victoria, and west of Melbourne. The total coal reserves of Australia were estimated in 1971 at 24,000,000,000 tons of which 7,000,000,000 tons in New South Wales and Queensland are described as coking coal.

Commercial gas fields

The fuel supply situation in Australia changed fundamentally with the discovery of commercial gas fields and the opening of pipelines to capital cities in 1969. The Roma area of Queensland supplies gas to Brisbane; the Gippsland offshore area supplies Melbourne; gas from the Gidgealpa–Moomba area is brought to Adelaide and may be supplied to Sydney; while the Dongara field, in the Perth Basin, is being connected by pipeline with Perth. Over 8,000,000,000 cubic feet of gas were used in Australia in 1969, the first year of major production, and over 110,000,000,000 cubic feet in 1972. A gas reserve of over 500,000,000,000 cubic feet in the Ordovician sandstone (some 450,000,000 years old) of the Mereenie field, 150 miles west of Alice Springs, remains unused, pending decisions on pipeline construction and other matters. Recoverable reserves of natural gas in Australia were estimated at 37,750,000,000,000 cubic feet in the early 1970s.

The first commercial oil discoveries in Australia were made, in 1961 and 1964, respectively, at Moonie and Alton, in the Surat Basin. Cumulative production by the end of the decade amounted to over 15,000,000 barrels, although production was declining in these two fields in the early years of the 1970s. The Barrow Island oil field, located off the coast of northwestern Australia, by the early 1970s had an average daily production rate of 43,000 barrels from 348 producing wells, with estimated

G.R. Roberts—Photo Researchers



Open-cast coal mine, Muswellbrook, Hunter Valley.

reserves of 200,000,000 barrels. Crude oil production from the offshore fields in Bass Strait off the Gippsland coast of Victoria began in 1969, and within a year the first two fields were producing at an average daily rate of 105,000 barrels, with an increase to 330,000 barrels scheduled. The recoverable oil reserves of the known

structures were estimated at 1,800,000,000 barrels in the early 1970s.

The output of indigenous crude oil in Australia in 1969 met only 10 percent of the total Australian demand; but by the early 1970s indigenous production was meeting more than 60 percent of the continent's crude oil requirements. In addition, exploration drilling was in progress at the rate of about 100 wells per year.

Uranium. Post-World War II exploration efforts, linked to the world demands of a growing atomic energy industry, led to the discovery and opening of uranium mines in northwest Queensland, South Australia, and the Northern Territory. The completion of export contracts concluded at that period had brought active mining to a standstill by the early 1970s, but the prospect of new overseas markets led to a revival of exploration and to new discoveries.

Major uranium deposits

In northwest Queensland, the Mary Kathleen uraninite deposit is situated between Mt. Isa and Cloncurry. It has reserves of over 3,000,000 tons of rich ore, ready for open-cut mining. The Radium Hill mine, located near Olary in South Australia, was based on a deposit discovered in 1906 and worked from 1954–60, producing about A$35,000,000 worth of uranium in the form of davidite. Uraninite deposits have long been known in the South Alligator River area, and deposits at Narbalek, on the lower reaches of this river in Arnhem Land, 150 miles east of Darwin, have reserves estimated at 10,000 tons. The announcement of the first large discovery of a sedimentary uranium deposit in Australia was made during the same period. This deposit lies, at a shallow depth, under the Lake Frome plains in South Australia, and is clearly derived from the uranium-bearing Mt. Painter Complex in the adjoining northern Flinders Ranges.

Iron ore. Iron ore is produced in Australia for iron and steel production and for export: the physical and chemical properties of the iron ores control the value of the ore for modern technological processes. These processes facilitate the use of otherwise valueless fine-grained ore, although strict control of phosphorus and nickel content is required; upgrading (known technically as beneficiation) of low-grade jaspilite-hematite ores could add thousands of million of tons to the available reserves.

After a long period of concern about the continued availability of reserves for the Australian steel industry from the two known iron-ore-producing areas, an export ban, which had been imposed in 1938, was lifted in 1960. Indications that major additional resources existed proved correct, and with the stimulus of large contracts for the export of iron ore (with Japan as the main destination), reserves of many thousands of tons of limonite and hematite ores were developed.

The old ore-producing areas are in the Middleback Range on Eyre Peninsula in South Australia, and in Yampi Sound and Koolyanobbing Range in Western Australia. The Middleback Range contain 12 ore bodies of banded hematite-jaspilite iron formation. High-grade ore production had exceeded 100,000,000 tons by the 1970s and reserves were estimated at 200,000,000 tons. The Yampi Sound deposits, 85 miles north of Derby, have reserves of about 65,000,000 tons.

New iron ore discoveries

The new discoveries are on a vast scale: the Mt. Goldsworthy deposit, 60 miles east of Port Hedland, contains 65,000,000 tons of hematite lode. The newly developed Hamersley iron province contains thousands of millions of tons of ore in iron formations. The largest high-grade ore deposits are Mt. Tom Price, with reserves of about 500,000,000 tons of ore; Mt. Whaleback; and Mt. Newman. Other ores occur in vast quantities in the Robe River area; reserves are estimated at 5,000,000,000 tons, with an iron content of 56–57 percent. The Savage River iron deposits, in Tasmania, were also being developed in the early 1970s for export to Japan. They contain reserves of 100,000,000 tons of magnetite ore.

Ferroalloy metals. Tungsten has been produced in Australia since the end of the 19th century, but since 1920 its contribution to world output has been minor. It is produced from wolframite and scheelite found in Queensland, New England, King Island (in Bass Strait), and the Northern Territory. Manganese has been mined from many small deposits, but over 500,000 tons of ore are produced annually from the recently discovered deposits on Groote Eylandt, in the Gulf of Carpentaria. One of the most significant of the many new economic mineral occurrences recently discovered in Australia is that of nickel. The Kambalda deposits, 35 miles south–southeast of Kalgoorlie, were discovered in 1964, and many other similar deposits in the old goldfields area of the Western Australian Shield were explored in the late 1960s and early 1970s. Recent finds in the Windarra area, in the northern part of the nickel belt, have attracted much attention. In addition, there are large areas of lower grade nickel ores in the Musgrave Block, near the borders of Western Australia, South Australia, and the Northern Territory.

Nonferrous base metals. The Australian continent has long been one of the world's principal producers of lead and zinc. The Broken Hill lode, in western New South Wales, discovered in 1883, produced over 85,000,000 tons of ore in the next 75 years. At Mt. Isa, in western Queensland, there are important lead-zinc and also copper ore bodies, discovered in 1923. By the early 1970s, another lead-zinc deposit was being developed in the McArthur River area of northern Australia. Copper is widely distributed in the Precambrian and Paleozoic rocks of the continent, although most occurrences are small. The principal producing mines are at Mt. Isa and Mt. Morgan in Queensland, Mt. Lyell in Tasmania, and Tennant Creek in the Northern Territory. The South Australian copper mines, at Wallaroo–Moonta and Burra, were of great importance to the early economic development of that state, up to about 1920. With changed mineral economics and improved mining technology, it was found practicable to reopen the Burra and Kanmantoo Mines, near Adelaide, in 1971. Vast resources of bauxite for aluminum production have been discovered at Weipa (on the Gulf of Carpentaria) and at Gove (Arnhem Land) in northern Australia. Tin is produced in both eastern and western Australia, and in the Darwin area; it occurs in lodes connected with granites, and as alluvial deposits. Production declined to a few thousand tons during the early part of the 20th century. Rutile (titanium oxide) and zircon are heavy minerals that are now extensively mined from beach sands, mainly on the east coast.

Copper production

Precious metals. Gold discoveries, which first occurred near Orange, in New South Wales, in the mid-19th century, and later in Victoria, north Queensland, and Western Australian, were important stimulants for the growth of population in Australia. From a peak production of nearly 4,000,000 fine ounces in 1904, output had declined by the 1970s to about 700,000 fine ounces, most of it coming from the Kalgoorlie–Norseman area of Western Australia. Silver occurs in the rich lead–zinc ores in Broken Hill and Mt. Isa, and small amounts of platinum and palladium have been found during the search for nickel.

Nonmetallic deposits. As can be expected from its size, the Australian continent has abundant deposits of such industrial minerals as clays, mica, salt, dolomite, building materials of all kinds, refractories, abrasives, talc, and asbestos. An intensive search for phosphates to offset the declining production of Nauru and Ocean Island led to the discovery of large deposits in the Cloncurry–Mt. Isa area, but the economics of marketing these resources were still under investigation in the early 1970s. Gem minerals occur in many localities, but mechanized industrial prospecting and mining had barely got under way by the early 1970s. The Australian white and "black" opals, mainly from Andamooka and Coober Pedy in South Australia and White Cliffs and Lightening Ridge in western New South Wales, are world famous. The sapphires and topaz from Queensland and the New England district, are also well known.

Mining of gems

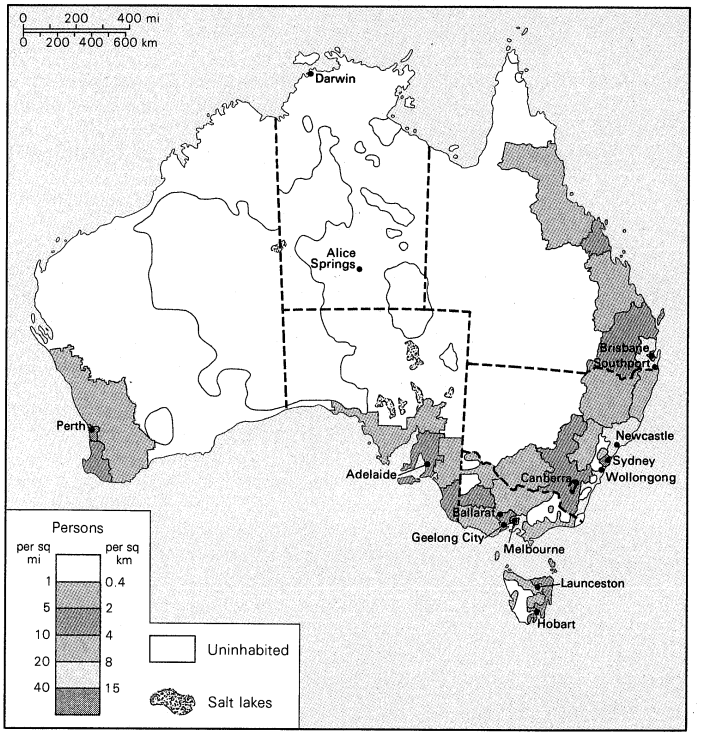
Water resources. If not the oldest continent, Australia is certainly the driest. The average annual rainfall is 17 inches, with less than 10 inches falling in more than

one-third of the continental area. The more significant index of average variability from average annual rainfall is over 20 percent, and reliable rainfall occurs only in the southeast, southwest, and in the north around Darwin. In the interior, the low and unreliable rainfall, the high evaporation rate, and the flat topography combine to reduce the streamflow. The catchment area of the Murray River system, the largest in Australia, is almost a third that of the Nile, but the average annual flow is only one-sixth (12,000,000 acre-feet). The control of water runoff and the general development of storage dams are of great importance in the management of water resources in Australia. An elaborate development scheme for the waters of the southeastern highlands for storage, irrigation, and power, the Snowy Mountains hydro-electric scheme, is nearing completion as one of the greatest projects of this kind ever undertaken. Some of the waters of the Snowy River, which takes the meltwaters of snow from the eastern highlands straight to the sea, are turned into the Murray Basin to be used for irrigation. The total storage capacity of the 15 large dams of the Snowy Mountains scheme will be well over 6,000,000 acre-feet.

The Snowy Mountains scheme

In the Great Artesian Basin there are over 18,000 recorded water bores. The quantity of water drawn from the basin apparently exceeds the rate of recharge, but the need for further investigations and reassessment of data on the hydrology of the Great Artesian Basin is recognized. The underground water resources of the Murray Basin and of other sedimentary basins are significant but insufficiently known. Considerable underground water supplies are available for the development of Alice Springs in central Australia, and drilling has also supplied sufficient water for the mining developments in the northwest. An important problem—at least until desalination becomes economically feasible—is the quality of underground water; the water from the Great Artesian Basin bores is used only for livestock. The groundwater at Kalgoorlie, in Western Australia, is as salty as the sea, and at Norseman salinity is twice as high. Investigations are in progress concerning better use of Australia's water resources as knowledge of their limited amount increases.

Biological resources. The natural biological resources of the Australian continent, as distinct from introduced plants and animals, are limited. Native timber resources in the form of eucalypts are used for papermaking, and the jarrah and karri forests of southwestern Australia provide valuable sawmilling timber. In many parts of the country the virgin native plants still provide the principal pasture component, but in the more arid parts of the country overstocking has destroyed much of this resource and damaged the soil. The arid part of the continent carries only one-third of the total livestock, but introduction of types of cattle more suited to the country and a greater awareness of the limits of its carrying capacity have improved the utilization of these regions. In the zones of better climate, introduced pasture plants such as lucerne and clover, together with the use of superphosphate as fertilizer and of trace elements for soil improvement, has greatly improved Australian pasture resources.



Population density of Australia.

THE OUTLOOK

Australia is a vast continent with a geological history which is long and almost as complex as that of other continents. It has produced a distinctive environment for plants, animals, and man. The Australian environment has clearly recognizable disadvantages: over much of the continent the surface is flat, the soils are leached, sandy or salty, and the climate is arid. Isolation from other major landmasses, possibly itself an effect of geological events (continental drift) has been a significant factor, until modern times of rapid communications. The flora and fauna, in isolation, has adapted itself to its environment in a remarkable manner and has become highly distinctive.

The complex geological history of Australia has provided man not only with a unique challenge but also with rich mineral resources, to the knowledge of which intensive exploration is making significant additions every year. The greatest remaining challenge is that of resources management and conservation of the environment, with the aim of enhancing the productivity of the land, rationalizing the use of the limited water resources, and assuring suitable conditions for the survival of the indigenous flora and fauna and of human populations. (M.F.G.)

BIBLIOGRAPHY

General: DEPARTMENT OF NATIONAL DEVELOPMENT, *Atlas of Australian Resources* (1st series 1952–60, 2nd series 1962–), an official compendium of separate maps on geology, geography (physical and human), and resources (natural and industrial) of Australia, with commentaries; G.W. LEEPER (ed.), *The Australian Environment*, 4th ed. rev. (1970), a brief outline, mainly on land forms, climates, soils, water and irrigation vegetation, crops and pastures, and animal production; R.O. SLATYER and R.A. PERRY (eds.), *Arid Lands of Australia* (1969), symposium papers on many aspects of arid Australia and particularly on land and resources management; V. SERVENTY, *Australia's National Parks* (1969), a beautifully illustrated book on the more important national parks, arranged geographically and dealing with fauna, flora, and geography.

Geology: D.A. BROWN, K.S.W. CAMPBELL, and K.A.W. CROOK, *The Geological Evolution of Australia and New Zealand* (1968); D. HILL and A.K. DENMEAD (eds.), "The Geology of Queensland," *J. Geol. Soc. Aust.*, vol. 7 (1960); G.H. PACKHAM (ed.), "The Geology of New South Wales," *J. Geol. Soc. Aust.*, vol. 16, pt. 1 (1969); L.W. PARKIN (ed.), *Handbook of South Australian Geology* (1969); A. SPRY and M.R. BANKS (eds.), "The Geology of Tasmania," *J. Geol. Soc. Aust.*, vol. 9, pt. 2 (1962). The first of these is a general text for students, on regional and historical geology, divided into chapters on systems from the Precambrian to the Quaternary. The others are the most up-to-date compilations of known geological data for four of the six Australian States. Further information may be obtained from periodical publications of the Bureau of Mineral Resources (Canberra), the State Geological Surveys, and the Geological Society of Australia.

Physical geography: C.F. LASERON, *The Face of Australia* (1953), a readable account of the evolution of Australia's scenery, although unsound in some areas and now outdated; C.R. TWIDALE, *Geomorphology, with Special Reference to Australia* (1968), a general and fairly elementary text, containing numerous references to Australian features and examples of well-known forms. Perhaps the best technical accounts are contained in the series of texts published or to be

published by the ANU Press, Canberra: E.C.F. BIRD, *Coasts* (1968); J.L. DAVIES, *Landforms of Cold Climates* (1969); C.D. OLLIER, *Volcanoes* (1970); C.R. TWIDALE, *Structural Landforms* (1971); and J.A. MABBUTT, *Desert and Savana Landforms*, J.N. JENNINGS, *Karst*, and I. DOUGLAS, *Humid Landforms* (all in press).

Flora: GEORGE BENTHAM, *Flora Australiensis*, 7 vol. (1863–78); S.F. BLAKE and A.C. ATWOOD, "Geographical Guide to the Floras of the World," pt. 1, *Misc. Publs. U.S. Dep. Agric. 401* (1942); W.E. BLACKWELL and B.J. GRIEVE, *How to Know Western Australian Wildflowers*, 3 pt. (1954–65); N.T. BURBIDGE, "The Phytogeography of the Australian Region," *Aust. J. Bot.*, 8:75–211 (1960); D.W. GOODALL (comp.), "Bibliography of Statistical Plant Sociology," *Excerpta Bot.*, sect. B., 4:253–316 (1962).

Fauna: A. KEAST, R.L. CROCKER, and C.S. CHRISTIAN (eds.), *Biogeography and Ecology in Australia* (1959); W.D.L. RIDE, *A Guide to the Native Mammals of Australia* (1970); S. and K. BREEDEN, *The Life of the Kangaroo* (1966); N.W. CAYLEY, *What Bird Is That: A Guide to the Birds of Australia*, 5th ed. rev. and enl. by A.H. CHISHOLM *et al.* (1968); E. WORRELL, *Reptiles of Australia* (1963); W.R. EASTMAN and A.C. HUNT, *The Parrots of Australia* (1966); T.C. ROUGHLEY, *Fish and Fisheries in Australia*, rev. ed. (1966); I.M. MACKERRAS (ed.), *The Insects of Australia: A Textbook for Students and Research Workers* (1970).

Mineral resources: JOHN MCANDREW (ed.), *Geology of Australian Ore Deposits*, 2nd ed. (1965); I.R. MCLEOD (ed.), *Australian Mineral Industry: The Mineral Deposits*, Bureau of Mineral Resources, Geology and Geophysics, Bull. No. 72 (1965); Z. KALIX, L.M. FRASER, and R.I. RAWSON (eds.), *Australian Mineral Industry: Production and Trade, 1842–1964*, Bureau of Mineral Resources, Geology and Geophysics, Bull. No. 81 (1966); H.G. RAGGATT (ed.), *Fuel and Power in Australia* (1969).

(M.F.G./C.R.T./R.T.La.)

Australia, Commonwealth of

The Commonwealth of Australia is located in the Southern Hemisphere between the Indian Ocean and the South Pacific, on an island continent that has been called both the Oldest Continent and the Last of Lands. Neither description is accurate. It is the oldest continent only in the sense that much of Australia is formed of rocks laid down during the Precambrian (4,600,000,000 to 570,000,000 years ago), and has altered relatively little since life first appeared on earth. It is the last of lands only in the sense that it was the last continent (if we except Antarctica) to be discovered and explored by Europeans. Thousands of years before the explorers Abel Tasman and James Cook sailed into the South Pacific, the Aborigines had crossed the land bridge from Asia formed by the Malaysian archipelago and had spread throughout the mainland and Tasmania. They remained, however, a sparse, primitive, and nomadic people. When Captain Arthur Phillip of the British Royal Navy landed with the First Fleet at Botany Bay in 1788, the true beginning of modern Australia, there were probably not more than 300,000 Aborigines altogether, and there was no structure on the continent more solid than a bark hut.

Today the Commonwealth of Australia is a prosperous, independent nation of some 12,700,000 people united under one government, of whom about 80 percent are of British stock and only 1 percent nonwhite. A British dominion and a member of the British Commonwealth of Nations, it had become the most important monument to the British Empire in the Southern Hemisphere. Australians were fortunate in that they did not have to share their continent—which, with almost 3,000,000 square miles (8,000,000 square kilometres), is only a little smaller than the United States—with any other nation. It was all their own. Their nearest neighbours, apart from Papua-New Guinea, which was administered by Australia, were the Indonesians to the north, the Polynesians and Melanesians of the Pacific islands to the east, and New Zealand, which, like Australia, was a British dominion and a member of the British Commonwealth, to the southeast. On the other hand, Australia was remote from its two principal allies: it is 12,000 miles from Australia to Britain via the Indian Ocean and the Suez Canal; it is 7,000 miles across the Pacific to the West Coast of the United States.

Like Canada and the United States, contemporary Australia is a political federation with a central government (the Commonwealth) and six states (New South Wales, Victoria, Queensland, South Australia, Western Australia, and Tasmania), each of which has its own government enjoying a limited sovereignty. The Commonwealth also directly administers two internal territories, the Northern Territory and the Australian Capital Territory, and the external territories of Papua-New Guinea, Norfolk Island, Cocos Islands, Christmas Island, Ashmore and Cartier Islands, Coral Sea Islands, Heard and McDonald Islands, and the Australian Antarctic Territory.

This article describes the landscape, people, economic development, and political, constitutional, and cultural aspects of the contemporary Commonwealth of Australia, taken as a whole. The article AUSTRALIA treats in detail geological evolution and the physical environment; AUSTRALIA, HISTORY OF, deals with the historical development of the nation.

THE LANDSCAPE

Regional variations. Australia is both the flattest continent and the driest. Seen from the air it is hard to believe that its vast plains, sometimes the colour of dried blood, more often tawny like a lion's skin, are not one huge desert. One can fly the 1,900 miles to Sydney from Darwin in the north, or the 2,000 miles from Perth in the west, without seeing a single town or anything but the most scattered and minute signs of human habitation. A good deal of the central depression and western plateau is indeed desert. Yet appearances can be deceptive. The red and blacksoil plains of Queensland and New South Wales have long supported the world's greatest wool industry, while recent discoveries have revealed that some of the most arid and forbidding areas of Australia conceal great mineral wealth.

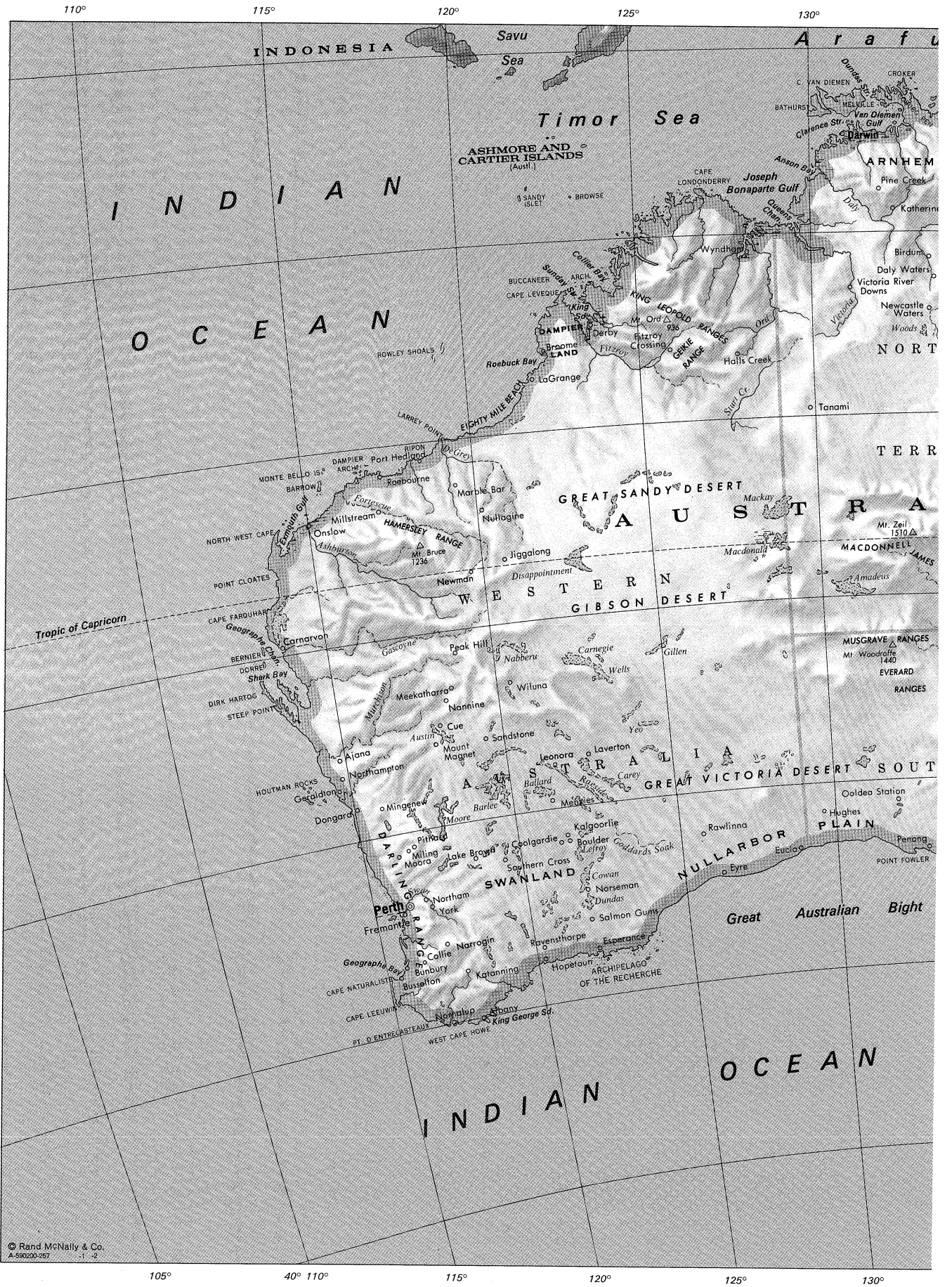
Moreover, the coastal rim is, almost everywhere, an exception to these rules. In particular the east coast, where European settlement began and where the majority of Australians now live, is hilly, well watered, and fertile.

Inland from the coast runs a chain of highlands known as the Great Dividing Range, from Cape York in northern Queensland to the southern seaboard of Tasmania. From the coast itself this range, which may be anything from 20 miles (32 kilometres) to 200 miles (320 kilometres) distant, often appears as a bold range of mountains, though few of its peaks exceed 5,000 feet (1,500 metres). In fact, it is more like the escarpment of a giant plateau, formed of gently rolling hills, which then slopes imperceptibly down to the western plains. There are similar, though smaller, stretches of hilly, well-watered land all around the rim of the continent except on the south coast where the Nullarbor Plain stretches to the sea; but everywhere the rainfall diminishes rapidly as one penetrates further from the coast.

The "Outback"

To Australians the land beyond the Great Dividing Range and the coastal rim is the Inland, or the "Outback." For them it still retains some of the mythical quality it had for the first explorers searching for inland seas and great rivers. It is their Frontier: the land of hope and adventure. Yet in fact it is still very sparsely populated and perhaps always will be. The real heart of Australia lies in the industrial cities of the east and west coasts.

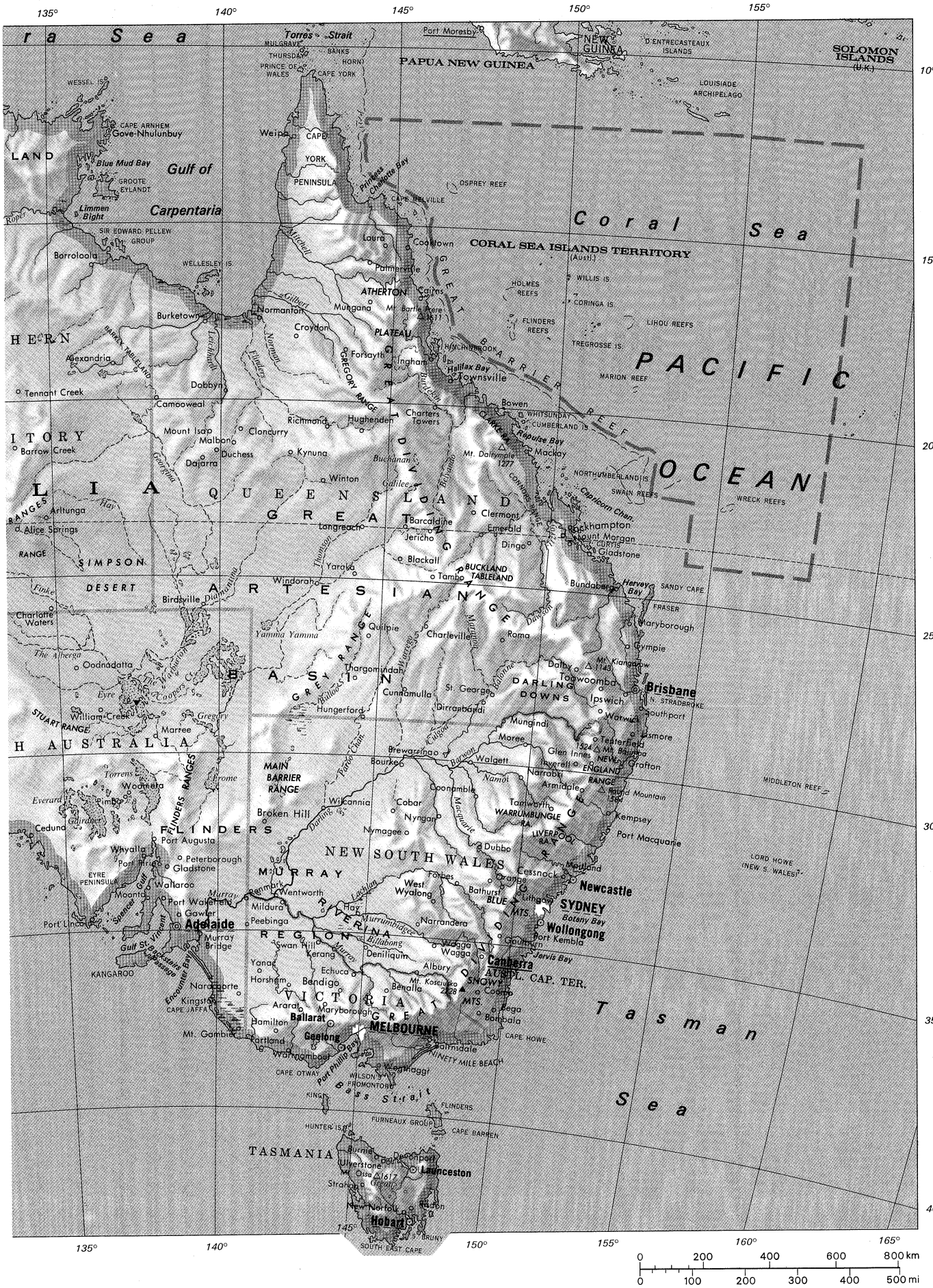
In this huge continent there are wide variations in scenery and climate. The thickly wooded ranges of the Great Divide have little in common with the treeless, sun-dried plains of the Inland. There is a vast difference between the red rocks and monumental hills of central Australia and the tropical rain forests and sugar plantations of north Queensland. Yet visitors to Australia usually detect a certain uniformity created, perhaps, more by the red earth, the brilliant light, and the drab, olive-coloured leaves of the ubiquitous eucalyptus than by any real resemblance. And if visitors from the Northern Hemisphere are at first repelled, as the English novelist D.H. Lawrence was, by "the vast, uninhabited land and by the grey charred bush . . . so phantom-like, so ghostly, with its tall, pale trees and many dead trees, like corpses," they should remember that to Australians born in the country



COMMONWEALTH OF AUSTRALIA

Size of symbol indicates relative size of town ○ ⊙ ◎

Elevations in metres


135°
140°
145°
150°
155°
160°
165°
10°
15°
20°
25°
30°
35°
40°
Torres Strait
Port Moresby
PAPUA NEW GUINEA
NEW GUINEA
D'ENTRECASTEAUX ISLANDS
SOLOMON ISLANDS (U.K.)
LOUISIADE ARCHIPELAGO
MULGRAVE
THURSDAY
BANKS
PRINCE OF WALES
CAPE YORK
HORN)
WESSEL IS.
CAPE ARNHEM
Gove-Nhulunbuy
LAND
Gulf of
Carpentaria
Blue Mud Bay
GROOTE EYLANDT
Limmen Bight
SIR EDWARD PELLEW GROUP
Weipa
CAPE YORK PENINSULA
Princess Charlotte Bay
OSPREY REEF
CAPE MELVILLE
Coral Sea
CORAL SEA ISLANDS TERRITORY (Austl.)
Laura
Cooktown
Mitchell
Palmerville
Cairns
ATHERTON
Mungana
Mt. Bartle Frere 5,611
GREAT DIVIDING RANGE
HOLMES REEFS
WILLIS IS.
CORINGA IS.
FLINDERS REEFS
LIHOU REEFS
TREGROSSE IS.
WELLESLEY IS.
Borroloola
Burketown
Normanton
Gilbert
Croydon
PLATEAU
Forsayth
Ingham
HYCHINBROOK
Halifax Bay
Townsville
GREAT BARRIER REEF
MARION REEF
PACIFIC OCEAN
Alexandria
BARKLY TABLELAND
Leichhardt
Flinders
Norman
Tennant Creek
Dobbyn
Camooweal
Mount Isa
Malbon
Cloncurry
Richmond
Hughenden
Charters Towers
Bowen
WHITSUNDAY
CUMBERLAND IS.
Repulse Bay
Mackay
Barrow Creek
Duchess
Kynuna
Dajarra
Georgina
Buchanan
Galilee
Diamantina
Mt. Dalrymple 1277
NORTHUMBERLAND IS.
SWAIN REEFS
WRECK REEFS
Capricorn Chan.
Winton
QUEENSLAND
GREAT ARTESIAN BASIN
Arltunga
RANGES
Hay
Alice Springs
RANGE
SIMPSON DESERT
Barcaldine
Clermont
Emerald
Rockhampton
Mount Morgan
CURTIS
Gladstone
Langreachc
Jericho
Thomson
Blackall
Dingo
Yaraka
Tambo
BUCKLAND TABLELAND
Bundaberg
Harvey Bay
SANDY CAPE
FRASER
Finke
Diamantina
Windorah
Birdsville
Charlotte Waters
Quilpie
Yamma Yamma
Charleville
Roma
Maryborough
Gympie
The Alberga
Oodnadatta
The Warburton
The Coopers Creek
Eyre
GREY RANGE
Thargomindah
Wallam
Bulloo
Cunnamulla
St. George
Dirranbandi
DARLING DOWNS
Dalby
Mt. Kiangaroo 1140
Toowoomba
Brisbane
N. STRADBROKE
Ipswich
Southport
Warwick
Lismore
Tenterfield
1524 Mt. Barney
NEW ENGLAND RANGE
Grafton
Glen Innes
Inverell
Narrabri
Armidale
Round Mountain 1564
MIDDLETON REEF
LORD HOWE (NEW S. WALES)
Kempsey
Port Macquarie
Tamworth
WARRUMBUNGLE RA.
LIVERPOOL RA.
Dubbo
Nyngan
Cobar
Coonamble
Namoi
Macquarie
Darling
Broken Hill
Wilcannia
Nymagee
Brewarrina
Bourke
Walgett
Hungerford
Cuddapan
Stuart Range
Marree
Torrens
Woomera
Everard
Frome
Gairdner
Ceduna
Whyalla
Port Augusta
Port Pirie
Peterborough
Gladstone
EYRE PENINSULA
Spencer Gulf
Wallaroo
Moonta
Port Wakefield
Gawler
Port Lincoln
Adelaide
MAIN BARRIER RANGE
FLINDERS RANGES
NEW SOUTH WALES
MURRAY
RIVERINA
REGION
Murray
Renmark
Wentworth
Mildura
Peebinga
Lachlan
Murrumbidgee
Narrandera
Forbes
West Wyalong
Orange
Bathurst
Lithgow
BLUE MTS.
Lessnock
Maitland
Newcastle
SYDNEY
Botany Bay
Wollongong
Port Kembla
Gulf St. Vincent
Backstairs Passage
Encounter Bay
Murray Bridge
Swan Hill
Kerang
Billabong
Murray
Wagga
Wagga
Goulburn
Jervis Bay
Canberra
AUST. CAP. TER.
KANGAROO
Yanac
Horsham
Bendigo
Echuca
Deniliquin
Albury
Mt. Kosciusko 2228
SNOWY MTS.
Cooma
Bega
Bombala
Naracoorte
Kingston
CAPE JAFFA
Ararat
Maryborough
Ballarat
Hamilton
Mt. Gambier
Portland
Geelong
MELBOURNE
VICTORIA
GREAT
Tasman Sea
CAPE HOWE
Port Phillip Bay
Warrnambool
Sorrento
Ninety Mile Beach
CAPE OTWAY
Wonthaggi
WILSON PROMONTORY
KING I.
Bass Strait
FLINDERS
FURNEAUX GROUP
HUNTER IS.
CAPE BARREN
TASMANIA
Burnie
Devonport
Ulverstone
Launceston
Mt. Ossa 1617
Strahan
Gordon
Derwent
New Norfolk
Hobart
Tasman Sea
SOUTH EAST CAPE
BRUNY
0 200 400 600 800 km
0 100 200 300 400 500 mi

MAP INDEX

Political subdivisions

Australian Capital Territory.....35·30s 149·00e
New South Wales.....33·00s 146·00e
Northern Territory.....20·00s 134·00e
Queensland....20·00s 145·00e
South Australia.....30·00s 135·00e
Tasmania.....42·00s 147·00e
Victoria.....38·00s 145·00e
Western Australia.....25·00s 122·00e

Cities and towns

Adelaide.....34·55s 138·35e
Ajana.....27·57s 114·38e
Albany.....35·02s 117·53e
Albury.....36·05s 146·55e
Alexandria.....19·05s 136·40e
Alice Springs..23·42s 133·53e
Ararat.....37·17s 142·56e
Arltunga.....23·26s 134·41e
Armidale.....30·31s 151·39e
Bairnsdale.....37·50s 147·38e
Ballarat.....37·34s 143·52e
Barcaldine.....23·33s 145·17e
Barrow Creek..21·33s 133·53e
Bathurst.....33·25s 149·35e
Bega.....36·40s 149·50e
Benalla.....36·33s 145·59e
Bendigo.....36·46s 144·17e
Birdsville.....25·54s 139·22e
Birdum.....15·39s 133·13e
Blackall.....24·25s 145·28e
Bombala.....36·54s 149·14e
Borroloola.....16·04s 136·17e
Boulder.....30·47s 121·29e
Bourke.....30·05s 145·56e
Bowen.....20·01s 148·15e
Brewarrina.....29·57s 146·52e
Brisbane.....27·28s 153·02e
Broken Hill.....31·57s 141·27e
Broome.....17·58s 122·14e
Bunbury.....33·19s 115·38e
Bundaberg....24·52s 152·21e
Burketown.....17·44s 139·22e
Burnie.....41·04s 145·54e
Busselton.....33·39s 115·20e
Cairns.....16·55s 145·46e
Camooweal....19·55s 138·07e
Canberra.....35·17s 149·08e
Capoompeta...29·23s 152·01e
Carnarvon.....24·53s 113·40e
Ceduna.....32·07s 133·40e
Cessnock.....32·50s 151·21e
Charleville.....26·24s 146·15e
Charlotte Waters.....25·55s 134·55e
Charters Towers.....20·05s 146·16e
Clermont.....22·49s 147·39e
Cloncurry.....20·42s 140·30e
Cobar.....31·30s 145·49e
Collie.....33·21s 116·09e
Cooktown.....15·28s 145·15e
Coolgardie.....30·57s 121·10e
Cooma.....36·14s 149·08e
Coonamble....30·57s 148·23e
Croydon.....18·12s 142·14e
Cue.....27·25s 117·54e
Cunnamulla...28·04s 145·41e
Dajarra.....21·42s 139·31e
Dalby.....27·11s 151·16e
Daly Waters....16·15s 133·22e
Darwin.....12·28s 130·50e
Deniliquin.....35·32s 144·58e
Derby.....17·18s 123·38e
Devonport.....41·11s 146·21e
Dingo.....23·39s 149·20e
Dirranbandi...28·35s 148·14e
Dobbyn.....19·48s 140·00e
Dongara.....29·15s 114·56e
Dubbo.....32·15s 148·36e
Duchess.....21·22s 139·52e
Echuca.....36·08s 144·46e
Emerald.....23·32s 148·10e
Esperance.....33·51s 121·53e
Eucla.....31·43s 128·52e
Eyre.....32·15s 126·18e
Fitzroy Crossing.....18·11s 125·35e
Forbes.....33·23s 148·01e
Forsayth.....18·35s 143·36e
Fremantle.....32·03s 115·45e
Gawler.....34·37s 138·44e
Geelong.....38·08s 144·21e
Geraldton.....28·46s 114·36e
Gladstone.....23·51s 151·16e
Glen Innes.....29·44s 151·44e
Goulburn.....34·45s 149·43e
Gove-Nhulunbuy...12·10s 136·45e
Grafton.....29·41s 152·56e
Gympie.....26·11s 152·40e
Halls Creek....18·13s 127·40e
Hamilton.....37·45s 142·02e
Hay.....34·30s 144·51e
Hobart.....42·53s 147·19e
Hopetoun.....33·57s 120·07e
Horsham.....36·43s 142·13e
Hughenden....20·51s 144·12e
Hughes.....30·42s 129·31e
Hungerford....29·00s 144·25e
Ingham.....18·39s 146·10e
Inverell.....29·47s 151·07e
Ipswich.....27·36s 152·46e
Jericho.....23·36s 146·08e
Jiggalong.....23·25s 120·47e
Kalgoorlie.....30·45s 121·28e
Katanning.....33·42s 117·33e
Katherine.....14·28s 132·16e
Kempsey.....31·05s 152·50e
Kerang.....35·44s 143·55e
Kingston.....36·50s 139·51e
Kynuna.....21·35s 141·55e
LaGrange.....18·41s 121·45e
Lake Brown....30·57s 118·20e
Launceston....41·26s 147·08e
Laura.....15·34s 144·28e
Laverton.....28·38s 122·25e
Leonora.....28·53s 121·19e
Lismore.....28·48s 153·17e
Lithgow.....33·29s 150·09e
Longreach.....23·26s 144·15e
Mackay.....21·09s 149·11e
Maitland.....32·44s 151·33e
Malbon.....21·04s 140·18e
Marble Bar....21·11s 119·44e
Marree.....29·39s 138·04e
Maryborough..37·03s 143·45e
Maryborough..25·32s 152·42e
Meekatharra...26·36s 118·29e
Melbourne.....37·49s 144·58e
Menzies.....29·41s 121·02e
Mildura.....34·12s 142·09e
Miling.....30·30s 116·21e
Millstream.....21·35s 117·04e
Mingenew.....29·11s 115·26e
Moonta.....34·04s 137·35e
Moora.....30·39s 116·00e
Moree.....29·28s 149·51e
Mount Gambier.....37·50s 140·46e
Mount Isa.....20·44s 139·30e
Mount Magnet.28·04s 117·49e
Mount Morgan.23·39s 150·23e
Mungana.....17·07s 144·24e
Mungindi.....28·58s 148·59e
Murray Bridge.35·07s 139·17e
Nannine.....26·53s 118·20e
Naracoorte....36·58s 140·44e
Narrabri.....30·19s 149·47e
Narrandera....34·45s 146·33e
Narrogin.....32·56s 117·10e
Newcastle.....32·56s 151·46e
Newcastle Waters.....17·24s 133·24e
Newman.....23·28s 119·40e
New Norfolk...42·47s 147·03e
Normanton....17·40s 141·05e
Nornalup.....35·00s 116·49e
Norseman.....32·12s 121·46e
Northam.....31·39s 116·40e
Northampton..28·21s 114·37e
Nullagine.....21·53s 120·06e
Nymagee.....32·04s 146·20e
Nyngan.....31·34s 147·11e
Onslow.....21·39s 115·06e
Oodnadatta....27·33s 135·28e
Ooldea Station.30·27s 131·50e
Orange.....33·17s 149·06e
Palmerville....15·59s 144·05e
Peak Hill.....25·38s 118·43e
Peebinga.....34·56s 140·55e
Penong.....31·55s 133·01e
Perth.....31·56s 115·50e
Peterborough..32·58s 138·50e
Pimba.....31·15s 136·47e
Pine Creek....13·49s 131·49e
Pithara.....30·24s 116·40e
Port Augusta...32·30s 137·46e
Port Hedland..20·19s 118·34e
Port Kembla...34·28s 150·54e
Portland.....38·21s 141·36e
Port Lincoln....34·44s 135·52e
Port Macquarie....31·26s 152·55e
Port Pirie.....33·11s 138·01e
Port Wakefield.34·11s 138·09e
Qulpie.....26·37s 144·15e
Ravensthorpe..33·35s 120·02e
Rawlinna.....31·01s 125·20e
Renmark.....34·11s 140·45e
Richmond.....20·44s 143·08e
Risdon.....42·48s 147·20e
Rockhampton..23·23s 150·31e
Roebourne.....20·47s 117·09e
Roma.....26·35s 148·47e
Saint George...28·02s 148·35e
Salmon Gums..32·59s 121·38e
Sandstone.....27·59s 119·17e
Southern Cross.....31·13s 119·19e
Southport.....27·58s 153·25e
Strahan.....42·09s 145·19e
Swan Hill.....35·21s 143·34e
Sydney.....33·52s 151·13e
Tambo.....24·53s 146·15e
Tamworth.....31·05s 150·55e
Tanami.....19·59s 129·43e
Tennant Creek.....19·40s 134·10e
Tenterfield.....29·03s 152·01e
Thargomindah.28·00s 143·49e
Toowoomba...27·33s 151·57e
Townsville.....19·16s 146·48e
Ulverstone.....41·09s 146·10e
Victoria River Downs.....16·24s 131·00e
Wagga Wagga.35·07s 147·22e
Walgett.....30·01s 148·07e
Wallaroo.....33·56s 137·38e
Warrnambool..38·23s 142·29e
Warwick.....28·13s 152·02e
Weipa.....12·41s 141·52e
Wentworth.....34·07s 141·55e
West Wyalong..33·55s 147·13e
Whyalla.....33·02s 137·35e
Wilcannia.....31·34s 143·23e
William Creek..28·55s 136·21e
Wiluna.....26·36s 120·13e
Windorah.....25·26s 142·39e
Winton.....22·23s 143·02e
Wollongong....34·25s 150·54e
Wonthaggi.....38·36s 145·35e
Woomera.....31·31s 137·10e
Wyndham.....15·28s 128·06e
Yanac.....36·08s 141·26e
Yaraka.....24·53s 144·04e
York.....31·53s 116·46e

Physical features and points of interest

Amadeus, *dry lake*.....24·50s 130·45e
Anson Bay.....13·20s 130·06e
Arafura Sea...10·00s 134·00e
Arnhem, Cape.12·21s 136·21e
Arnhem Land, *aboriginal reserve*.....13·10s 134·30e
Ashburton, *river*.....21·40s 114·56e
Atherton Plateau.....17·00s 145·00e
Austin, *dry lake*.....27·40s 118·00e
Bajimba, Mount, *mountain*.....29·18s 152·07e
Backstairs Passage.....35·42s 138·05e
Ballard, *dry lake*.....29·27s 120·55e
Balonne, *river*..27·47s 147·56e
Banks, *island*..10·12s 142·16e
Barkly Tableland, *upland*.....19·00s 138·00e
Barlee, *dry lake*.....29·10s 119·30e
Barrow, *island*.20·48s 115·23e
Bartle Frere, Mount, *mountain*.....17·23s 145·49e
Barwon, *river*..30·00s 148·05e
Bass Strait....39·20s 145·30e
Bathurst, *island*.....11·37s 130·23e
Belyando, *river*.21·38s 146·50e
Bernier, *island*.24·52s 113·08e
Billabong, *river*.....34·55s 143·05e
Blue Mountains....33·33s 150·17e
Blue Mud Bay..13·26s 135·56e
Botany Bay....33·59s 151·12e
Browse, *island*.14·07s 123·33e
Bruce, Mount, *mountain*.....22·36s 118·08e
Buccaneer Archipelago, *islands*.....16·17s 123·20e
Buchanan, *lake*.....21·28s 145·52e
Buckland Tableland, *upland*.....24·35s 147·55e
Bulloo, *river*....28·43s 142·30e
Burdekin, *river*.19·39s 147·30e
Cape Barren, *island*.....40·25s 148·12e
Cape York Peninsula.....14·00s 142·30e
Capricorn Channel.....23·28s 152·00e
Carey, *salt lake*.29·05s 122·15e
Carnegie, *dry lake*.....26·10s 122·30e
Carpentaria, Gulf of.....14·00s 139·00e
Clarence Strait.12·00s 131·00e
Clarke Range, *mountain range*.....20·50s 148·33e
Cloates, Point..22·43s 113·40e
Coburg Peninsula.....11·20s 132·15e
Collier Bay.....16·10s 124·15e
Connors Range, *mountain range*.....21·40s 149·10e
Coopers Creek.28·29s 137·46e
Coral Sea.....15·00s 155·00e
Cowan, *salt lake*.....31·50s 121·50e
Croker, *island*..11·12s 132·32e
Culgoa, *river*...29·56s 146·20e
Cumberland Islands.....20·40s 149·09e
Curtis, *island*...23·38s 151·09e
Dalrymple, Mount, *mountain*.....21·02s 148·38e
Daly, *river*.....13·20s 130·19e
Dampier Archipelago, *islands*.....20·35s 116·35e
Dampier Land, *peninsula*.....17·30s 122·55e
Darling, *river*...34·07s 141·55e
Darling Downs, *physical region*.....27·30s 150·30e
Darling Range, *mountain range*.....32·00s 116·30e
Dawson, *river*..23·38s 149·46e
DeGrey, *river*...20·12s 119·11e
D'Entrecasteaux, Point.....34·50s 116·00e
Diamantina, *river*.....26·45s 139·10e
Dirk Hartog, *island*.....25·48s 113·00e
Disappointment, *salt lake*.....23·30s 122·50e
Dorre, *island*...25·09s 113·07e
Dundas, *dry lake*.....32·35s 121·50e
Dundas Strait,.11·20s 131·35e
Eighty Mile Beach.....19·45s 121·00e
Encounter Bay.35·35s 138·44e
Everard, *dry lake*.....31·25s 135·05e
Everard Ranges, *mountain range*.....27·05s 132·28e
Exmouth Gulf..22·00s 114·20e
Eyre, *lake*.....26·40s 139·00e
Eyre Peninsula.34·00s 135·45e
Farquhar, Cape.....23·37s 113·37e
Finke *river*.....26·20s 136·00e
Fitzroy, *river*...17·31s 123·35e
Flinders, *physical region*.....32·30s 140·00e
Flinders, *river*..17·36s 140·36e
Flinders, *island*.40·00s 148·00e
Flinders Ranges, *mountain range*.....31·25s 138·45e
Fortescue, *river*.....21·00s 116·06e
Fowler, Point...32·02s 132·29e
Fraser, *island*..25·15s 153·10e
Frome, *dry lake*.30·48s 139·48e
Furneaux Group, *islands*.....40·10s 148·05e
Gairdner, *lake*..31·35s 136·00e
Galilee, *lake*....22·21s 145·48e
Gascoyne, *river*.....24·52s 113·37e
Geikie Range, *mountain range*.....18·07s 125·46e
Geographe Bay.....33·35s 115·15e
Geographe Channel.....24·40s 113·20e
Georgina, *river*.23·30s 139·47e
Gibson Desert..24·30s 126·00e
Gilbert, *river*...16·35s 141·15e
Gillen, *dry lake*.26·11s 124·38e
Goddards Soak, *swamp*.....31·20s 123·30e
Great, *lake*.....41·52s 146·45e
Great Artesian Basin, *physical region*.....25·00s 143·00e
Great Australian Bight.....35·00s 135·00e

MAP INDEX (continued)

Great Barrier Reef.....18·00s 146·50e
Great Dividing Range, *mountain range*.....25·00s 147·00e
Great Sandy Desert.....21·30s 125·00e
Great Victoria Desert.....28·30s 127·45e
Gregory, *dry lake*.....28·55s 139·00e
Gregory Range, *mountain range*.....19·00s 143·05e
Grey Range, *mountain range*.....27·00s 143·35e
Groote Eylandt, *island*.....14·00s 136·40e
Halifax Bay....18·50s 146·30e
Hamersley Range, *mountain range*.....21·53s 116·46e
Hay, *river*.....25·14s 138·00e
Hervey Bay....25·00s 153·00e
Hinchinbrook, *island*.....18·23s 146·17e
Horn, *island*...10·37s 142·17e
Houtman Rocks.....28·35s 113·45e
Howe, Cape...37·31s 149·59e
Hunter Islands.40·32s 144·45e
Indian Ocean..38·00s 122·30e
Jaffa, Cape....36·58s 139·40e
James Range, *mountain range*.....24·06s 132·30e
Jervis Bay.....35·05s 150·45e
Joseph Bonaparte Gulf.....14·15s 128·30e
Kangaroo, *island*.....35·50s 137·06e
King, *island*....39·50s 144·00e
King George Sound.....35·03s 117·57e
King Leopold Ranges, *mountain range*.....17·30s 125·45e
King Sound...17·00s 123·30e
Kosciusko, Mount, *mountain*.....36·27s 148·16e
Lachlan, *river*..34·21s 143·57e
Larrey Point...19·58s 119·07e
Leeuwin, Cape.34·22s 115·08e
Lefroy, *dry lake*.31·15s 121·40e
Leichhardt, *river*.....17·35s 139·48e
Leveque, Cape.16·24s 122·56e
Limmen Bight..14·45s 135·40e
Liverpool Range, *mountain range*.....31·40s 150·30e
Londonderry, Cape.....13·45s 126·55e
Lord Howe, *island*.....31·33s 159·05e
Macdonald, *dry lake*.....23·30s 129·00e
Macdonnell Ranges, *mountain range*.....23·45s 133·20e
Mackay, *salt lake*.....22·30s 129·00e
Macquarie, *river*.....30·07s 147·24e
Main Barrier Range, *mountain range*.....31·25s 141·25e
Maranoa, *river*.27·50s 148·37e
Melville, Cape..14·11s 144·30e
Melville, *island*.11·40s 131·00e
Mitchell, *river*..15·12s 141·35e
Monte Bello Islands.....20·25s 115·32e
Moore, *salt lake*.....29·50s 117·35e
Mulgrave, *island*.....10·07s 142·08e
Murchison, *river*.....26·01s 117·06e
Murray, *river*...35·22s 139·22e
Murray Region.34·00s 143·00e
Murrumbidgee, *river*.....34·43s 143·12e
Musgrave Ranges, *mountain range*.....26·10s 131·50e
Nabberu, *dry lake*.....25·36s 120·30e
Namoi, *river*....30·00s 148·07e
Naturaliste, Cape.....33·32s 115·01e
New England Range, *mountain range*.....30·00s 151·50e
Ninety Mile Beach.....38·13s 147·23e
Norman, *river*..17·28s 140·49e
North Stradbroke Islands.....27·35s 153·28e
Northumberland Islands..21·40s 150·00e
North West Cape.....21·45s 114·10e
Nullarbor Plain.....31·00s 129·00e
Ord, *river*.....15·30s 128·21e
Otway, Cape...38·52s 143·31e
Ossa, Mount, *mountain*.....41·54s 146·01e
Pacific Ocean..20·00s 156·00e
Paroo Channel.31·28s 143·32e
Port Phillip Bay.....38·07s 144·48e
Prince of Wales, *island*..10·40s 142·10e
Princess Charlotte Bay.14·25s 144·00e
Queens Channel.....14·46s 129·24e
Raeside, *dry lake*.....29·30s 122·00e
Recherche, Archipelago of the, *islands*.34·05s 122·45e
Repulse Bay...20·36s 148·43e
Ripon, *island*...20·07s 119·12e
Riverina, *physical region*.....35·30s 145·30e
Roebuck Bay. 19·04s 122·17e
Roper, *river*....14·43s 135·27e
Round Mountain.....36·15s 148·34e
Rowley Shoals.17·30s 119·00e
Saint Vincent, Gulf.....35·00s 138·05e
Sandy Islet, *island*.....14·03s 121·49e
Sandy Cape...41·25s 144·45e
Shark Bay.....25·30s 113·30e
Simpson Desert.....25·00s 137·00e
Sir Edward Pellew Group, *islands*.....15·40s 136·48e
Snowy Mountains....36·30s 148·20e
South Bruny, *island*.....43·23s 147·17e
South East Cape.....43·39s 146·50e
Spencer Gulf. 34·00s 137·00e
Steep Point....26·08s 113·08e
Stuart Range, *mountain range*.....29·10s 134·56e
Sturt Creek....20·08s 127·24e
Sunday Strait..16·25s 123·18e
Swain Reefs...21·40s 152·15e
Swan, *river*....32·03s 115·45e
Swanland, *physical region*.....32·00s 120·00e
Tasmania, *island*.....42·00s 147·00e
Tasman Sea...37·30s 156·00e
The Alberga, *river*.....27·06s 135·33e
The Warburton, *river*.....27·55s 137·28e
Thomson, *river*.25·11s 142·53e
Thursday, *island*.....10·35s 142·13e
Timor Sea.....13·00s 125·00e
Torrens, *salt lake*.....31·00s 137·50e
Torres Strait...10·25s 142·10e
Van Diemen, Cape.....11·10s 130·23e
Van Diemen Gulf.....11·50s 132·00e
Victoria, *river*..15·12s 129·43e
Warrego, *river*..30·24s 145·21e
Warrumbungle Range, *mountain range*.....31·27s 149·10e
Wellesley Islands.....16·42s 139·30e
Wells, *dry lake*.26·43s 123·10e
Wessel Islands.11·30s 136·25e
West Cape Howe.....35·08s 117·36e
Whitsunday, *island*.....20·17s 148·59e
Wilson's Promontory...38·55s 146·20e
Woodroffe, Mount, *mountain*.....26·20s 131·45e
Woods, *lake*....17·50s 133·30e
Yamma Yamma, *lake*..26·20s 141·25e
Yeo, *dry lake*....28·04s 124·23e
York, Cape....10·42s 142·31e
Zeil, Mount, *mountain*.....23·24s 132·23e

the bush is friendly and familiar. Australia is not a pretty country, but it has a unique and haunting beauty that exerts a powerful fascination on those who get to know it.

If there is no real uniformity in the Australian landscape, there is certainly uniformity among the Australian people. No marked regional differences have emerged in the 200 years of European development, and steadily improving means of transport and communication have constantly worked to erase such differences as did exist. Today there is a strong similarity in the speech, manners, and customs of all Australian states, and everywhere the culture of white Australia is immediately recognizable as characteristic of Anglo-Saxon culture in Britain and North America.

The pattern of settlement. Australia did not yield easily to development by white men. Even on the relatively favoured eastern seaboard, the first settlers found it hard to cross the rugged mountains of the Great Divide. When they did so they had to fight an endless battle against savage droughts, sudden floods, and fierce bush fires. Though they had little to fear from the Aborigines or from the gentle and harmless Australian animals, the land itself often appeared as a harsh and implacable foe. The long struggle by these settlers to tame the Australian Outback helped to form the tough and independent character of modern Australians just as the struggles of the pioneers and frontiersmen helped to mold the national character of the United States.

By the second half of the 20th century, the automobile, the airplane, and the radio had greatly eased the harshness of life in country districts. Although many families still live in astonishing isolation by the standards of most other countries—it is nothing for a grazier to live 50 miles from the nearest country town and perhaps 10 or 15 miles from his nearest neighbour—most of them are linked by telephone and by well-surfaced roads. Even those who live on the great sheep and cattle stations beyond the Darling in New South Wales or in northwest Queensland are linked by radio and can call the radio doctor and the flying ambulance in emergency. Many of the bigger stations own their own airplanes.

Sheep and cattle stations

In Australia a sharp distinction has been made in the past between the grazier, who runs sheep or cattle on his "station" or "property," and the small farmer who grows wheat or fruit or raises dairy cattle on the coast or in the irrigation areas. That distinction had been blurred in the early 1970s by the subdivision of the biggest properties and by the fact that many graziers, faced by rising costs and the falling price of wool, had been forced to grow wheat or other crops to supplement their income. But Australia is still one of the last countries where one can find huge properties of 100,000 acres or more where sheep and cattle graze over land unbroken by the plough. Life on a big sheep station is still a privileged one where a comfortable income, the pride of owning wide acres, and the spacious dignity of life on these sunburnt plains more than compensates for the loneliness and the inevitable hazards of drought and bush fires.

The life of the small farmer in Australia is less attractive and, consequently, less envied. Most of them struggle against the same hazards of drought and fire without the reassurance of broad acres to fall back on. Very often their homes are modest, and their cash income may not equal that of an industrial worker on the basic wage. Many of them could not survive or compete with imported produce without generous government subsidies.

Because of the vast distances and sparse settlement, nothing like the European village has ever developed in Australia. Instead, there are country towns that serve a wide area and vary a great deal in size and amenities.

Many of the Outback towns are dusty little settlements with one wide main street, a store, two or three hotels, and not much else. But in the areas of closer settlement nearer the coast, many substantial country towns have grown to a point where they offer excellent medical and educational services and first-class shopping. Very few towns in the Outback, however, have more than 25,000 inhabitants.

Commonwealth of Australia, Area and Population

	area		population	
	sq mi	sq km	1966 census	1971 census*
States				
New South Wales	309,433	801,428	4,238,000	4,590,000
Queensland	667,000	1,727,522	1,674,000	1,823,000
South Australia	380,070	984,377	1,095,000	1,173,000
Tasmania	26,383	68,332	371,000	390,000
Victoria	87,884	227,619	3,220,000	3,496,000
Western Australia	975,920	2,527,621	848,000	1,027,000
Territories				
Australian Capital Territory	939	2,432	96,000	144,000
Northern Territory	520,280	1,347,519	57,000	86,000
Total Australia	2,967,909	7,686,849†	11,599,000	12,728,000†

*Preliminary. †Figures do not add to total given because of rounding.
Source: Official government figures.

The great paradox of Australia, however, is that in this huge continent with its small population, relatively few people live in the country at all. By 1971 some 60 percent of the population lived in the seven capital cities, 25 percent in smaller urban areas, and only about 15 percent in rural areas. The average density was then four persons per square mile, and the Northern Territory had a population of only 86,000. Against this, 5,000,000 people lived in Sydney and Melbourne. Many Australians regard this balance of the population as unhealthy and deplore the concentration of people in the conurbations of Sydney and Melbourne. Most Australian politicians pay lip service to the principle of decentralization even though they have done little to put it into practice.

Urbanization

Yet there are compensations in this imbalance. Both Sydney and Melbourne are large, modern, sophisticated cities that can compare in the services and amenities they offer with any city in the world except, perhaps, London, New York, and Paris. Although neither is the federal capital, they are still the most important cities in Australia, the natural centres both of business and the arts. After Sydney and Melbourne the biggest cities (1971 census) are the state capitals of Queensland (Brisbane: 817,000); South Australia (Adelaide: 809,000), Western Australia (Perth: 640,000), and Tasmania (Hobart: 130,000).

All of these cities grew naturally as centres of local trade and commerce. Canberra, on the other hand, was artificially created as the federal capital of Australia. Although building did not begin until 1913, by the early 1970s it had become the most rapidly growing city in Australia, with a population in excess of 140,000.

The original plan for Canberra was designed by Walter Burley Griffin, an American architect of the Chicago School, as the result of a competition held for the purpose. Though Griffin's plan was later modified, Canberra is still a conspicuously well-planned city, with broad avenues lined with trees and an artificial lake in the centre of the city. It has often been praised as a model of town planning but, more than half a century after its inception, very few of the important public buildings had been completed.

For all its advantages of site and plan, Canberra has not quite escaped the main criticism levelled against Australian cities—that they have allowed suburbia to run riot. Because of the availability of land and the determination of the average Australian to own his own house and garden, all the main cities stretch for miles around their centres, putting an inevitable strain on public transport and services. The rising price of land, particularly in Sydney, has slowly brought a change to apartment dwelling and higher density housing, but, by the decade of the 1970s, it was still true that most Australians lived in suburbs, and some critics have found in this an explanation for a certain suburban outlook in contemporary Australia.

THE PEOPLE

Origins. The population of Australia is remarkably homogeneous. By the early 1970s some 80 percent were of British stock, though this proportion was declining each year with the increase of migrants from other European countries. (In another generation intermarriage may make this distinction virtually meaningless.) Other Europeans made up the balance, and only one percent of the total population was nonwhite. This included about 130,000 Aborigines, of whom only 46,000 were full bloods. There was also an Asian minority of about 37,000, most of whom were students or visitors, though there were a few thousand Chinese Australians, the descendants of Chinese coolies and diggers who entered Australia during the gold rush of 1851 before the White Australia policy was adopted.

Until the end of World War II almost all immigrants to Australia came from the British Isles. This should not be equated with England. Among these immigrants there was always a high proportion of Scots, Welsh, and especially of Irish. The ancestry of present-day Australians has been estimated as roughly 50 percent English, 20 percent Irish, 10 percent Scottish, and 2 percent Welsh. (The remaining 18 to 20 percent are of non-British origin, largely from continental Europe.) There is a strong Celtic strain in Australia.

After World War II it was decided to try to increase the population of Australia by encouraging migrants from other European countries, including refugees who had been rendered homeless by the war. Missions were sent to many of these countries and Government assistance was extended to European as well as British immigrants. As a result, a steady stream of migrants has come to Australia from Europe since 1945 at an annual rate of from 100,000 to 180,000 a year. About half of these, however, still came from Britain. For example, of the 16.39 percent of the population recorded in the 1966 census as being born in Europe, the distribution was as follows: U. K. and Republic of Ireland, 7.87 percent; Italy, 2.31 percent; Greece, 1.21 percent; with other countries each contributing less than 1 percent.

This migration program has been remarkably successful. On the whole, the European immigrants have been easily assimilated and most apply for Australian citizenship after the statutory five years residence. Few return to their homelands, though, curiously enough, it has been estimated that one of every six British families who came to Australia during this period returned to the United Kingdom.

The White Australia policy

This homogeneity, of course, has been achieved only by enforcing a White Australia policy. Strictly speaking, there is no such policy. The phrase does not appear in any act of Parliament. In fact, however, since 1901 all Australian governments have seen to it that black, yellow, and brown immigrants (politely called non-Europeans) were not admitted to permanent residence in Australia. Since the late 1950s exceptions have been made for the wives, children, and parents of Australian citizens, and since 1966 non-Europeans have been able to apply for admission as migrants. Non-European people, mostly Asians with skilled qualifications, are now settling in Australia at the rate of 3,500 annually; and people of partly non-European descent are settling at the rate of 6,000 a year.

Naturally this policy, based on racial discrimination, has not won the approval of all other nations. It is still, however, supported by all the major political parties in Australia, though recently minority groups have urged that it should be modified still further to allow a substantial quota of Asian immigrants.

Australia has one nonwhite minority that is too often forgotten. That is the Aborigines, of whom about 150,000 had, by the 1970s, survived two centuries of white persecution and indifference. (There is no true definition of an

Aborigine. Only about 46,000 full bloods remained, but anyone with a recognizable amount of Aboriginal blood tended to be counted as an Aborigine.)

Aborigines

These people have had little share in Australia's growth and prosperity. Most of the full bloods are to be found in the Northern Territory, the north of Queensland, and Western Australia, where some of them still live in tribal societies and a few—a very few—still follow the nomadic, hunting life of their ancestors. Others find appropriate employment as stockmen on the cattle stations. But most of them tend to live on mission stations or government reserves, depending on handouts or earning a little money as casual workers.

The same is largely true of the part-Aborigines in the southern states, although they have had greater opportunities for employment. There are no Aborigines in Tasmania, where they were exterminated early in the 19th century, and, by the decade of the 1970s, only an estimated 4,000 in Victoria, 8,000 in South Australia, 24,000 in New South Wales, and 26,000 in Western Australia.

Official policy is to help the Aborigines become an integral part of Australian community life. So far, however, they have been held back by white prejudice or indifference, by their own deep suspicions, and, most of all, by the vicious circle of poverty, ignorance, and disease in which they are trapped. Their birthrate is high, but infant mortality is three and a half times that for white Australian children. Those who survive tend to suffer from malnutrition and other diseases, which gravely handicap their development at school and afterward.

Recently, however, public opinion has become concerned at the plight of the Aborigines, and the federal and state governments have, since the mid-1960s, taken several measures that give some hope of improvement. All acts discriminating against Aborigines have been repealed except in Queensland; Aborigines vote in all federal and state elections. There has been a marked improvement in their treatment and conditions in many areas. With the exception of Victoria, however, no government has yet granted the Aborigines the right to own the reserves on which most of them live.

Major religious groups

If Australia is overwhelmingly white and predominantly Anglo-Saxon, it is far from being overwhelmingly Protestant. The Irish immigrants have seen to that. Sectarian feeling between Catholics and Protestants has at times played an important part in Australian political and social life, and it cannot yet be said to be entirely extinct. From the beginning, the Roman Catholic Church has played an important part in Australia and by the 1970s its adherents numbered some 26 percent of the total population, as compared with some 34 percent for the Anglican Communion.

Demographic outlook. The total population of Australia passed 12,500,000 at the 1971 census. The continued growth of the Australian population must depend on two factors, natural increase and immigration, both of which may fluctuate according to political, social, and economic circumstances.

By 1970 the birth rate was 20.5 per 1,000, the death rate 9.0 per 1,000, and the excess of immigrants over emigrants approximately 125,000. The rate of annual increase, including migration, was 2.15 percent. If one were to assume that all these figures remained constant, then the population of Australia, in round figures, would be approximately 15,500,000 in 1981, 19,000,000 in 1991, and 23,000,000 in 2001. Any variation in the vital statistics figures would, of course, affect the totals. If, for instance, the excess of immigrants over emigrants was increased to 150,000 a year, then the total population would be 25,000,000.

It would, however, be rash to assume that either the birth rate or the immigration rate will remain constant. The crude birth rate began to fall in 1961 and reached a low point of 19.4 per thousand in 1965/66. It then began to rise again and by 1970 was 20.5.

The immigration rate is an even more hazardous subject. The view that prevailed from the end of World War II to the close of the 1960s—that Australia should increase its population as rapidly as possible both to develop its economic resources and to enable it to survive in an increasingly dangerous world—has recently been challenged by those who argue that quality is more important than quantity and point to the dangers of overcrowding and pollution that have threatened so many other nations.

Moreover, even if the view prevails that Australia should seek as many immigrants as it can absorb economically, it cannot be assumed that a constant supply of suitable immigrants will always be available. During the 1960s there was actually a slight decline in the number of migrants coming from the continent of Europe, though this short-fall was made up by migrants from Britain, from North and South America, and by limited immigration from such countries as Turkey, Egypt, and Lebanon.

There would presumably be no lack of prospective migrants if the door were thrown open to immigration from Asia, but this would mean abandoning the White Australia policy and would raise the fundamental question whether Australia is to continue as a white nation of European race, religion, and culture or whether it should deliberately seek to create a multiracial society in which the races of Europe and Asia commingle. (J.D.Pr.)

THE NATIONAL ECONOMY

Australia's traditional world reputation as a predominantly farming country has changed. Its meteoric rise over the 1960s from a mining nonentity to a major mining nation, and the wild fluctuations of mining share prices on Australian stock exchanges in 1969, attracted much world attention. Industrialized countries have realized that beneath this vast continent and its offshore areas lies a wealth of mineral riches. By 1970 Australia had already become one of the chief sources of minerals for the leading industries of Japan.

The minerals boom since the mid-1960s, the impressive progress of industrialization between 1950 and 1970, and the growth of service industries and other features of a sophisticated economy have added new dimensions of economic strength and transformed the nation's traditional image of a predominantly agricultural and pastoral country with vast open spaces into that of a bustling industrial nation. In the Pacific Basin area, Australia plays a vital role in trade, commerce, and investment. It also has one of the world's fastest economic growth rates; since the early 1960s the gross national product has expanded at a faster rate, in real terms, than that of the other major industrial countries, with the exception of Japan. By the 1970s, Australian living standards in terms of per capita income were comparable to those of the advanced economies, ranking third after the U.S. and Canada.

Australia exports to all the important world markets and it imports goods from most countries, particularly the leading industrialized nations.

Resources. *Mineral resources.* Australia is endowed with resources of all the major minerals, particularly coal, oil, natural gas, iron ore, lead, zinc, copper, bauxite, uranium, tin, gold, silver, nickel, and beach-sand minerals. Among these, there were spectacular discoveries in the 1960s of iron ore, bauxite, nickel, oil, and natural gas. Australia's iron ore and bauxite reserves are among the world's largest. Its reserves of lead, zinc, copper, and the rare minerals rutile, zircon, and ilmenite are also very large by world standards.

Although many discoveries have been made, it is generally believed that the surface of this vast continent has only been scratched and that the mineral exploration boom should continue at least into the 1980s with regular and spectacular discoveries adding to the nation's strongly appreciating mineral wealth.

Australia's significant mineral reserves are located in Western Australia (iron ore, nickel, bauxite, gold), Queensland (bauxite, lead, zinc, and silver), New South Wales (coal, lead, zinc, silver, and beach-sand minerals), and off the coast of Victoria (oil and natural gas).

Australia is the world's largest wool producer and a ma-

jor supplier of cereals, dairy products, meat, sugar, and fruit. About 95 percent of the wool and over a third of most of the other rural products are exported. In contrast with most other countries, by 1970 over 90 percent of the utilized land area in Australia remained in its natural state or was capable of very limited improvement; *i.e.*, was land used solely for rough grazing. The area cultivated for agriculture and intensive grazing was slightly more than 8 percent of all the utilized land. The main limiting factor in this respect is lack of water, but unsuitable soil and topography are also important determinants. By the 1970s Australia's 256,000 rural holdings had a combined land area of about 1,210,000,000 acres or about 60 percent of the total land area. (A rural holding in Australia is defined as a piece of land of one acre or more in extent used for the production of agricultural products or for the raising of livestock and the production of livestock products.)

Sheep

Animal resources. The sheep industry accounted for about 46 percent of Australia's total farming area. By 1970 Australian sheep numbers totalled over 180,000,000 head, and they constituted about one-sixth of the world's wooled sheep, producing around a third of the world's wool supplies. Merino sheep constitute the basis of the Australian wool industry and they make up over 70 percent of total sheep numbers. The Merino, which produces wool of a very fine quality, is particularly well adapted to the extreme vagaries of climate in Australia's wool producing areas. Australia's other types of sheep are crossbred and comeback varieties and dual purpose Australian and British breeds, reared mainly for fat lamb production.

Despite the adverse effects of recurrent droughts, Australian wool production more than doubled between 1950 and 1970, with peak seasonal production reaching almost 2,000,000,000 pounds in weight. This rise in production over the period was attributable to a number of factors which included: success of a myxomatosis campaign in the early 1950s; pasture land flock improvement; improved methods of pest and disease control; and better management of sheep farms.

Sheep are run throughout Australia under a wide range of climatic conditions and environments, but about a third of Australia's sheep are reared in those areas with an annual rainfall of less than 15 inches that are generally known as the pastoral zone, where lack of water and suitable fodder usually limits development. The dry conditions in the pastoral zone make the adaptable Merino the main breed. In areas of higher rainfall, up to 25 inches, sheep are often reared in conjunction with wheat and other cereals. About 40 percent of Australia's flock are found in these wheat-sheep zones, and breeds other than Merinos constitute a high proportion of the flocks. The rest of Australia's sheep are reared in the areas of relatively high and reliable rainfall, which produce most of the country's superfine wool.

Cattle

The breeding and fattening of cattle is usually carried out in different climatic zones. The better quality pastures in the relatively high rainfall areas are usually reserved for the fattening of cattle, which are often purchased seasonally and moved to rich natural pastures. Most of Australia's beef cattle are raised in the states of Queensland and New South Wales in regions of warm and coarse pastures. Victoria is the most important producer of mutton and lamb, using crossbred sheep raised in areas of high rainfall and fertility. Most of this type of grazing in Victoria is part of a mixed farming operation, where wheat is often the other principal product.

In the northern parts of Australia herds of the more recently introduced zebu and Brahman cattle are thriving and increasing; they are well adapted to the generally poor tropical pastures and resistant to heat and insects. A recent introduction from the United States, the Santa Gertrudis breed, is also offering good prospects of further development in these areas. Australia's dairy industry is mainly located in the temperate zones of high rainfall in coastal New South Wales and Victoria.

Other agricultural resources. Wheat is the most important grain crop grown in Australia; other grain crops are barley and oats. Wheat is usually grown in the medium rainfall belt in all states and it has become increasingly integrated with sheep grazing and cultivation of other crops. Wheat, barley, and oats are often grown on the same farm for grain and green fodder or hay for livestock. Most of these cereal crops are grown for grain. Sugar, Australia's most important crop after wheat, is grown from cane in coastal areas of Queensland and in northern New South Wales. The former area accounts for over 95 percent of the Australian sugar crop. The sugar industry is based on small farms, and covers an area exceeding 500,000 acres.

Other important crops grown in Australia include: tobacco, citrus fruit, grapes, apples, cotton, bananas, potatoes, and sorghum.

Forest and fishery resources. Australia is estimated to have about 60,000,000 acres of commercial or potentially commercial areas of forest, in addition to large regions of low grade forest suitable only for the production of small quantities of forest products for use in the immediate vicinity.

The main commercial forests are in the areas of high rainfall on the coast or near the coastal highlands of southeastern and eastern Australia, Tasmania and the southwest coast of Western Australia. The main types of trees in Australian forests are eucalyptus, a broad-leaved genus which provides timbers of great strength and durability for building and packaging. The rain forests of the wetter regions are the other important type with many important species of broad-leaved trees.

Australia's annual fish catch exceeds 120,000,000 pounds weight, not counting lobsters, crayfish, crabs, and prawns. It has been estimated that there are about 2,000 species of fish (including freshwater species) in Australia and the waters surrounding it. The resources include mullet, cod, bream, perch, tuna, snapper, whiting, flathead, abalone, mackerel, and Australian salmon. New South Wales and Western Australia are the most important producing states, the latter being famous for its crayfish. There has not been much change in the types of fish caught in Australian coastal waters over the years, but lobster and crayfishing has been expanding strongly in recent decades.

Hydroelectric and other power resources. By the early 1970s Tasmania was one of the few states with sufficient water resources to permit the continuous operation of large hydropower stations. Its hydroelectric resources have been estimated at about 50 percent of Australia's total. Although the use of hydropower is expected to grow, the country's potential for this form of energy is small in relation to its total area.

Coal is the most important source of energy in all the other states of Australia. About 71 percent of the total installed electric capacity in Australia is thermal power equipment. By 1970 the installed generating capacity of the main public supply electricity systems totalled almost 12,000,000 kilowatts, about 60 percent of which was located in the highly industrialized states of New South Wales and Victoria.

Sources of national income. *Agriculture.* Although the rural sector still makes a substantial contribution to Australia's national income its importance declined sharply during the 1960s. By the end of the decade the gross annual value of rural production totalled about $A2,230,000,000, or about nine percent of gross national product, a substantial drop compared with a 15 percent contribution to GNP in 1959. Rural industry generally was in a very depressed state by the early 1970s, the exceptions being such sectors as beef, cattle, sugar, and vineyards. The wool, wheat, and dairying industries have been hardest hit by falling world prices for their products and lack of export markets. The industries are in such a depressed condition that the Australian federal government was, by the early 1970s, considering measures to reorganize and rationalize them. These would involve the consolidation of numerous small uneconomic farms into more efficient units, and also the curtailing of production of such crops as wheat because of the inability of disposing of all the production.

Depression in rural industry

Although rural production has increased since World War II at twice the rate of increase in population, nonrural industries, particularly manufacturing, mining, and the service sectors of the economy have shown much faster growth rates and substantially increased their contributions to national income. Until 1969 wool was Australia's biggest single export earner, but it was subsequently displaced by minerals. In the latter portion of the 1960s, the value of wool exports nevertheless averaged about 27 percent of the total value of Australia's merchandise exports, with total annual value of wool exports approaching $800,000,000.

By the end of the 1960s Australia's wheat production totalled 540,000,000 bushels, compared with an annual average of about 180,000,000 bushels a decade earlier. Wheat is Australia's second most important rural crop but, by the early 1970s, the country faced a perhaps temporary crisis of overproduction following the lack of orders from the most important export customer, the Chinese People's Republic. Forestry and fishing contribute only an insignificant proportion, less than one percent, of Australia's gross national product.

Mining and quarrying. The story of Australia's mineral industry is one of strong expansion and development. New discoveries now being developed or yet to be exploited, particularly iron ore, bauxite, and nickel, indicate that minerals will play an increasingly important role in the Australian economy. The contribution to national income and exports from the nation's booming mineral industry is increasing at a spectacular rate. The total value of mineral production for 1970, exceeding A$1,000,000,000, compared with a 1963 figure of only A$416,000,000. The two most important contributors to mineral output are the coal and iron ore industries, which accounted for about 25 percent and 15 percent respectively of the total value of mineral production by 1970. Other minerals of economic significance are lead and zinc, the construction materials group (road metal, gravel, clays, limestone and building stone), copper, gold, tin, limestone, clays, salt, and such rare minerals as zircon.

Minerals are already Australia's largest single source of export income, and by 1970 mineral exports totalled over A$950,000,000 or about a quarter of total export income, compared with A$342,000,000 or about 12 percent of the total, as late as 1967. If the trends at the start of the decade were any indication, mineral exports might well rise to A$2,000,000,000 a year by the late 1970s.

Manufacturing industry. A vital sector of the Australian economy, manufacturing industry, has grown spectacularly since 1950, and by 1970 accounted for over a quarter of the nation's gross national product. Australia's limited home market will not support certain types of industry, such as the manufacture of highly sophisticated machines, heavy fabricating, certain types of electrical equipment, some industrial chemicals, and highly specialized scientific instruments and equipment. With the exception of these special classes of goods, however, the domestic market is well catered for with products of local manufacturers, which include the full range of items required in a modern community. Most of the industries are soundly based; only a few are carried out at an excessive cost, and require high tariff protection.

About a quarter of the nation's workforce, or around 1,300,000 people is employed in the 63,000 factories of Australia's manufacturing industries. The value of production of manufacturing industries has increased from A$406,000,000 around the year 1940 to a 1970 annual figure of over A$7,500,000,000. New capital expenditure by manufacturing industry was running at an annual rate of A$800,000,000 by the early 1970s, about double the rate a decade earlier.

Since World War II the expansion of manufacturing industry has been widely spread throughout the industrial structure, but it has been particularly strong in the engineering, vehicle, chemical, and construction materials industries.

The motor vehicle industry is a vital sector of the Australian economy and numerous other industries are dependent upon demand created by its level of activity. Its importance is underlined by the fact that Australia has the world's third highest per capita motor vehicle ownership, while most of these vehicles are locally produced or assembled. Total investment in the motor vehicle industry was estimated to be worth about A$600,000,000 by 1970, and it employs about 12 percent of the nation's workforce.

The motor vehicle industry

Iron and steel production was for long a virtual monopoly in the hands of one company, the Broken Hill Proprietary Co. Ltd., but there were plans for another producer to enter the market in the 1970s. Other important manufacturing industries include: oil refining, chemicals, textiles, and domestic appliances.

Energy. The main source of power used in Australian homes, factories, and other buildings is electricity. There was a strong rise in domestic use of electricity over the 1950s and 1960s because of the increased use of domestic appliances, extension of electricity supply into rural areas, and the high rate of home building, which has been running at an annual rate of over 100,000. Between 1958 and 1968, for example, the annual domestic consumption of electric power per consumer rose by about 50 percent. Thermoelectric power, based on either brown or black coal, is the main source of electrical energy in Australia.

Financial services. Australia has a well-developed banking system broadly similar to that operating in Britain. The system is made up of the Reserve Bank of Australia (the central bank), the trading banks, the savings banks, the Commonwealth Development Bank of Australia, and the Australian Resources Development Bank.

The federal government-owned Commonwealth Banking Corporation operates a trading bank, a savings bank, and a development bank. Some of the savings banks are owned by state governments. These savings banks operate within the borders of a particular state only and some offer checking-account facilities. The seven major trading banks and their savings-bank subsidiaries form the backbone of the Australian banking system.

The major banks in Australia operate under a branch-banking system, and most of them have branches in all cities and towns in Australia. The Australian Resources Development Bank was set up to provide finance mainly for the development of Australia's mineral projects. One of its aims is to provide finance in order to maintain a substantial proportion of Australian equity in these mineral enterprises. The Australian government planned to set up an organization known as the Australian Industries Development Corporation by the mid-1970s. This was expected to perform functions broadly similar to the Resources Bank.

By the 1970s there were over 4,600 branches and 1,900 agencies of check-paying banks in Australia.

The banking system in Australia is an important repository for savings, the most important source of finance, and the medium through which most commercial and financial transactions are settled.

The trading banks provide two types of facilities: current accounts operated on by checks and payable on demand, and fixed deposits, which are lodged for specific periods. By the start of the 1970s monthly deposits of the check-paying banks totalled some A$6,400,000,000 and deposits of the savings banks totalled almost A$7,000,000,000. The Reserve Bank applies its monetary policy to regulate the economy through the banking system, which usually bears the brunt of monetary and credit restraints during periods when the authorities consider it necessary to dampen down inflationary pressures in the economy (see below *Management of the economy*).

The next most important sources of finance in Australia are the finance companies, which grew strongly in the decades of the 1950s and 1960s, largely as a result of the restraining influence of official monetary policy on the banking system. All of the major Australian banks have substantial shareholdings in finance companies that in some instances are subsidiaries of the banks. In recent years a number of foreign banks also have acquired inter-

Finance companies

ests in finance companies in order to enter the Australian financial market, because Australian government policy prohibits the establishment of foreign banks in Australia, the exceptions being the long-established ones.

Australia has a well-developed short-term money market. There are also several merchant banks performing such functions as underwriting, company flotations, mergers and takeovers of companies, portfolio management, lending, and general financial and allied services.

Other important institutions in the financial system are the insurance companies (fire, marine, and general), life insurance companies, pastoral finance companies, and the building societies.

Besides funds available for expansion from undistributed profits, Australian companies raise large sums, averaging about A$600,000,000 a year by the start of the 1970s, by the issue of shares and fixed-interest borrowing.

Each state capital has its independent stock exchange, but a nationwide coverage is achieved by reciprocal arrangements with brokers in other capitals and agents in country centres.

The system is well-developed and sophisticated, and the stock exchanges provide a ready market for shares of listed companies and government securities. Quotations are distributed widely and quickly by special teleprinter channels and by radio and newspaper services.

Foreign trade. By 1970 Australia's exports made up about 14 percent of the gross national product; they had been growing at an annual rate of around 9 percent in previous years. At the start of the 1970s total annual export income was estimated at more than A$4,100,000,000, with imports around A$3,900,000,000. Minerals contributed about a quarter of Australia's export income; other important sources of export income were wool, wheat, meat, and manufactured goods. Japan was Australia's most important export customer, accounting for about a quarter of Australian exports, in value terms. Other important customers were the United Kingdom, the United States, the Common Market countries, New Zealand, and the rest of Asia, as a whole.

Management of the economy. The Australian economy is essentially one based on private enterprise, although government plays an important role in influencing economic conditions.

Most capital expenditure in Australia is undertaken by private business enterprises and persons. There is only limited public ownership of business enterprises. Major enterprises owned and run either directly by federal or state governments or by government corporations include the post office, Australia's international airline, one of the two major domestic airlines, the Commonwealth Banking Corporation, the Australian National Shipping Line, and the electricity and gas distribution organizations in the various states.

The government regulates economic activity through fiscal policies embodied in its annual budget and through monetary policy, which is administered by the Reserve Bank. The Reserve Bank and the federal government's Treasury Department are the government's chief source of economic advice.

An important feature of the Australian economic scene has been the strongly rising trend in government and other public authority expenditure. The chief source of this expenditure is the federal government, whose annual budget expenditures since the early 1960s have been running at an annual rate in excess of A$6,000,000,000. Public authority spending comprises over a quarter of gross national expenditure. While the private sector remains the dominant segment of the Australian economy, the public sector has grown strongly.

Taxation

In Australia taxes are levied by federal, state, and local governments. The main taxation authority is the federal government, which levies income tax, customs and excise duties, sales tax, payroll tax, estate duty, and gift duty. There are also a number of minor taxes imposed for specific purposes. There is no general duplication of taxes by the three tiers of government. The states impose a wide variety of taxes that include stamp duties, taxes on motor vehicles, land taxes, and probate duties.

The trade-union movement is well organized in Australia. Most workers are members of trade unions; in a number of instances membership is virtually mandatory. Industrial disputes are common and often disruptive in terms of loss of production. Australia has a unique arbitration system for settling disputes, but there have been signs of this system breaking down, and the trend is increasingly toward direct negotiations with employers (see below *Administration and social conditions*).

Contemporary economic policies

During the 1960s the principal aims of government economic policy were relatively high economic growth, full employment, stability of costs and prices, and balance of payments equilibrium. By the early 1970s there had been a considerable degree of success in the achievement of these aims. For instance, Australia had not experienced a major recession in the eight years following the one in 1961–62. Over that period the Australian economy grew at an annual rate of about 5 percent in real terms, employment was maintained at relatively high levels, the annual rate of inflation (about 2¾ percent) was lower than in most industrialized countries, and there had been no serious balance of payments crisis. Australia emerged from the 1960s and entered the 1970s with new dimensions of economic strength stemming from the continued powerful growth of its mineral industry. The balance of payments was very much stronger than it had been around 1960, export income had risen strongly, exports were more diversified with a greatly reduced reliance on rural exports, export markets were also more diversified, and there were prospects for further strengthening of the balance of payments.

Problems and prospects. During the 1970s Australia has the potential for achieving a rate of economic growth unsurpassed in any decade since World War II. Gross national product is expected to grow at an annual rate of at least 6 percent in real terms, compared with about 4.8 percent per annum during the 1960s. The balance of payments has been, and will continue to be, the key factor determining the rate at which the Australian economy can expand. Since the 1950s it has been the main limiting factor. During the 1970s, however, it is not generally expected to be an important obstacle to a high rate of economic expansion because exports of minerals and manufactured goods will almost certainly grow strongly. Also, domestic oil production should lead to a substantial savings in the import bill. While Australia will continue to depend on overseas capital inflow to balance its external payments during the 1970s, the extent of the dependence is likely to be considerably lessened. Mineral export earnings should be contributing over a third of export income by the late 1970s, compared with about 24 percent in 1970 and less than 10 percent about five years previously.

All told, as Australia entered the 1970s, its economic outlook was bright. (E.I.U.)

TRANSPORTATION

Overall patterns. Because of the great size and small population of Australia, transport has always been costly and has absorbed an unduly high proportion of the total work force. Moreover, the main lines of road and rail transport were laid down in the second half of the 19th century, when Australia was a collection of separate colonies, each of which looked to Britain for most of its trade. The transport system was therefore designed to maintain this trade with roads and railways radiating from the main ports. Little thought was given to internal transport between the separate colonies. An unfortunate relic of this is the existence of three different railway gauges in the continent. It was not until 1970 that it became possible to go by train from Sydney on the east coast to Perth in the west without changing trains.

Fragmentation of transport system

In spite of these historical and geographical handicaps, Australia is fairly well equipped with roads and railways, especially in the southeastern states. It has some 2.7 miles (4.3 kilometres) of railway per 1,000 persons as against 2.1 miles (3.4 kilometres) in the United States, and more than 350 motor vehicles per 1,000 as against 280 in the United States. It is one of the most highly motorized countries in the world.

Australia is almost entirely lacking in internal waterways. For a short period during the 19th century the Murray–Darling river system was widely used to transport produce (mostly wool) from the country districts of New South Wales and Victoria to the coast, but the variable flow of water in these rivers always made this method hazardous and unreliable. With the coming of the railways it was quickly abandoned.

On the other hand, Australia is in many ways ideally suited to air transport. The great distances, lack of mountains, and fine prevailing weather makes flying safe and economical. Australia took to the air age with enthusiasm. By the 1970s Australia's internal airlines carried about 6,000,000 passengers a year.

For the historical reasons already outlined, roads and railways are mainly the responsibility of the six state governments, while shipping and air transport between the states is the responsibility of the federal government. Since 1946 the Australian Transport Advisory Council, including the federal minister for shipping and transport and the six state ministers for transport, has provided machinery for coordinating transport on national lines.

Component systems. As has been noted above, the main road networks in Australia radiate from the ports, and especially from the capital cities of Sydney, Melbourne, Brisbane, Adelaide, and Perth. In recent years most of the states, aided by lavish federal grants, have also made remarkable progress in sealing thousands of miles of country roads. By 1970 Australia had some 550,000 miles (885,000 kilometres) of roads and tracks. Of these, slightly in excess of 117,000 miles (188,000 kilometres) were paved and sealed with concrete or bitumen, about 133,000 miles (214,000 kilometres) were paved but unsealed, and some 300,000 (483,000 kilometres) were merely earth roads graded to form a reasonably smooth surface.

The most serious lack is still adequate highways between the state capitals. The Hume Highway between Sydney and Melbourne and the Pacific Highway between Sydney and Brisbane were both far too narrow for the heavy traffic they were carrying in the early 1970s. The only road across the continent from east to west, which links Adelaide and Perth, is still a hazardous adventure for private cars. The only road across the continent from north to south, between Darwin and Adelaide, is not much better.

The provision of adequate expressways and throughways in the great conurbations of Sydney and Melbourne has also fallen behind. Like the industrialized nations of the West, Australia is finding it hard to keep up with the insatiable demands of the automobile.

Railways

The state governments own and operate their own railway systems. The Commonwealth government operates the Trans-Australia Railway, the Central Australia Railway, and the North Australia Railway. Vast sums of money were spent on building these railways in the latter half of the 19th century. By the end of the 1960s the total capital invested in fixed assets by all government railways was approximately $A1,577,000,000.

Today all these railways have a hard struggle to pay their way against the competition of road and air services, but rail transport still plays an important part in the development of the Australian economy. New capital is being spent in providing standard gauge lines between the different state capitals and industrial centres and in building new branch lines where mineral discoveries justify it, as in the case of northwestern Australia. Sydney is also extending its underground and suburban railway system built at the turn of the century.

There are about 66 ports of commercial significance in Australia, of which the great majority are on the east coast. The rest of Australia is notably lacking in good natural harbours. The most important port is Sydney, which has one of the finest harbours in the world. By 1970 Sydney (with Botany Bay) discharged and shipped some 17,000,000 freight tons annually, followed in order by Melbourne, Newcastle, Port Kembla, Fremantle, Geelong, Whyalla, Brisbane, and Port Adelaide. In addition, semi-artificial harbours were being constructed on the northwest coast to handle shipments of iron ore from Western Australia.

Shipping is still the lifeblood of Australia, which at the start of the 1970s imported some 28,200,000 tons and exported some 77,000,000 tons annually. Most of this shipping was in foreign hands, although Australian companies had a monopoly of interstate trade around the Australian coast. Australian registered tonnage represented only 0.9 percent of the total tonnage entering Australian ports from overseas, but the Australian National Line was attempting to extend its share by the purchase of new ships.

Air transport

Australia has one international airline, Qantas Airways, Ltd., which is owned by the federal government. It is one of the world's leading air carriers. By 1971, 17 other international airlines operated regularly in and out of Australia. Sydney has long been the site of the main international airport and a new terminal was opened there in 1970. The same year Melbourne opened a new international airport at Tullamarine.

In the same period ten airlines operated domestic services throughout Australia and Papua-New Guinea. The two major companies are Trans-Australia Airlines (TAA), owned by the federal government, and Ansett-Airlines of Australia (Ansett), owned by private enterprise. Competition between the two is strictly controlled and regulated by federal legislation. Australia is well served by its internal airways, which offer a service that compares favourably in safety, cheapness, and extent with any in the world.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. The constitution of Australia may be described crudely as an amalgam of the constitutional forms of Great Britain and the United States. Like Britain it is a monarchy, and the king or queen of Britain is also the king or queen of Australia. Like Britain, too, the governments both of the Commonwealth and of the states are chosen from the majority party in Parliament. Like the United States, on the other hand, it is a federation, and the duties of the federal government and the division of powers between the Commonwealth and the states are laid down in a written constitution that can be altered only by a referendum that gains the consent of a majority of all the electors and a majority in at least four out of the six states.

Any disputes arising out of the constitution are decided by the High Court of Australia.

Although the queen of Britain is also queen of Australia, Australia is wholly independent. Except when the queen is in Australia, her functions, which, by the 1970s, were almost entirely formal and decorative, are exercised by a governor-general, who resides in the federal capital in Canberra, and by governors in each of the six states. Though formally the governor-general and the governors are appointed by the queen, they are invariably recommended by the Australian governments concerned, and in recent years there has been a growing tendency to choose Australians.

The constitution defines the form and duties of the federal government in some detail. The most important of these are defense, foreign policy, immigration, customs and excise, and the Post Office. Those powers not given to the federal government in the constitution—the residual powers—are left to the states. The state governments are the direct descendants of the colonial governments of the 19th century. Originally they owed their powers to an act of the British Parliament, but these acts are short documents that lay down only the form that the parliaments shall take, not the sort of legislation they may pass.

State and territorial government

In theory the states are sovereign, with real and important powers. They are responsible for justice, education, health, internal transport, and, indeed, for most of the things that most closely concern the citizen. In practice, however, their powers have been greatly weakened by the federal government's decision in 1942, at the most dangerous point of World War II, to collect all income tax itself and then to reimburse the states on an agreed formula. This temporary arrangement has become permanent,

in spite of many political and legal challenges, so that today the states, deprived of any equivalent "growth tax," have become largely dependent on federal grants for their revenues.

The federal government directly administers the internal territories (the Northern Territory and the Australian Capital Territory) and the External Territory of Papua-New Guinea through resident administrators. The Northern Territory has a legislative council consisting of some elected and some nominated members. Papua-New Guinea has an elected House of Assembly as the first stage toward self-government.

Australia is a true parliamentary democracy. Both the federal upper house (the Senate) and the lower house (the House of Representatives) are directly elected by universal adult suffrage, with a voting age of 21. All state lower houses are similarly elected.

The political process

Preferential voting (placing the candidates in order of preference on the ballot paper) is used for all Parliamentary elections. In elections for the House of Representatives the alternative vote system is used; the 60 senators (10 from each state) are elected by proportional representation. Voting in both federal and state elections is compulsory.

Since federation the political struggle for power in Australia has been between the Australian Labor Party and a coalition of anti-Labor parties under different names. For more than two decades after 1949 the federal government was formed by a coalition between the Liberal Party, which broadly represents the interests of private enterprise, and the Country Party, which represents the interests of the farmers and graziers in rural constituencies. Australia entered the 1970s, for example, with the Liberal Party holding 46 seats in the House of Representatives, the Country Party 20, and the Australian Labor Party 59.

The Australian Labor Party is a typical Western social-democratic party, based firmly on the support of the trade unions, and normally preferring practical reforms to socialist theories. It has, however, always had a left-wing retaining a belief in Marxist principles.

There is one other political party of some significance. In 1956 the right-wing of the Australian Labor Party split away, largely influenced by Roman Catholic fears of Communist influence on the ALP (Australian Labor Party), and formed the Democratic Labor Party (DLP). The DLP, which depends chiefly on Catholic votes in Victoria, Queensland, and New South Wales, failed to win a seat in the House of Representatives by the end of the ensuing decade, but entered the 1970s holding four seats in the Senate. By exploiting the system of preferential voting it had also been able to prevent the Australian Labor Party's return to power.

The Australian Communist Party is insignificant in numbers and has no members in the federal parliament. It has retained a disproportionate influence in some trade unions.

Voting in state elections does not necessarily follow voting in federal elections. During the long rule of the Liberal–Country Party coalition in the federal parliament after 1949, Labor governments at different times held office in five of the six states. But with a few minor exceptions, the lines of battle are the same though local issues play a larger part. State Liberal and Labor governments once elected, though on different sides politically, often find themselves in alliance against the Commonwealth either in defending states' rights or in seeking more money from the federal government.

Justice. The law of Australia is based on the common law of England, and many laws are identical with those laid down in acts of the British Parliament. The administration of the law is largely in the hands of the states, each of which has a series of courts culminating in a state supreme court. Between them these courts have a comprehensive jurisdiction that extends to all matters of state, and to most matters of federal, jurisdiction. The states are also responsible for the police.

The High Court of Australia, established by the constitution, is the federal supreme court. It has a general appellate jurisdiction over all other federal and state courts, and has a special duty to decide disputes involving the interpretation of the constitution. The High Court of Australia has a high reputation among legal authorities both inside and outside Australia.

The federal Parliament has created two other federal courts, the Federal Court of Bankruptcy and the Commonwealth Industrial Court. Each federal territory has its own judiciary.

The armed forces. Australia has a proud military tradition dating from the landing of the first Australian Imperial Force on the Gallipoli Peninsula during the Dardanelles campaign of World War I. Since then Australian armed forces have served with distinction in World Wars I and II, in Malaysia, Korea, and Vietnam. The Royal Australian Navy, the Army, and the Royal Australian Air Force are separate services with distinctive uniforms and separate commands.

The Royal Australian Navy is a small but compact, flexible force with a special emphasis on antisubmarine warfare. By 1972 it had one operational aircraft carrier, three guided-missile destroyers, five Daring class destroyers and six destroyer escorts. Four new Oberon class submarines had been delivered.

The first aim of the Army is to provide a highly trained field force for limited warfare in Southeast Asia. Compulsory national service was introduced in 1964. Young men in their 20th year are selected by ballot. Recruits serve two years full-time followed by three years in the reserve. By 1970 the strength of the Regular Army including National Servicemen was 44,000. The Australian Army has made a special study of jungle warfare in which its reputation is undisputed.

The Royal Australian Air Force has three bomber squadrons, four fighter squadrons, five transport squadrons, two utility helicopter squadrons, and two maritime reconnaissance squadrons. By the early 1970s the fighter squadrons were equipped with Mirage and Avon-Sabre fighters, largely built in Australia. Twenty-four F-111C strike-reconnaissance bombers had been ordered from the United States to replace obsolete Canberra bombers.

Social services. Australia may be described as a modified welfare state. The federal and state governments have not yet gone so far as some European countries in providing free education, health, and other services for its citizens but have gone further than the United States. Government benefits include pensions for the aged, invalids, and widows; unemployment and sickness benefits; maternity allowances; and child endowment. Benefits are paid from the National Welfare Fund, which is financed from consolidated revenue. By 1970 welfare services accounted for some 20 percent of the federal budget.

On the other hand, there is a strong tradition in Australia one should buy or build his own house. There are relatively few state or municipal houses and apartments for rent, though there are various federal and state financial schemes to help citizens purchase their own homes.

Education

Responsibility for education lies primarily with the states. Free education at government schools is provided at the primary and secondary levels, though there is also a considerable private sector formed by schools run by the churches (mostly Roman Catholic) and a few grammar schools for children of the wealthy. The principle of giving some state aid to Roman Catholic schools has now been accepted by all the main political parties.

Tertiary education is not free, but the federal government awards Commonwealth scholarships to some 8,500 university students each year and another 1,500 to students at other colleges of advanced education. The federal government also makes grants toward the recurring costs of universities. Australia now has 14 universities and two university colleges. The total enrollment had exceeded 100,000 students before the end of the 1960s.

Health services are everywhere excellent. There is no national health service, but the federal government pays a portion of a citizen's medical expenses providing he insures with a recognized insurance organization. Dental care is at present outside this medical benefits scheme.

Australia's most interesting contribution to the welfare

Wage fixing

of the ordinary man is the arbitration system, which has roused much interest in other countries. In brief this is an attempt, unique to Australia and New Zealand, to fix wages and working conditions by law.

The Constitution gave the federal government the right to undertake conciliation and arbitration in industrial disputes. Armed with this power, the first federal government set up the Commonwealth Conciliation and Arbitration Commission, which began by establishing a "basic wage" necessary to keep in "frugal comfort" the average family of an unskilled worker.

From this beginning an elaborate system has been built up, involving both federal and state arbitration courts, conciliation commissioners, and wage boards, the aim of which is to prevent or mitigate industrial disputes.

If a dispute cannot be solved by collective bargaining or conciliation, then either the employers or the trade union concerned may take the dispute to the relevant court for a judicial decision that has the force of law. Strikes are not forbidden, but a union that strikes in defiance of a judicial award may be held to be in contempt of court and fined accordingly. In practice, therefore, the judges on the Commonwealth Conciliation and Arbitration Commission, after hearing argument from both sides, fix minimum wages and conditions over a large section of Australian industry.

This system has frequently been criticized as excessively cumbrous and, more recently, complaints have been heard that the judges who make the awards are not always equipped to appreciate the economic consequences of their decisions. On the whole, however, it has worked reasonably well over the years though it has certainly not prevented industrial unrest in Australia. It is still supported, with reservations, by both employers and trade unions.

Both Australia's welfare services and arbitration system spring from a deep concern for the common man. From the earliest days Australians have been strongly egalitarian in outlook, quick to resent any claims to privilege either by a class or an individual. This has not prevented class distinctions in Australia—there is a stronger class system than is often admitted—or wide differences in wealth, but it has greatly eased the conflicts inherent in a capitalist society. Australian trade unions are as militant as any in the world in the pursuit of higher wages and better conditions, but the Australian working man entered the 1970s well paid, well cared for, and living in a prosperous, democratic, and expanding country. He was widely judged to be no wage-slave, but rather the master of his own fate.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. For the first 100 years of white settlement the arts were naturally neglected. Men were too busy exploring and developing this harsh land to have much time or energy left for the graces of life. This neglect was never total. There were always Australians who wrote poems or novels or who painted the landscape of their adopted country, but for the most part they were content to take their art and culture secondhand from England.

Early developments in the arts

The first sign of an Australian consciousness in the arts was the emergence of a small group of writers and artists in the 1890s associated with the Sydney *Bulletin*. Most of them—though not all—tended to express the radical, egalitarian, and nationalist views that were then beginning to stir political life in the colonies. Of these the most important were Joseph Furphy (1843–1912), author—under the pseudonym Tom Collins—of the long novel *Such Is Life* (1903), and Henry Lawson (1867–1922), whose short stories are still worth reading. The *Bulletin* also provided an opportunity for a remarkable collection of artists working in black-and-white, amongst whom Norman Lindsay, who lived until 1970, was the best known.

But Australia's great leap forward in the arts did not take place until World War II, when the growing wealth and sophistication of the cities, and isolation from Britain and Europe, provided a powerful stimulus. This movement began with painting and poetry and later spread to all the arts, with the possible exception of the theatre. Since 1954 the federal and state governments have subsidized the arts on an increasing scale. Both, however, tended to impose a censorship, partially on imported films and books, that was unnecessarily rigid when judged by contemporary standards overseas. In spite of this restriction, by the 1970s the cultural scene was vigorous and lively, though only a few Australian painters, writers, and musicians had won international recognition.

The arts today. The first of the arts to attract attention outside Australia was painting. Starting in Melbourne, a group of painters—of whom George Russell Drysdale, Sidney Nolan, and Arthur Boyd are the best known—developed an original school of Australian painting by adapting contemporary techniques to depict Australian myths and Australian landscape. Since then the influence of that school in Australia itself has dwindled, and young Australian painters now prefer to paint in the international styles of their contemporaries in Paris, London, and New York; but their work is marked by a freshness and vigour characteristic of their young nation. The number of galleries showing contemporary work in Sydney and Melbourne is a remarkable proof of this vigour.

Australian poets have also revealed a new maturity and sophistication, though, unlike artists in other fields, they have tended to turn their backs on the more revolutionary developments overseas and to prefer traditional metrical forms. This may be due to the powerful influence of Alec Derwert Hope and James McAuley, both professors of literature as well as poets, who have consciously set and maintained classical standards in their own work. Kenneth Slessor and Judith Wright are two older poets who have written movingly on Australian history and landscape.

Australian novelists have, on the whole, been less original and creative, though Patrick White (1912–) has won world acclaim for his lyric and visionary contemporary novels *The Tree of Man* (1955), *Voss* (1957), and *Riders in the Chariot* (1961).

The performing arts are more dependent on financial support and suffer from the fact that outstanding artists can always go overseas to Britain or the United States where a wealthier market offers greater rewards. This is particularly true of actors and singers. Australia seems to produce an inexhaustible supply of fine voices, many of whom, from Dame Nellie Melba to Joan Sutherland, have won international fame. In recent years, government subsidies have enabled Australia to maintain excellent national ballet and opera companies. The Australian Broadcasting Company also maintains symphony orchestras in all the capital cities, of which Sydney's is probably the best. By the early 1970s there was reason to hope that the completion of the Melbourne Victorian Arts Centre and Sydney's spectacular Opera House (designed by the Danish architect Jørn Utzon) would enable these cities to become international centres for opera, ballet, and theatre.

The Australian popular arts follow too closely developments in Britain and America to deserve special mention, and the remarkable Aboriginal arts, notably dancing and painting on bark, must be considered dead though efforts are being made to preserve and revive them.

Cultural institutions. The first serious attempt to organize the arts in Australia was the formation of the Elizabethan Theatre Trust in 1954. It was this body that formed the Australian National Opera company and Australian Ballet company. Since then the main work of advising the federal and state governments on the performing arts has been taken over by the Australian Council for the Arts, which was formed in 1967. It is this body that is responsible for deciding how the federal government's grant (A$2,750,000 annually by the end of the decade) should be divided among the various claimants. It has been the policy of the Australian Council for the Arts to give the bulk of federal subsidies to national companies, like the ballet and opera companies, and to encourage the states to accept responsibility for drama and music. Other official bodies are the Australian Broadcast-

The Australian Council for the Arts

ing Commission, the Music Board, the Commonwealth Art Advisory Board, and the Commonwealth Literary Fund.

The state governments have long provided museums and art galleries in their respective states. In 1967 the federal government decided to build a National Gallery in Canberra to house the national collection. Since it is accepted that it is now impossible for such a gallery to compete with the great galleries of Europe and America in collecting European painting and sculpture from the past, this gallery will concentrate on Australian art, past and present, art of the Pacific and Asian areas, and art on a worldwide basis beginning with the 20th century.

Press and broadcasting. The press of Australia is free, independent, competitive, and vigorous. It has long reached high standards in such papers as *The Sydney Morning Herald* and the Melbourne *Age*, both of which are over 125 years old, though Australia has also always had and still has popular papers that are more noted for sensationalism than for sober reporting or informed comment. All of these papers circulate only in the states in which they are published. In 1964 *The Australian* was first published with the aim of becoming a national paper. It is printed in several states and has won a small but influential readership. The Australian *Financial Review* also has a national readership.

Broadcasting and television are shared between the Australian Broadcasting Commission (ABC), a national service dependent partly on fees for licenses and partly on federal government grants, and a number of commercial radio and television stations operating under licenses granted by the postmaster-general and dependent on advertising for their revenues. In general it may be said that while the ABC provides a high quality service, especially in the broadcasting of music, it attracts only a small part of the available audience. The commercial stations, on the other hand, attract large audiences with rather second-rate programs. Commercial television in Australia is heavily dependent on material imported from Britain and the United States.

CONCLUSION

Australia has been called "the lucky country." It is indeed fortunate in its superb climate (in spite of drought), its peaceful history (it is the only continent where there has never been a war), its natural resources, and the unity and uniformity of its people. More recently Australians have also begun to recognize the exceptional privilege of living in a largely unspoiled country where there is still fresh air to breathe and clean seas to swim in. Australia has not, perhaps, escaped the reverse side of this good fortune—a certain complacency and mental laziness—but time can be counted on to cure these faults.

For the future is not likely to be as tranquil as the past. Nothing can now disguise the essential fact that Australia is the last outpost of European civilization in the South Pacific. Now that both Britain and the United States are withdrawing from Southeast Asia, it is left facing problems of great difficulty and complexity.

The first of these is whether it should involve itself in the wars and politics of its neighbours in Southeast Asia or should stand aloof, relying on its distant British and American allies, its own strength and, perhaps, on nuclear weapons. So far Australia has chosen involvement, though not without great hesitation.

The second is whether it should try to maintain its racial purity as a white nation or deliberately seek assimilation with Asia by taking in increasing numbers of Asian immigrants. No one can pretend that these are simple questions to answer. Yet Australians have courage, resourcefulness, and adaptability, which should enable them to meet any dangers and difficulties that may arise.

BIBLIOGRAPHY

General: C.M.H. CLARK, *Australian Hopes and Fears* (1958); W.K. HANCOCK, *Australia* (1930); C. MCGREGOR, *Profile of Australia* (1966); J.D. PRINGLE, *Australian Accent* (1958). The best descriptions of Australia's landscape may be found in the novels, *Kangaroo* by D.H. LAWRENCE (1950); and *Flying Fox and Drifting Sand* by F.N. RATCLIFFE (1938).

History: C.M.H. CLARK, *A History of Australia*, 2 vol. (1962–68).

Constitution: J. QUICK and R.R. GARRAN, *The Annotated Constitution of the Australian Commonwealth* (1901); G. SAWER (ed.), *Federalism* (1952).

Government and politics: L.F. CRISP, *The Parliamentary Government of the Commonwealth of Australia* (1949); J.D.B. MILLER, *Australian Government and Politics* (1954); L. OVERACKER, *The Australian Party System* (1952); G. SAWER, *Australian Government Today* (1952).

Economics and development: J.B. CONDLIFFE, *The Development of Australia* (1964); P.H. KARMEL and M. BRUNT, *The Structure of the Australian Economy* (1962); H.W. ARNDT, *A Small Rich Industrial Country: Studies in Australian Development, Aid and Trade* (1968); J.O.N. PERKINS, *Australia in the World Economy*, 2nd ed. (1971).

Foreign policy and defence: A. WATT, *The Evolution of Australian Foreign Policy: 1938 to 1965* (1967); T.B. MILLAR, *Australia's Defence* (1965).

Statistics: *Official Yearbook of the Commonwealth of Australia* and *Australia Handbook* (annual publications); various census bulletins and quarterly bulletins issued by the Commonwealth Bureau of Census and Statistics.

(J.D.Pr.)

Australia, History of

This article deals mainly with the history of Australia since its discovery and settlement by Europeans. The ancestors of the Aborigines of Australia arrived in the land some 25,000 years ago. They came from Southeast Asia, but whether they derived from one or more racial stocks is still unknown. In the 18th century AD, Aborigines numbered about 300,000 on the northern and eastern coasts and in the Murray River Valley; some 4,000 lived in Tasmania. For a more detailed treatment of the indigenous people of Australia, see AUSTRALIAN ABORIGINAL CULTURES. This article is divided into the following sections:

I. Australia to 1900
- Early exploration and colonization
  - Early contacts and approaches
  - Oceanic exploration
  - European settlement
  - An authoritarian society
- The great shift: 1830–60
  - Settlement
  - Politics
  - The economy
  - Culture
- Several small democracies: 1860–1900
  - Politics
  - The economy
  - The colonies
  - Social movements

II. Australia since 1900
- Nationhood and war: 1901–45
  - The economy
  - Politics and government
  - Culture
  - Growth of the commonwealth
  - The states
- Australia since 1945
  - Social and economic history
  - Culture
  - The economy
  - Domestic politics
  - Foreign politics

I. Australia to 1900

EARLY EXPLORATION AND COLONIZATION

Early contacts and approaches. Prior to documented history, there may have been further Asian contacts. Chinese astronomers are said to have made observations in Australia during the 6th century BC; firmer evidence argues for a Chinese landing near Darwin in 1432. The incursion of Islām into Southeast Asia came within 300 miles (480 kilometres) of Australia, and adventure, wind, or current might have carried some individuals the extra distance. Both Arab and Chinese documents tell of a southern land, but with such inaccuracy that they scarcely clarify the argument. Similarly, the "jave la Grande" shown on some 16th-century European maps may or may not be a genuine representation of Australia. Bugis

(Macassar) seamen certainly fished off Arnhem Land, in the Northern Territory, from the late 18th century and may have done so for generations. Perhaps only effective resistance by the Aborigines against the Bugis reserved Australia for white colonization.

The Portuguese. The quest for wealth and knowledge might logically have pulled the Portuguese to Australian shores; and the assumption has some evidential support, including a reference that Melville Island, off the northern coast, supplied slaves. Certainly the Portuguese debated the issue of a *terra australis incognita* ("unknown southern land")—an issue in European thought in ancient times and revived from the 12th century onward. Yet hard, clinching evidence of contact is lacking.

The Spanish. Viceroys of Spain's American empire regularly sought new lands. One such expedition, from Peru in 1567, commanded by Álvaro de Mendaña, discovered the Solomon Islands; excited by finding gold, Mendaña hoped that he had found the great southern land and that Spain would colonize there. In 1595 Mendaña sailed again but failed to rediscover the Solomons. One of his officers—Pedro Fernández de Quirós, a man of the Counter-Reformation, who desired that Catholicism should prevail in the southland, of whose existence he was certain—won the backing of King Philip III for an expedition under his own command. It left Callao, Peru, in December 1605 and reached the New Hebrides. Quirós named the island group Austrialia del Espiritu Santo, and he celebrated with elaborate ritual. He (and some later Catholic historians) saw this as the discovery of the southern land. But Quirós' exultation was brief; troubles forced his return to Hispanic America. The other ship of the expedition, under Luis de Torres, went on to sail through Torres Strait but almost certainly failed to sight Australia; and all Quirós fervour failed to persuade Spanish officialdom to mount another expedition.

Oceanic exploration. The exploration and settlement of Australia began early in the 17th century.

The Dutch. Late in 1605 Willem Jansz of the Netherlands sailed from Bantam in search of New Guinea. He reached Torres Strait a few weeks before Torres himself and unknowingly saw, and also named, part of the Australian coast—Cape Keer Weer, on the west of Cape York Peninsula. More significantly, from 1611 some Dutch ships sailing from the Cape of Good Hope to Java inevitably carried too far east and touched Australia: the first and most famous was Dirck Hartog's "Eendracht," from which men landed and left a memorial at Shark Bay, Western Australia, October 25–27, 1616. Pieter Nuyts explored almost 1,000 miles of the southern coast in 1626–27, and other Dutchmen added to knowledge of the north and west.

The first landing

Most important of all was the work of Abel Tasman, who won such respect as a seaman in the Dutch East Indies that in 1642 Governor General Anthony van Diemen of the Indies commissioned him to explore southward. In November, having made a great circuit of the seas, Tasman sighted the west coast of what he called Van Diemen's Land (latterly Tasmania). He then explored New Zealand before returning to Batavia. A second expedition of 1644 contributed to knowledge of Australia's northern coast; thenceforth, New Holland was the name for the landmass.

The British. But the Netherlands spent little more effort in exploration, and the other great Protestant power in Europe, England, took over the role. In 1688 the English pirate William Dampier relaxed on New Holland's northeastern coast. On returning to England, he published his *Voyages* and persuaded the Admiralty to back another venture. He traversed the western coast for 1,000 miles (1699–1700) and reported more fully than anyone previously, but in terms so critical of the land and its people that another hiatus resulted.

The middle decades of the 18th century saw much writing about the curiosities and possible commercial value of the southern seas and *terra australis incognita.* This was not restricted to Great Britain, but it had especial vigour there. The British government showed its interest by backing several voyages. Hopes flourished for a mighty empire of commerce in the eastern seas.

This was the background for the three voyages of Capt. James Cook on behalf of the British Admiralty. The first, that of the "Endeavour," left England in August 1768 and had its climax April 20, 1770, when Lieut. Isaac Hicks sighted southeastern Australia. Cook landed several times, most notably at Botany Bay, and at Possession Island in the north, where on August 23 he claimed the land, naming it New South Wales. Cook's later voyages (1772–75, 1776–79) added only a little to Australian exploration but were both symptom and cause of strengthening British interest in the eastern seas.

Captain Cook's voyages

Later explorations. Cook's voyages led to settlement but did not complete exploration of the Australian coasts. Marion Dufresne of France skirted Tasmania in 1772, seeing more, at least of the western coast, than had Tasman. The Comte de La Pérouse, another French explorer, made no actual discoveries in Australia, but his visit to Botany Bay early in 1788 was notable. In 1791 the British navigator George Vancouver traversed and described the southern shores discovered by Pieter Nuyts years before; A.-R.-J. de Bruni, chevalier d'Entrecasteaux, of France also did significant work, especially in southern Tasmania.

Two Britons—George Bass, a naval surgeon, and Matthew Flinders, a naval officer—were the most famous postsettlement explorers. Together they entered some harbours on the coast near Botany Bay in 1795 and 1796; and Bass ventured farther south in 1797–98, pushing around Cape Everard to Westen Port. Flinders, too, was in that region early in 1798, charting the Furneaux Islands. Late that year Flinders and Bass in the "Norfolk" circumnavigated Tasmania, establishing that it was an island and making further discoveries. Several other navigators, including merchantmen, filled out knowledge of the Bass Strait area; most notable was the discovery of Port Phillip in 1802.

Meanwhile, Flinders had returned home and in 1801 was appointed to command an expedition that would virtually complete the charting of Australia. Over the next three years Flinders proved equal to this task. Above all, he left no doubt that the Australian continent was a single landmass. Appropriately, Flinders urged that the name Australia replace New Holland, and this change received official backing from 1817.

France sponsored an expedition, similar in intent to Flinders', at the same time. Under Nicolas Baudin, it gave French names to many features (including "Terre Napoléon" for the southern coast) and gathered much information but did little completely new. It was on the northern coast, from Arnhem Land to Cape York Peninsula, that more work was needed. Two Admiralty expeditions—under P.P. King (1817–22) and J.C. Wickham (1838–39)—filled this gap.

European settlement. The British government determined on settling New South Wales in 1786, and colonization began early in 1788. The motives for this move have become a matter of some controversy. The traditional view is that Britain thereby sought to relieve the pressure upon its prisons, a pressure intensified by the loss of its American colonies, which hitherto had accepted felons. Convicts went to the settlement from the outset, and official statements put this first among the colony's intended purposes. But some historians argue that this glossed a scheme, likely to provoke concern both within Britain and at the diplomatic level, to provide a bastion for British trade in the eastern seas. Supporters of the commercial-strategic view emphasize that Cook had extended hopes that the South Pacific would provide essential naval stores, especially mast timber and flax. (Its lack of documentary support notwithstanding, the argument makes considerable sense.)

British motives for settlement

Whatever the deeper motivation, plans went ahead, with Lord Sydney (Thomas Townshend), secretary of state for home affairs, as the guiding authority. Arthur Phillip served as commander of the expedition; he was to take possession of the whole territory from Cape York to Tasmania, westward as far as 135° and eastward to in-

clude adjacent islands. Phillip's power was to be near absolute within his domain. The British government planned to develop the region's economy by employing convict labour on government farms, while former convicts would subsist on their own small plots.

The first settlement

The First Fleet sailed May 13, 1787, with 11 vessels, including six transports, aboard which were about 730 convicts (570 men and 160 women.) More than 250 free persons accompanied the convicts, chiefly marines of various rank. The fleet reached Botany Bay on January 19–20, 1788. Crisis threatened at once. Botany Bay was poor in soil and water and even as a harbour. Phillip therefore sailed northward on January 21 and entered a superb harbour, Port Jackson, which Cook had marked but not explored. He moved the fleet there; the flag was hoisted on January 26 and the formalities of government begun on February 7. Sydney Cove, the focus of settlement, was deep within Port Jackson, on the southern side; around it was to grow the city of Sydney.

Phillip at once established an outstation at Norfolk Island. Its history was to be checkered—settlement was abandoned in 1813 and revived in 1825 only to provide a jail for convicts who further misbehaved in Australia. (The island served a new purpose from 1856 as a home for the descendants of the "Bounty" mutineers, who by then were too numerous for Pitcairn Island.)

Phillip remained as governor until December 1792, seeing New South Wales through its darkest days. The land was indifferent, disease and pests abounded, few convicts proved able labourers, and the Aborigines often were hostile. The nadir came in autumn 1790 as supplies shrank; the arrival of a second fleet brought hundreds of sickly convicts but also the means of survival.

An authoritarian society. While much change proceeded throughout this period, authoritarian and hierarchical elements remained strong. The reception of convicts continued and was a major fact in social and economic life. Entrepreneurs strove hard but did not yet develop a staple industry. Farmers and graziers began to fill out an arc 150–200 miles (240–320 kilometres) around Sydney: this area was designated as the Nineteen Counties in 1829, and settlement beyond that limit was discouraged. Following the discovery of Bass Strait, and in concern to secure southern waterways, new settlements were made in the south.

From Britain, David Collins sailed in 1803 to settle Port Phillip; his sojourn there was unhappy, and in mid-1804 he moved to the River Derwent in southern Tasmania, already settled (September 1803) by a group from Sydney under John Bowen. Collins resettled the amalgamated parties at Hobart. In November 1804 William Paterson founded a settlement in northern Tasmania, the precursor of Launceston. These settlements united in 1812; they were still under supervision from Sydney, although only nominally from 1825. Among penal outstations settled from Sydney were those at Newcastle (1804) and Moreton Bay (1824), the forerunner of Brisbane.

British possession of the whole continent

Britain extended its possession over the whole of the continent in the mid-1820s, again suspecting French (or even American) intervention. The western boundary of the governor's commission shifted to 129° in 1825 to include Bathurst and Melville islands in the far north, and there was a small settlement in this region (1824–29). At Western Port, east of Port Phillip, another settlement was made (1826–27), while in January 1827 Edmund Lockyer began permanent settlement at Albany, Western Australia. His instructions stated that Britain now claimed all Australia.

Structure of the government. As remarked above, the constitutional structure was authoritarian. The governors were all military officers. There were no representative institutions; but the Judicature Acts of 1823 and 1828 provided for executive and legislative councils, with the major officers of government serving in both and an equal number of private individuals, chosen by nomination, in the latter. More significant at this stage was articulation of a judicial system, especially the establishment of supreme courts (New South Wales, 1814; Tasmania, 1824); normal trial by jury did not obtain.

Sociopolitical factions. Within this rigid structure, sociopolitical factions developed and affected events. Most important in the early years was the leadership of the New South Wales Corps, stationed at Sydney from 1791. Some officers of the corps sought power and profit with an avidity that led to clash after clash with the early governors; this culminated on January 26, 1808, when John Macarthur, a former officer of the corps, led a rebellion that deposed Gov. William Bligh (served 1806–08), earlier famous for the "Bounty" mutiny. In due course, the imperial government reacted and recalled the corps; but Gov. Lachlan Macquarie (served 1810–21) also clashed with the colony's "exclusives"—former officers and a handful of wealthy free immigrants. Conversely, he associated himself with the "emancipist" faction—a group that argued in favour of erstwhile convicts having a particular claim upon government and the colony's resources.

The "emancipists"

Macquarie's attitude disturbed the imperial government. After an official inquiry (1819–21) by J.T. Bigge, it encouraged the migration of men of some standing and wealth to both New South Wales and Tasmania. Such men received substantial grants of land and appeared the natural leaders of social and economic development. The emancipists continued to be strong, however, especially through the leadership of W.C. Wentworth (himself the son of a convict woman), whose newspaper, the *Australian* (founded 1824), was the spearhead of opposition, especially to Gov. Ralph Darling (served 1825–31). In Tasmania factions never formed so clearly, but there, also, the press led criticism of the government.

The convicts. By 1830 about 58,000 convicts, including almost 50,000 men, had come to Australia (the rate increasing rapidly after 1815). Many were more or less habitual urban thieves. There were a few political prisoners, while a substantial proportion of the Irish convicts (a third of the total) had become offenders through sociopolitical unrest. In Australia the convicts were either employed by government or "assigned" to private employers. Broadly speaking, conditions were not especially harsh or repressive, and "tickets of leave" and pardons provided relatively quick routes to freedom; assignment to the new settlers of the 1820s, however, often had an element of slavery. Most convicts committed some further misdeeds, although only about one-tenth were charged with serious offenses. Those found guilty went to secondary penal stations, the (sometimes exaggerated) horror spots of Australian history—Macquarie Harbour, Newcastle, and Moreton Bay in this period and, later, Norfolk Island and Port Arthur. The convicts gave Australia a large *Lumpenproletariat*, yet success stories were common enough, and many led decent lives. There were only a few militant protests—the most remarkable was an uprising among Irish convicts outside Sydney in March 1804. Altogether, the convict impact was less grim and ugly than might be expected.

Economic activity. The maintenance of convicts was essentially the economic resource of the colony for many years; this function caused very considerable expenditure by the British government. Wealth was won by supplying government stores with food and grain, or by controlling internal trade (especially in rum), which ultimately depended on the government maintaining the settlements, or by both. The officers of the New South Wales Corps were skilled in filling these roles, although civil officers, private settlers, former convicts, and even serving convicts all had their own means of doing business, and the amount of petty commercial activity was large.

The growth of commerce

Farming was pursued on a widely ranging scale. John Macarthur (see above *Sociopolitical factions*) was most notable of those who early believed that wool growing would be a major economic resource; he himself received a substantial land grant in 1805 to pursue this hope, and he persuaded Bigge of its validity. By 1830 these hopes were yet to be decisively confirmed. Sealing and whaling returned a little more, although the richest seal fields (especially in Bass Strait) were soon thinned; and not until the 1820s did colonists have the wealth to engage seriously in whaling, although British and Americans early

used Australian ports for this purpose. Maritime adventure led early colonists to make contact with Pacific islands, most importantly Tahiti.

Exploration. The period saw some notable exploration by land. From early days in Sydney men sought a way over the mountains, 50–100 miles west. The task was accomplished in 1813; the journalist W.C. Wentworth (see above *Sociopolitical factions*) led the party. A surveyor, George William Evans, followed their route to Bathurst (founded 1815) and reported rich pastoral country. John Oxley further mapped the inland plains and rivers, especially the Lachlan and Macquarie, and also explored the southern coasts of future Queensland (1823), while Allan Cunningham was the great pioneer of that state's hinterland (1827). Meanwhile, in 1824–25, Hamilton Hume and William Hovell went overland southward to the western shore of Port Phillip. Charles Sturt, in 1828–30, won still greater fame by tracing the Murray–Murrumbidgee–Darling river system down to the Murray's mouth.

Culture. The writings of explorers and pioneers were Australia's first contributions to literary culture. While catering to the European appetite for natural history, they sometimes achieved literary grace. Pictorial illustrations of the new land, some by convicts, also dated from the earliest years. David Collins' *An Account of the English Colony in New South Wales* (1798) and W.C. Wentworth's *Description of New South Wales* (1817) were literate, informed, and impressive. Wentworth showed skill as a versifier, too, especially in his *Australasia* (1823). Newspapers were founded as early as 1803, and they contributed to cultural as well as political history. Outstanding was the architecture of Francis Greenway, a former convict, who, under Macquarie's patronage, designed churches and public buildings that remain among the most beautiful in Australia.

THE GREAT SHIFT: 1830–60

The three decades between 1830 and 1860 comprised the period of most rapid change in Australia's history. The impact was most evident in politics, but the economy and culture were no less affected. Patterns then established persisted.

Settlement. Four of Australia's six states were formed between 1829 and 1859. A British naval captain, James Stirling, examined the Swan River in 1827 and interested English capitalist-adventurers in colonization. Two years later he returned to the Swan as governor of the new colony of Western Australia. The Colonial Office discouraged schemes for massive proprietorial grants; still the idea persisted, with Thomas Peel—kinsman of the future prime minister Sir Robert Peel—investing heavily. But colonization was grim work in a hot, dry land, with the government reluctant to spend a penny. Western Australia's story for decades was survival, not success.

Yet the same enthusiasm quickly generated around proposals to establish a colony in South Australia, inspired by a British social reformer, Edward Gibbon Wakefield, who argued that if land were sold at a "sufficient" price, its owners would be forced to maximize its value by cultivation, while labourers would have to lend their energies to that task before being able to become landowners themselves. Highly doctrinaire, Wakefieldianism appealed to the liberal intelligentsia and to dissenting groups in England who also backed nascent South Australia. The first colonists arrived in July 1837, and Adelaide was settled soon afterward. The colony experienced many hardships, but lasting significance resulted from its founders' stress on family migration, equality of creeds, and free market forces in land and labour.

The northern and southern portions of New South Wales formed separate colonies. Settlement into the Port Phillip district in the South proceeded very quickly from the mid-1830s, colonists coming from both north of the Murray and from Tasmania; Melbourne's history began in 1835 and boomed immediately. Throughout the 1840s there were calls for constitutional independence, granted in 1851, when the Port Phillip District took the name Victoria. Northward, the Moreton Bay District was never quite so buoyant; and the creation of Queensland had to wait until 1859. Short-lived settlements included Port Essington (1838–49) and Gladstone (1847).

Development of responsible self-government

Politics. All colonies except Western Australia gained responsible self-government. New South Wales led the way when an imperial act of 1842 created a two-thirds elective legislature. The Australian Colonies Government Act (1850) extended this situation to Victoria, South Australia, and Tasmania. The act made allowance for further revision of the colonial constitutions, and in 1855–56 this took effect in the four colonies, Tasmania then abandoning the name Van Diemen's Land. Queensland followed, at its separation from New South Wales. All had bicameral legislatures, with ministers responsible to the lower houses, which by 1860, except in Tasmania, were elected on a near-democratic adult-male franchise.

While the imperial power often responded to colonial cries for self-rule, there were some tense moments. Virtually all colonists abhorred paying taxes for imperial purposes, including the costs of maintaining convicts locally; a good many disliked convictism altogether; most disputed the imperial right to dictate land policy; and many, especially in South Australia, disapproved of the imperial government directing that aid to be given to religious denominations. These were the main issues, coalescing with that of increased self-rule, which stoked sometimes quite violent debate.

From the outset of the period, the imperial government fostered a freer market in land and labour throughout the colonies, not merely in South Australia. Thus grants of land ceased in 1831, replaced by sale. Attempts to create a satisfactory system caused much friction, with colonists generally hostile to any demand for payment. In New South Wales in 1844, new regulations even prompted talk of rebellion.

Friction over convicts

With regard to labour, colonists agreed with imperial encouragement of free migration, but friction arose over the convicts. British opinion in the 1830s became increasingly critical of the assignment of convicts to private employers as smacking of slavery; it was abolished in 1840, and with it transportation of convicts to the mainland virtually ceased, although more than before were sent to Tasmania. The end of assignment removed the chief virtue of transportation from the colonists' viewpoint and so contributed to a very vigorous movement against its continuation. The British government terminated it for eastern Australia in 1852; in Western Australia transportation commenced in 1850, at the colonists' behest, and continued until 1868. Altogether some 151,000 convicts were sent to eastern Australia and nearly 10,000 to Western Australia.

In the early 1850s the most dramatic political problem arose from the gold rushes (see below *The economy: Minerals*). Diggers (miners) resented tax imposition and the absence of fully representative institutions. Discontent reached a peak at Ballarat, Victoria, and in December 1854, at the Eureka Stockade, troops and diggers clashed, with loss of life. The episode is the most famous of the few occasions in Australia's history involving violence among Europeans.

Common suspicion of the imperial authority modified but did not obliterate internal tension among the colonists. Divisions of ideology and interest were quite strong, especially in Sydney, where a populist radicalism criticized men of wealth, notably the big landholders. The coming of self-government marked a leftward, although far from revolutionary, shift in the internal power balance.

The economy. This period saw the first and greatest booms of the two bonanzas of Australian economic growth—wool and minerals.

Wool. Only now did men, money, markets, and land availability interact to confirm that Australia was remarkably suited for growing fine wool. Occupation of Port Phillip was the most vital part of a surge that carried sheep raising 200 miles and deeper in an arc from beyond Adelaide in the south, north, and east to beyond Brisbane. The "squatter" pastoralist became an arche-

type of Australian history. Although pastoralism contributed to depression in the early 1840s, the industry kept growing, and the whole eastern mainland benefitted consequently.

Pastoral expansion had one grim corollary—disaster for the Aborigines. Relations between the two races were always tense. The early governors all ordered and even practiced benevolence, but to little effect. When the European spread inland, the Aborigine fought hard, but that only strengthened the European's arm. The Tasmanians moved rapidly toward extinction, but the Australians, with further room for withdrawal, evaded that fate. Philanthropists, missionaries, and administrators stood aghast but found no remedy.

Minerals. The first significant mineral discovery was that of copper in South Australia (1842 and 1845). The discovery had the effect, to be repeated time and again, of suddenly redeeming an Australian region from stagnation. Much more remarkable, however, were a series of gold discoveries made from 1851 on, first in southern New South Wales, but then, and overwhelmingly, throughout Victoria. As a result, Australia changed from a land of exile to one of golden attraction. The Victorian economy benefitted from the flood of men and money, although the smaller colonies suffered. Despite the Eureka Stockade incident, the diggers proved quite moderate and responsible.

Culture. Governments and citizens paid considerable heed to improvement of soul and mind. From the mid-1830s, generous aid helped all Christian churches to expand. The Church of England had the highest nominal allegiance, but in the eastern mainland colonies Roman Catholicism was notably strong; Methodism had vigorous advocates throughout; Congregationalism and other forms of dissent dominated in South Australia; and Presbyterianism had its chief strength in Victoria. Most churches attended to education, especially the provision of superior schools, while the state struggled to provide a primary system. The universities of Sydney and Melbourne were founded in 1850 and 1853, respectively. Mechanics' institutes, museums, and botanical gardens were also built.

Architects created much beauty in early Australia. Artists and writers were active; drama and music developed in all towns. The first Australian novel, *Quintus Servinton* (1831), was by a convict, Henry Savery; Henry Kingsley's *Geoffrey Hamlyn* (1859) is often judged the first major Australian novel (though Kingsley was an English citizen). John West's *History of Tasmania* (1852) was a work of remarkable scope and insight.

Further exploration

Various forms of science had their investigators, but land exploration remained the richest field of discovery. Sir Thomas Livingstone Mitchell confirmed Sturt's work on the river systems (see above *An authoritarian society: Exploration*) and first opened the way from New South Wales to the rich lands of western Victoria (1836). The West Australian coastal regions were mapped by George Grey (1837–40) and by Edward John Eyre, who went overland from Adelaide to Albany (1840). Eyre and Sturt both vainly attempted to reach midcontinent from Adelaide; this was at last achieved in April 1860 by John McDonall Stuart, who in 1862 went still farther, to Darwin. Meanwhile, the central north and the northeast had been penetrated from Sydney; the most famous explorer was Ludwig Leichhardt, who led two successful expeditions (1844, 1846–47) before disappearing in an attempted traverse from the Darling Downs to Perth. An equal and more celebrated tragedy ended the expedition of Robert O'Hara Burke and William John Wills, who crossed from Melbourne to the Gulf of Carpentaria in 1860–61 but starved to death on the return. Later explorations of Western Australia in the 1870s added the names of John Forrest and Ernest Giles to the pantheon of explorer-heroes.

SEVERAL SMALL DEMOCRACIES: 1860–1900

During this period the colonies had little formal relation with each other; instead they concentrated their attention inward on their capitals. New Zealand had more in common with the eastern colonies than did Western Australia. Federation came about in 1901, but the more striking fact was its tardiness. Nevertheless, the colonies did follow similar, if independent, paths.

Politics. Democracy was largely fulfilled, save that the upper houses remained elitist in franchise and membership. Governments changed rapidly over long periods, but the constitutions survived. Political groupings were extremely intricate, often personal or power seeking in origin, but allowing some expression for liberal or conservative ideology.

Social reform

The liberals made the colonies quite advanced in matters of social reform, if not the average man's paradise that some glib publicists pictured. Breaking up the large "squatter" estates and replacing them with yeoman farming was a constant concern, meeting many difficulties yet achieving some effect where market and environment allowed. Reformers put much faith in education and strove toward providing adequate primary schooling for all. "Free, secular, and compulsory" was a slogan and roughly the final result; this entailed savage controversy with the Roman Catholic Church, which scorned the godless schools and made enormous efforts to provide its own. Other forms of state aid to religion tapered away. Factory legislation and rudimentary social services presaged the welfare state; restriction of nonwhite, especially Chinese, immigration belonged to this context, for Europeans feared these labourers would reduce living standards, but the restriction was also a matter of sheer racism.

The economy. Overall the economy prospered. Wool and metals continued as the great export income earners. Pastoralism flourished, especially up to the mid-1870s; despite land legislation, this was the heyday of the squatter "aristocracy." Expansion of sheep and cattle growing into the more distant hinterland continued the heroic-pioneer theme of earlier years. Railway construction aided rural industry and proceeded remarkably fast, notably in the 1880s: between 1875 and 1891 the mileage rose from 1,600 to above 10,000 and reached as far as 500 miles (800 kilometres) inland. Most of the required capital was raised overseas on behalf of governments, contributing to the extremely important role played by the public sector in economic growth.

Gold and copper discoveries

Mining. Victoria's gold and South Australia's copper maintained their significance as new techniques allowed more sophisticated exploitation. Gold was found in southern Queensland in the later 1860s, then in the Northern Territory, and in tropical Queensland: the Palmer River goldfield pulled men to the far north in the mid-1870s. By then Cobar, in central New South Wales, had proved the most important of many new copper fields. Tin also became significant, Mt. Bischoff in Tasmania being the world's largest lode at its discovery in 1871. The 1880s was predominantly the decade of silver; western New South Wales proved richest, and in 1883 Charles Rasp, a German migrant, first glimpsed the varied riches of Broken Hill, which were to make that city almost fabulous and to prompt the establishment of Broken Hill Proprietary Company Ltd.—in time, Australia's largest private enterprise. Also from 1883 dated another big and ramifying discovery, the gold of Mount Morgan, Queensland. Gold also became Western Australia's great bonanza in the early 1890s, the Kalgoorlie and Coolgardie fields winning world attention; the copper of Mt. Lyell, Tasmania, was another highlight of that decade. These discoveries were both product and instigator of much wider activity, creating speculation, mobility, boom, and slump of extraordinary impact.

Industry. Urban expansion and the growth of secondary industry, while less distinctively Australian and contributing little to export income, were remarkable. By the criteria of investment, employment, and relative acceleration, the growth of secondary industry outstripped that of primary industry. Secondary industry multiplied its growth some ten times over during the period, so that manufacturing and construction accounted for 25 percent of the national product in the 1880s. The population ratio shifted decisively from country to town, establish-

ing an extreme capital-city concentration and eventually putting Melbourne and Sydney among the world's large cities. Urban building and services attracted much capital; and most manufacturing was directed to providing food, furniture, and clothing for the relatively affluent townsman. City speculation contributed more than its share to a general overcapitalization, which, in company with worldwide forces, produced a depression in the early 1890s, the main impact of which was in the urban-secondary sector.

The colonies. The history of the respective colonies sharpens some points in this general background. In the later 19th century, regional characteristics consolidated, and they changed little at least until the 1960s.

Victoria. Victoria retained the impetus of the 1850s for a full generation. This was most evident in its capital, Melbourne, which had a vigorous cultural and social life. Ardent and ideological liberalism was evident in the colony's education controversy and, with greater novelty, in its adoption of protection as a means of developing its industries and living standards. Disputes between the upper (conservative) and lower (radical) houses of the parliament were frequent and sharpened political feeling. A famous clash occurred in 1865–66 over a protective tariff; another, in 1877–78, over payment of members—the liberal cause triumphed on both occasions.

Adoption of protection

New South Wales. With its longer background, New South Wales changed less in this period. Its master politician, Henry Parkes, first came into prominence in the 1840s. Parkes was involved in sectarian disputes, which were especially vigorous in the colony. Another major theme of political debate was protection versus free trade—the latter retaining greater favour, in contrast to Victoria. Sydney had its share of scandals and scalawags, especially late in the period, contributing to a rumbustious image.

Queensland. Physical expansion westward and northward dominated the history of Queensland. Cattle and sugar became industries of substantial importance. A class of small farmers was established that aspired to settle the tropics, which were usually considered unsuitable for small-scale farming by Europeans. Conversely, the established "kings" of the region used Kanakas (labourers from the Pacific islands). Thus the continued immigration of such labour provoked hot debate, which was not resolved until after federation, when the young commonwealth imposed an absolute prohibition. The Kanaka question contributed to regional feeling antagonistic to the capital.

South Australia. South Australia enjoyed less prosperity than its eastern neighbours. Agriculture remained significant in its economy but saw much disappointment; in the decade around 1870 farmers pushed out into semi-arid country, hopeful that rain would follow the plough, only to learn with cruel certainty that it did not. Settlement drew back toward Adelaide, the sole sizable town. Landholding did prompt South Australia's most famous contribution to reform: that land transfer proceed simply by registration, rather than through cumbrous title deeds. Another notable contribution was the institution of woman suffrage (1894), which effectively brought nationwide application of the principle at federation; appropriately, South Australia was the home of Catherine Helen Spence, the most remarkable Australian woman of the time, who published a significant novel, *Clara Morison* (1854), and became active in many social and political movements.

Woman suffrage

In 1863 South Australia took over the administration of the area thenceforth known as the Northern Territory, which earlier had remained technically part of New South Wales; the change entailed adjustment of boundaries. (The Northern Territory became the concern of the federal government in 1911.) South Australia set up its administrative centre at Darwin. In 1872 the construction of an overland telegraph line linked Australia with the outside world via Darwin.

Tasmania. The 1860s imprinted a "sleepy hollow" image on Tasmania, which persisted. The mineral discoveries at Mt. Bischoff and elsewhere were correspondingly important in reviving the economy. Nevertheless, living standards remained lower than in Victoria, and in Tasmania there were still property qualifications for voting in 1900. The colony contributed to democratic practice, however, by experimenting with proportional representation.

Western Australia. Western Australia ceased to receive convicts in 1868; it gained a partly elected legislature in 1870 and responsible government in 1890. Premier throughout the 1890s was John Forrest, as adept at politics as at exploration. Until the gold rushes, economic growth was slow and primitive; in the 1890s the colony was fastest in relative growth, and little short of that in absolute terms. Farming (in the southwest), town and railway building, and social legislation all followed.

Social movements. Working class and radical movements stretched back to the 1830s, although substantial trade union organization came only after midcentury.

Labour. The unions won some job benefits, including widespread adoption of the eight-hour day. The 1870s and 1880s saw extensive mass unionism, notably among miners and sheep shearers. Trades halls arose in the cities, and organization, extending beyond colonial boundaries, began to knit together. The unions early considered using political pressure and gaining political representation. This inclination strengthened in the early 1890s, helped by tougher times and by employers' stiffening resistance to union demands. Thus arose the labour parties, which gained quick success, especially in New South Wales and Queensland. At first the labourites' aim was simply to influence ministries, but for a few days in December 1899 Anderson Dawson was Labor premier in Queensland.

Rise of the labour parties

Other radicals reacted differently to the pressures of the 1890s. A few hundred of them set off for Paraguay in 1895 to establish there a Utopian "New Australia"; they failed. Republicanism was probably stronger in the 1890s than at any other time, sometimes accompanying a Marxist-like militance.

Movement toward federation. Federation was another ideal of the times. Most important politicians supported the cause, with more or less altruism. They could operate on more positive factors than common background and apparent common sense. Since the Crimean War (1853–56) Australians had feared incursion from the north by Europeans or Asians or both; the most emphatic result came early in 1883, when the government of Queensland, fearful of Germany, took possession of Papua, forcing Britain's reluctant connivance. Better defense was one motive for association, and so was the prospect of more effective Asian immigration restriction; intercolonial free trade was another desideratum. The Australian Natives Association (Australian-born comprised 64.5 percent of the population in 1901) rallied to the cause.

Yet the mills ground slowly. A federal council existed from 1885, but this was only a standing conference, without executive power. New South Wales never joined the council; the senior colony was jealous of a movement that would reduce its autonomy, the strength of which was in Victoria. Conventions met in 1891 and 1897–98 to prepare draft constitutions. These then went to referenda, which gained "yes" majorities. The Commonwealth of Australia came into existence on January 1, 1901.

The Commonwealth of Australia

The constitution was federal, with the states (as the colonies now became) forsaking only limited and specified power to the commonwealth government—these included defense, immigration, customs, marriage, and external affairs. While the lower house, the House of Representatives, was elected by single-member constituencies of roughly equal size, each state had an equal number of representatives in the upper house, the Senate. Ministers were to be members of Parliament. A high court would interpret the constitution.

Culture. Men of learning had contributed to the nationalist surge. Especially in the 1890s and through the Sydney *Bulletin*, verse and prose portrayed the "Outback" as the home of the true Australian—the bush worker:

tough, laconic, and self-reliant, but ever ready to help his mate. The *Bulletin* was nationalist, even republican, and much more radical than the federalist politicians. Henry Lawson and Joseph Furphy were the supreme writers of the nationalist school, and the latter's *Such Is Life* (1903) was an outstanding novel. Painters and poets also extolled the nationalist ideal.

Not all cultural achievement belonged to the nationalist context, however. Henry Kendall was a lyricist of nature, and A.L. Gordon wrote of horses and countryside with a skill that won him a memorial in Westminster Abbey. "Rolf Boldrewood" (Thomas Alexander Browne) wrote tales of outback adventure, while the great 19th-century Australian novel was Marcus Clarke's *For the Term of His Natural Life* (1874), based upon convict records and legends. The older universities remained small, but with some outstanding men on their faculties; the universities of Adelaide (1874) and Tasmania (1890) were new foundations. Ferdinand von Mueller was an outstanding botanist who worked primarily at the Botanic Gardens, Melbourne. That city was the hometown of the great coloratura soprano Nellie Melba (Helen Porter Mitchell, born 1861), not quite the first but certainly the most famous of numerous Australian singers to achieve international renown.

Popular culture

Popular culture followed the British model, with music halls, novelettes, and especially sport to the fore. "Australian Rules" football developed first in Melbourne and became strong throughout southern Australia; in cricket, a victory over the mother country in 1882 established one area of colonial equality. Admiration combined with fear to create a sporadic cult of the bushranger (rural desperado-thief): its most famous expression came with the capture of Ned Kelly's gang and Kelly's execution in 1880. Urban youths joined in gangs, or "pushes," and won the epithet larrikin.

II. Australia since 1900

NATIONHOOD AND WAR: 1901–45

The world's passions and conflict of the early 20th century were to shape the new nation's history, despite its physical distance from their epicentres. By many standards, this was the least attractive of the five major periods of Australian history. Nationalism strengthened, but it killed and sterilized rather more than it inspired; egalitarianism tended to foster mediocrity; dependence on external power and models prevailed. Yet creativity and progress survived.

The economy. Drabness was most evident in economic affairs. At the broadest level of generality, the period did little more than continue the themes of the 1860–90 generation. The most important such themes were the improvement of communications (railways reached their peak of 27,000 miles in 1941, and meanwhile came the motor boom) and increasing industrialization. In the primary field, there was significant expansion of exports, with wheat, fruits, meat, and sugar becoming much more important than hitherto. But just as manufactures received increasingly high tariff protection, so the marketing of these goods often depended on subsidy; and so the sheep's back continued to be the nation's great support in world finance. Metals, gold especially, were important in the early years; but thereafter this resource conspicuously failed to provide the vitality of earlier and later times. The worldwide economic depression of the 1930s affected Australia, especially its primary industries; otherwise the overall rate of growth, and probably of living standards, too, scrambled upward—a little more quickly than average in the mid-1920s and perceptibly so in the 1940s.

Politics and government. In national politics, men fought for office with increasing vigour and resource, while their administrative performances generally began well but then ebbed. Retrospect can espy a strangely regular pattern wherein four crisis episodes—the establishment of national policies, World War I, depression, World War II—alternate with three periods dominated by a political party enjoying its climacteric. The Labor Party filled this role in 1904–15; the Country Party, 1919–29; and the United Australia Party, 1931–41. At the state level politics were more confused, if not impenetrable.

Strengthening of the central government

A constant theme was the strengthening of the central government as against the states. This complemented the high degree of homogeneity, especially in personal and social matters, that extended through Australia's great physical spread; it was expressed primarily through the commonwealth's financial powers—at first especially relating to customs but later by direct taxation. From World War I both levels of government imposed such taxes, but in 1942 the federal government virtually annexed the field, with the high court's approval. The establishment of a national capital at Canberra, where Parliament first sat in 1927 after meeting in Melbourne since federation, symbolized this situation.

Culture. The period produced not only Furphy's *Such Is Life* (see above *Social movements: Culture*) but also the work of Henry Handel Richardson (H.H.R.; pseudonym of Ethel F.L. Richardson, later Robertson), another contender for "the great Australian novelist." In *The Fortunes of Richard Mahony* (three volumes, 1917, 1925, 1929), H.H.R. told the anguish of the central character, modelled on her father, as he sought to come to terms with Australian life. The tension of dual loyalties to Britain and Australia was a major concern also of Martin Boyd, whose long career as a novelist began in the 1920s. A more exclusively nationalist tone pervaded many tales of outback life and historical novel sagas. The first notable novel of urban life was Louis Stone's *Jonah* (1911); a later contributor to this genre was Vance Palmer (especially *The Swayne Family*, 1934), who, with his wife Nettie, won fame as a literary critic and selfless patron of the aspiring young.

The most significant contribution in poetry came from a group in Sydney influenced by the German philosopher Nietzsche and the late-19th-century French innovators. Outstanding was Christopher John Brennan, a major theorist of symbolism. While calling on the Australian background, these men gave a sophistication to their poetic world that lifted it far from outback balladry. Associated with this group was Norman Lindsay, an artist, novelist, sculptor, and seer.

In art, the rural landscape held domination. Revolutionary changes in European art were markedly slow in affecting Australia, but a few artists did produce some notable work of imaginative technique. Musical composition was hackneyed and mediocre, although in Percy Grainger Australia produced (but did not retain) a musician of remarkable originality and ability. Architecture promised an interesting chapter with the selection of the American Walter B. Griffin's design for the city of Canberra; in practice this was much mutilated, but Griffin did do some interesting work in both Melbourne and Sydney.

Education

One outstanding area to which the universities contributed was anthropology; a chief protagonist was A.R. Radcliffe-Brown (professor of anthropology at Sydney, 1925–31). Australians increasingly filled faculty posts, although most who did so had Oxbridge degrees, while some of the most able native intellects worked overseas. The University of Western Australia, founded in 1911, drew on one of the most substantial philanthropic bequests in Australian history (from the newspaperman Sir Winthrop Hackett) and initially charged no fees. Other university foundations were Queensland (1909) and colleges at Canberra and Armidale. The states developed their own secondary schools throughout the period, although the achievement was scarcely comparable to the development of primary education in the early period.

Australia was in the forefront of film making early in the century, but this early promise soon faded. A.B. Paterson's "Waltzing Matilda" became Australia's best known song—part folk hymn and part national anthem. Radio had an impact in Australia equal to that elsewhere; radio stations became a mark of medium-town status, and the Australian Broadcasting Commission became a major force in culture and journalism. Radio

helped make the 1930s probably the most sports-conscious decade in Australia's history. Cricket, tennis, swimming, boxing, and horse racing were areas of athletic excellence. Aviation moved from sport to enterprise to business; Charles Kingsford-Smith was the most famous hero and Qantas the most successful airline.

Early ministries

Growth of the commonwealth. The first two prime ministers were Edmund Barton (served 1901–03) and Alfred Deakin (served 1903–04), who had led the federation movement in New South Wales and Victoria, respectively. They were liberal protectionists. Their ministries established the "White Australia" (*i.e.*, exclusion of Asians) immigration policy, a tariff, an administrative structure, the high court, and a court of conciliation and arbitration that carried probably to the highest point anywhere in the world the principles of industrial arbitration and judicial imposition of welfare and justice through wage and working-condition awards.

In 1904 J.C. Watson led the first, brief Labor cabinet, followed by G.H. Reid's conservative free-trade ministry. Deakin led again (1905–08); then Andrew Fisher was Labor's second prime minister (1908–09), his ministry defeated when liberals and conservatives "fused" in Deakin's third term (1909–10). Then Labor won its first clear majority at election, which it barely lost in 1913 and regained, still under Fisher, in 1914. This kaleidoscope did not hinder—perhaps it even prompted—ambitious governmental policies. Social services were extended with old-age pensions (1908) and maternity grants (1912); protection rose markedly in a 1908 tariff; the Commonwealth Bank was established; and an army and navy developed.

The new nation was psychologically as well as physically prepared for war. Fear of attack became increasingly directed against Japan, and it prompted pressure on Great Britain for a firmer policy in the New Hebrides (since 1886 supervised jointly by Britain and France)—achieved in 1906–07. Although many Australians criticized Britain when the latter appeared negligent of local interests, the dominant note was overweening loyalty to the empire. Colonial troops had fought in both the Sudan and Boer wars. In 1914, when World War I began, politicians of all hues rallied to the imperial cause.

World War I. Some 330,000 Australians served in World War I; 60,000 died, 165,000 suffered wounds—few nations made such relatively heavy sacrifice. The most famous engagement of the Australia and New Zealand Army Corps (ANZAC) was in the Dardanelles campaign (1915); the day of the landing at Gallipoli—April 25—became a day of national reverence, honoured far beyond any other. Even before Gallipoli, Australian troops had occupied German New Guinea, and the Australian vessel "Sydney" sank the German cruiser "Emden" near the Cocos Islands (November 9, 1914). After the Dardanelles, Australians fought primarily in France—Ypres, Amiens, and Villers Bretonneux were among the battles, all redolent with slaughter. In Palestine, the Australian light horse and cavalry corps contributed to Turkey's defeat.

Effects of war

The war profoundly affected domestic affairs. In economic development, it acted as a supertariff, benefitting especially textiles, glassmaking, vehicles, and the iron and steel industry. Such products as wool, wheat, beef, and mutton found a readier market in Britain, at inflated prices. But the shock of war affected politics much more, especially by giving full scope to the furious energy of W.M. Hughes, who supplanted Fisher as Labor prime minister in October 1915. Soon afterward he visited Britain. There his ferocity as a war leader won acclaim, and he became convinced that Australia must contribute still more. He advocated military conscription for overseas service; but a referendum in October 1916 declared negatively for this proposal, and immediately afterward the Labor parliamentary caucus moved no confidence in Hughes's leadership. But he continued as prime minister of a "national" government, even after losing a second conscription referendum in December 1917. The referenda in particular and war stress in general made these years uniquely turbulent in Australian history. The Labor Party lost other men of great ability along with Hughes. The split cemented a long-standing trend for Roman Catholics to support the party. Hughes's enemies also included the small but growing number of extremists—most notably the Sydney section of the Industrial Workers of the World (IWW)—who opposed the war on doctrinaire grounds.

The postwar years. The aftermath of war echoed, but finally resolved, this turbulence. Some radicals hoped that returning servicemen would force social change; but instead, the Returned Servicemen's League became a bastion of conservative order, its supporters ready to use physical force against local people they considered "Bolsheviks." The Labor Party faltered, its members adopting a more radical socialist type of platform in 1921, but with far from uniform conviction. When the challenge came to Hughes's leadership early in 1923, it arose partly from the conservative-business wing of Hughes's own Nationalist Party (its representative S.M. Bruce becoming prime minister) and partly from the Country Party, which from late 1922 held a crucial number of parliamentary seats. Although led by wealthy landowners, the Country Party won support from many small farmers: it benefitted too from its former-soldier image and from widespread country-versus-city feeling. Its leader, E.C.G. Page, had considerable, if erratic, force.

Bruce continued as prime minister until 1929, with Page his deputy in Nationalist–Country coalitions. Bruce declared his policy to be the discovery of "Men, Money, Markets" and worked hard toward this end. The cost was high, however: tariffs, bounties, prices, and public indebtedness all rose. There was considerable administration innovation—*e.g.*, the Loan Council regulated all governments' overseas borrowing—and the successful Council for Scientific and Industrial Research (later, the Commonwealth Scientific and Industrial Research Organisation [CSIRO]) was established in 1926 to apply scientific expertise to developmental problems. The worldwide development of consumer industry had its impact: the revolution in transportation provided by the automobile is the best example, although full-scale car production was still in the future.

"Men, Money, Markets"

With much economic activity subsidized—the exception being one primary product (wool)—Australia was particularly vulnerable to the Great Depression of the 1930s. It struck hard: unemployment exceeded 25 percent of the work force and imposed a degree of social misery unknown in Australian history. The rate of recovery was uneven, manufactures doing better than primary industry. Population growth slowed; at the nadir, emigration exceeded immigration.

Politics reflected the impact. J.H. Scullin succeeded Bruce as prime minister in 1929, and his Labor ministry suffered the real squeeze of events; within the Labor Party there was considerable division as to how government should react to the Depression. Some favoured a generally inflationist policy, with banks facilitating credit issue and governments extending public works. Right-wing Labor men distrusted such a policy; radicals would have gone further, by renouncing interest payment on overseas loans. Orthodox opinion argued for deflationary policies—curtailed government expenditure, lower wages, balancing the budget, and the honouring of interest commitments. In June 1931 the commonwealth and the state governments agreed on a plan—the Premiers' Plan. Albeit having some inflationary features, this foreshadowed a one-fifth reduction in government spending, including wages and pensions—a considerable affront to Labor's traditional attitudes.

Against this background, the government disintegrated. Before the Premiers' Plan, some right-wingers, led by J.A. Lyons, had crossed to the opposition. In November some leftist dissidents voted against Scullin, forcing his resignation. In the elections that followed, Labor suffered a heavy defeat. The new prime minister was J.A. Lyons, whose followers had coalesced with the erstwhile Nationalists to form the United Australia Party (UAP). Lyons led a wholly UAP government until 1934 and UAP–Country coalitions until his death in 1939.

The governments of J.A. Lyons

The Lyons governments provided stability and not much more. Recovery was uneven and sporadic, quicker in manufacturing than in primary industry, aided more by market forces than by governmental planning. Two policies failed badly to fulfill expectations—the Imperial Economic Conference, held at Ottawa, Ontario, in 1932, improved trade slightly, but the integrated economic community for which some had hoped never developed; and Australia's "trade diversion policy" of 1936, which tried to redress the imbalance of imports from Japan and the United States, offended those countries and actually reduced exports further. A plan for national insurance, the Lyons governments' most ambitious social legislation, also aborted. These mishaps did not much bother the electorate; improvement, even if meagre, was enough to retain favour.

Internal division was the greater threat to the government. This became manifest after Lyons' death. The UAP elected Robert Menzies as its new leader (and therefore prime minister); but the decision was hard fought, and it was criticized publicly and vehemently by Page, still leader of the Country Party. Nevertheless, Menzies retained office; but internal division persisted, the coalition's parliamentary majority was tiny, and Menzies resigned in August 1941. A.W. Fadden, the new leader of the Country Party, then took office, but in October he gave way to John Curtin and a Labor ministry.

Communist and Fascist movements

While the electorate generally voted conservative, Australia shared the common Western experience of the interwar years in the rise of a small, vigorous Communist movement. Founded in 1922, the Australian Communist Party made most headway in the big industrial unions and in Sydney; it also had some influence and supporters among the intelligentsia, especially during the 1930s. Overall, the party suffered a fair share of internal factionalism but for the most part was able to present a united face to the public.

Fascism achieved no formal political recognition in Australia, but there were hints of sympathy toward Fascist attitudes—D.H. Lawrence wrote of such in his novel *Kangaroo*, based on a brief visit in 1922; and an "Australia First" movement began in literary nationalism but drifted into race mystique and perhaps even treason. An intellectual movement of more lasting force developed among a group of young Roman Catholic intellectuals in Melbourne in the mid-1930s. They developed a commitment to social justice and against Communism somewhat in the manner of G.K. Chesterton. This was the "Catholic Social Movement," which had considerable influence on Australian Catholics.

Whereas Australia had been virtually spoiling for war before 1914, passivity became the international keynote after 1920. At the Paris Peace Conference that formally concluded World War I, Hughes was his fire-eating self, especially in defense of Australia's interests in the Pacific; thus he won a mandate for erstwhile German New Guinea and Nauru (an atoll in the central Pacific) and effectually opposed a Japanese motion proclaiming racial equality, which he thought might presage an attack on Australia's immigration laws. In the League of Nations, Australia was an independent member from the outset. Bruce's succession as prime minister marked a new emphasis. Very much an Anglo-Australian, Bruce led the nation into a period when "the empire" became the object of yet more weighty rhetoric and more desperate hope than earlier. Australia did not ratify the Statute of Westminster (1931, embodying the 1926 Balfour Declaration as to the constitutional equality of the Dominions) until 1942. The UAP governments followed Britain closely concerning the totalitarian expansion of the 1930s; if Australian influence counted for anything, it was to strengthen appeasement of Germany and Japan. Although fear of Japan continued, that country's accession to the Fascist camp did not provoke a tougher governmental line. The Labor Party meanwhile was even more incoherent and variable in matters of foreign policy than were its social-democratic counterparts elsewhere in the Western world: isolationism and anti-Fascism were equal and opposing forces.

World War II. When war came again, however, the nation's response was firm—some 30,000 Australians died in World War II and 65,000 were injured. From early in the war, the Royal Australian Air Force was active in the defense of Britain. The Australian Navy operated in the Mediterranean (1940–41), helping to win the Battle of Cape Matapan (March 1941). Australian troops fought in the seesaw battles of North Africa. In mid-1941 Australians suffered heavy losses both in the Allied defeats in Greece and Crete and in the victories in the Levant. Meanwhile, the German general Erwin Rommel was scoring his greatest triumphs in North Africa; out of these emerged the successful Allied defense of Tobruk, substantially by Australians (April–December 1941).

After the Japanese attacked the United States naval base at Pearl Harbor, Hawaii (December 7, 1941), however, the focus shifted homeward. The Japanese victories of the following months more than fulfilled the fantasies that fear and hate long had prompted in Australia. On February 15, 1942, 15,000 Australians became prisoners of war with Singapore's fall, and four days later war came to the nation's shores, when Darwin was bombed. Then came a Japanese swing southward, by August threatening Port Moresby, New Guinea.

Relations with the United States

The United States became Australia's major ally. In a famous statement (December 1941), Prime Minister Curtin declared: "I make it quite clear that Australia looks to America, free from any pangs about our traditional links of friendship to Britain." A sharper note of independence from Britain came when Curtin insisted (February 1942) that Australian troops recalled from the Middle East should return to Australia itself and not help in the defense of Burma, as British prime minister Winston Churchill wished. Conversely, American needs prompted total response to Curtin's call. U.S. general Douglas MacArthur established his headquarters first in Melbourne and then in Brisbane; the Australian Navy assisted in the U.S. victory in the Battle of the Coral Sea in May, which retrospectively appears a turning point in the war; and the two nations' troops thereafter fought in many joint land battles. The American soldier became a common figure in the Australian capitals, forging the biggest single link in the social relations between the two countries (although not always a harmonious one, as competition for girls and grog sparked jealousy).

On land, the fortunes of war turned against the Japanese in August–September 1942, beginning with an Allied victory at Milne Bay, New Guinea. More prolonged—and of more heroic dimension in Australian eyes—was the forcing back of the Japanese from southern New Guinea, over the Kokoda Trail. Then followed a long attrition of Japanese forces elsewhere in New Guinea and the islands, with Australia playing a role secondary to American forces but nevertheless significant—initially in New Guinea, and at war's end especially in Borneo. Australian volunteers and conscripts fought in these campaigns, the government and people having accepted the legitimacy of sending conscripts as far as the Equator and between the 110th and 159th meridians.

The war brought some passion into domestic affairs, albeit less than that of 1914–18. Curtin's government exercised considerable control over the civilian population, "industrial conscription" being scarcely an exaggerated description. Overall this was accepted—partly because of the sheer crisis, partly because the government showed purposefulness and capacity. Curtin easily won the 1943 elections; thereafter his ministry and the bureaucracy gave considerable thought to postwar reconstruction, hoping to use war-developed techniques to achieve greater social justice in peace.

Wartime growth

The war carried industrialization to a new level. The production of ammunition and other matériel (including airplanes), machine tools, and chemicals all boomed. Meanwhile, primary production lost prestige, aids, and skills, so that the 1944 output was but two-thirds that of 1939–40. Urban employment was bountiful, and concentration in capitals became more marked than ever; many families had two or more income earners. Thus

affluence quickened; federal child endowment from 1940 and rationing of scarce products helped distribute this wealth. The gross national product, estimated at $1,089,000,000 in 1918–19 and $1,860,000,000 in 1938–39, rose to $2,936,000,000 in 1942–43.

The states. As noted above, the period saw a steady loss of state power to the federal centre and a general lessening of local consciousness. Nevertheless, repeated referenda showed the people reluctant to expand commonwealth powers formally. At many levels state loyalties and state governments still mattered most.

New South Wales. New South Wales resumed its primacy as the most populous, wealthy, and industrialized area. Most of the generalizations made in any account of Australia derive from the particular experience of New South Wales; the federal Labor and Country parties both drew their major strength from it. Its internal politics illustrated the common theme that Labor did much better in state than in commonwealth elections. The dominant Labor figure, J.T. Lang, as premier of New South Wales in 1925–27, emphasized child endowment and other welfare services. Again in office in 1930–32, Lang refused to endorse the play-safe policies with which the federal and other state governments responded to the Depression; feeling became high and a "New Guard" organization, led by military officers from World War I, pledged itself to save the state from Lang's radicalism. Finally, the state governor dismissed Lang, who broke also with the federal Labor Party and fought it for years afterward.

Federal parties' major strength

Victoria. Victoria offered much less drama; the dynamism of the colony's first half-century never returned. Deakin, Bruce, and Menzies were all Melbourne men, and Lyons had his power base there—indicative that this city remained the bastion of financial interests. Yet in state politics the Labor and Country parties generally supported one another and the Nationalists did poorly. The generation of electricity from brown coal deposits in eastern Victoria was especially important in maintaining economic development.

The smaller states. The smaller states probably lost economically by federation. Interstate free trade meant that there could be no local tariffs to offset Sydney's and Melbourne's advantages in industrial production. As fiscal matters generally became more complex, the states were left farther behind the commonwealth and more at its mercy. Tension developed, most famously in Western Australia, where in a referendum the people voted for secession in 1933. The federal government then established the Commonwealth Grants Commission, which thereafter administered subsidies intended to ensure approximate equality in living standards.

Each of the smaller states strove vigorously, if not very effectually, to close the gap between itself and New South Wales–Victoria. Chief reliance fell on sugar and cattle in Queensland; timber and gold in Western Australia; apples, other fruit, and hydroelectric power in Tasmania. South Australia sank into deepening gloom until the mid-1930s; revival then came through the success of a campaign to attract industry.

AUSTRALIA SINCE 1945

Social and economic history. The rate of change in Australia after 1939 was greater than that during the preceding 80 years (in both centuries the '50s were especially dynamic). Sheer affluence was an important factor. Australians had long been among the world's wealthy, and, relatively speaking, this situation remained much as before. In absolute terms, however, "middle" Australia acquired appreciably more consumer goods and comforts. Large-scale *embourgeoisement* was all the more significant in a society in which working class attitudes had always set overall norms. By 1970 automobile numbers rose to more than one-third of the population, and three-fourths of all families owned their domiciles. These basic possessions were the lodestones of much Australian life.

In the midst of Australia's prosperity and its relative freedom from tensions and problems, there were those who bewailed the mediocrity of its ruling elites—business, academic, administrative, political. This criticism had considerable force. Wealth seemed as often to dampen as to inspire a drive for excellence. Yet probably there were more exceptions to mediocrity than earlier. Canberra—with its buildings and its institutions, as developed from the mid-1950s—probably had the most excellence to offer: the federal public service was efficient and skilled, while the Australian National University and the National Library of Australia reached international calibre.

The affluence of the 1960s did not reach rural industry. Both the small farmer (his life often a grim compound of debt, drought, and disaster) and the big farmer (who continued to form a wealthy caste atop the social pyramid) seemed certain to diminish; and the life-style of the remainder might have to alter radically. Australian leaders had long worshipped industrialization, and the masses had preferred urban living; still the "bush" had meant much in national mythology, and the shrinkage of its economic importance was thereby significant for all.

Second to affluence in shaping a new Australia was the federal government's assistance of mass European-wide immigration from 1946 until restrictions were imposed in December 1974. Roughly 100,000 migrants entered Australia each year; fewer than one-third came from Britain, about one-sixth from Italy, and almost one-tenth each from Germany and The Netherlands. Thus not only did migration ally with increasing birth rates to increase the population, but the traditional homogeneity of British background (often exaggerated in vernacular, but sizable enough in fact) ended. The inner suburbs of Melbourne and Sydney became polyglot communities. For many years rhetoric insisted that immigration had vitally aided economic development; around 1970 criticism of migration began to develop, as part of a conservation-cum-population-control syndrome, and this included the charge that the costs of immigration at least matched its benefits.

Immigration

At other points, too, ties with Britain loosened. After 1941–42 the Royal Navy was no longer Australia's shield against the world. The United States and Japan became trading partners and investors of comparable weight—and the West Indies became a more stimulating opponent at cricket. Thus, no longer could an Australian speak of Britain as "home" without appearing ridiculous. The change was not absolute—Australia remained an active member of the British Commonwealth—but it was enough to leave a vacuum in Australian life.

Discoveries of minerals accelerated in the 1960s. Iron in Western Australia was the most considerable; bauxite, especially in north Queensland, and nickel, in Western Australia, ranked next. Uranium was a more exotic but transitory resource, while tin, copper, lead, and zinc all rallied. Petroleum and natural gas were discovered in substantial quantities in southern Queensland and off the coast of southeastern Victoria. Japanese capital was particularly important in the development of these resources.

Stock exchange speculation became commonplace in talk and action. The scale of business multiplied. In terms of economic activity and profitability, Australia placed high among the world's second-rank nations.

Government played a part of some significance. The Labor governments before 1949 believed in fiscal supervision as a matter of ideology, and this tendency survived the change of administration. "Protection all round" remained a basic policy, although it had some opponents, and in July 1973 the new Labor government reduced import tariffs by 25 percent. Tariff preferences for less developed nations were instituted in January 1974. Through membership in the General Agreement on Tariffs and Trade (GATT) and other organizations, the government sought beneficial trade agreements, especially for primary resources; marketing boards and commissions came even to include wool in 1970–71. Markets, both for wool and for wheat, were particularly scarce in 1971. Yet investors and managers, including those from overseas, went comparatively without restraint. Legislation directed against monopolies, for example, was insignificant, and overseas investors were not obliged to reinvest profits. Restrictions on foreign investment were imposed by the Labor ad-

Protective tariff policy

ministration early in 1973, but they were cancelled in August 1974.

Industrial relations became a matter of enormous complexity, involving executive and judicial arms of government as well as employer and employee organizations. They seemed to work, with generally less controversy than in earlier periods and in other places, although a strike in the petroleum industry in 1972 was especially disruptive. In May 1975 Australian workers began a trial period of wage "indexation," imposed by the Conciliation and Arbitration Commission, to be based on the consumer price index. The Australian Council of Trade Unions was one of the nation's powers.

Change in attitudes toward Asia

Attitudes toward Asia also changed, in ways difficult to generalize. Arrogant antipathy diminished; sympathy, interest, and knowledge increased. Relations with Japan became vital to the economy, many Asian students attended Australia's universities, and revisions of immigration policy in 1956, 1966, and 1973 caused an increasing number of Asians to be allowed to enter. Yet foreboding remained; its bases were Australia's heritage of Western history and Eastern geography and the end of European, especially Anglo-Saxon, world supremacy.

Not everyone shared the nation's affluence. Some European immigrants worked hard at unpleasant jobs. Overall, about one-tenth of the population might be considered underprivileged, with old-age pensioners suffering most. Social services, though expanded considerably by the Labor government in 1973–75, were not adequate to meet these problems. A new national health scheme called Medibank began operation in July 1975 under the Australian Health Commission.

The Aborigines' experience remained the nation's atrocious scandal. A mining boom in northern Australia continued the long story of territorial expropriation. A mere handful of Aborigines achieved distinction—the best known were Lionel Rose, who became the world bantamweight boxing champion in 1968, and Evonne Goolagong, women's tennis champion. In politics, the professions, or the academy, they had no counterparts. That the Aborigines' subordination derived more from custom than from force scarcely mitigated the situation. Early in 1973 an expanded Aboriginal welfare program was begun (mostly in Queensland), but a year later the program was declared a financial disaster.

Some Aborigines were among those who voiced the Western world's new discontents in the late 1960s. Rejectors of society, above all opponents of Australian participation in the Vietnam War in alliance with the United States, they followed the example of their dissenting U.S. counterparts. But the radical movement was not especially strong. (As with many other sociopolitical movements in Australian history, an analyst must hover between noticing their interest and remarking their weakness in terms of international comparison.)

State boundaries and loyalties suggested a more significant internal division. The crises of the nation's first half-century had forced unity, or at least the acceptance of federal power. The relative calm of later years was accompanied with growing antagonism to "dictation" from Canberra. Growing wealth in Western Australia especially reduced the economic dominance of the southeast corner.

Culture. Painting, sculpture, poetry, and the novel all flourished in the postwar period. The writing of Australian history produced some notable prose. Ray Lawler might have produced "the great Australian play" with *The Summer of the Seventeenth Doll* (1955), its heroes two Queensland sugarcane cutters. Affluence enabled architects to build more private homes of distinction.

Sydney Opera House

Construction of an opera house in Sydney was begun in 1957 and, after expenditure of unprecedented technical ingenuity and A$100,000,000, was finally completed in 1973. Melbourne had its cultural centre that housed an art gallery of splendid quality.

The commonwealth government assisted theatrical work; national opera and ballet companies developed considerable reputations. A Council for the Arts was established in 1967 (with increased authority in 1973) to administer state support for cultural activities. At the same time a small group of established writers and composers was invited to compose a new national anthem; "Advance Australia Fair" was chosen in 1974 to replace "God Save the Queen." The soprano Joan Sutherland achieved overseas fame almost comparable to that of the earlier soprano Nellie Melba.

Universities grew enormously after World War II with the admission of former servicemen. The effort outreached resources, and not until the late 1950s did federal government expenditure redress the situation. The large cities acquired two or three universities, while in the late 1960s colleges of advanced education began to diversify higher education. Meanwhile, governments had reversed the generations-old policy of refusing aid to nonstate (*i.e.*, denominational and especially Roman Catholic) schools. Considerable sums were spent on education, but many found much to criticize.

Scientific research

Medical research advanced notably. The first Australians to win Nobel Prizes, Macfarlane Burnet (1960) and J.C. Eccles (1963), both worked in this area. Astronomy also merits particular emphasis: the Mt. Stromlo observatory was another of Canberra's world-ranking institutions, and Australian scientists were in the forefront of radio and X-ray astronomy. This was one of several fields in which the Commonwealth Scientific and Industrial Research Organisation (see above *Growth of the commonwealth: The postwar years*) contributed much, departing from narrow limits of applied research. The CSIRO and the universities dominated research, but some minor bodies were important: Burnet directed one such, the Walter and Eliza Hall Institute of Medical Research.

Australia was one of the few countries to participate in all the modern Olympic Games; and in 1956, when Melbourne was host city, Australian swimmers dominated the scene. From time to time, golf, track and field sports (especially for women), cycling, and pugilism occupied the spotlight in Australia. European immigrants played soccer, but the old Australians still flocked to see "Australian Rules," especially in Melbourne.

Television, introduced in 1956 and so able to capitalize on interest in the Olympic Games, soon became dominant in everyday entertainment. Many programs came from Britain and the United States, but the Australian contribution slowly increased. Radio diminished as television boomed but later (partly from the use of the transistor) found a continuing audience. The cinema was forced to accept a more modest role. Popular live entertainment followed a different graph: it benefitted much from the boom, especially in Sydney, with clubs that provided liquor, gambling (poker machines), and variety shows.

Domestic politics. J.B. Chifley succeeded John Curtin as Labor prime minister in mid-1945, Curtin having died in office. A growing efficiency and enthusiasm in pursuit of social justice led to extended social welfare, national development, the establishment of the National University, and the provision of university scholarships. Public servants provided the expertise for more sophisticated supervision of the economy. Chifley won the elections of 1946 fairly comfortably, although the government's majority fell.

But the Australian Labor Party soon became the victim of the Cold War. During World War II Communists had increased their influence throughout Australian society and especially within the labour movement. Many believed that the Labor Party itself was Socialist. This judgment exaggerated the truth, but Chifley's belief in controlling the economy gave it some force; when he planned to nationalize all banks (1947), the reaction was accordingly intense. Although declared unconstitutional in 1948, this banking nationalization counted heavily against the government in the electorate's eye.

Meanwhile, Communist influence was in large measure responsible for many strikes; the culmination came in the coalfields of New South Wales in 1949. These strikes also aroused public concern and hostility. The Labor government suffered heavy defeat in the election of December 1949. The defeat was intensified by a restructured elec-

toral system: thenceforth the House of Representatives had about 120 members, representing single-member electorates of approximately equal size; the Senate had 60 members, each state electing five members every three years by proportional representation, the members serving for six-year terms.

Menzies and the Liberal Party

Liberal and Country party coalition governments ruled continuously after the 1949 election until 1972. The Liberal Party, formed in the mid-1940s, was largely based on the erstwhile United Australia Party. Its founding genius was Robert Menzies, and as its head he was prime minister from 1949 until 1966. Menzies and his governments provided the framework for the development of contemporary Australia.

Reverberations of the Cold War helped Menzies stay in office. In 1954 the defection from the Soviet Embassy in Canberra of the Soviet agent Vladimir Petrov led to a royal commission into alleged espionage within Australia. This not only revitalized the Communist threat, which Menzies often invoked, but prompted the Labor opposition to defend some of those implicated in the commission's inquiries. This in turn antagonized the right-wing and Roman Catholic elements in the Labor Party—the groups that had been the most vigorous elements within the Labor Party in opposing Communists in the trade unions. In 1955 this group established the Democratic Labor Party (DLP); although it won few parliamentary seats, the DLP took enough votes from the Labor Party to lessen the latter's chance of federal office.

The DLP contributed what little ideology there was in Australian politics after 1950. The Labor Party increasingly disowned any Socialist belief but found no alternative doctrine; two successive party leaders, Herbert Vere Evatt in 1951–60 and Arthur Aloysius Calwell in 1960–67, were frustrated in spending years as leaders of the opposition. The Communist Party split between Soviet and Chinese factions, each small and increasingly sterile.

In the late 1960s politics took a bizarre turn. Harold Holt, Menzies' successor as prime minister, had scarcely established himself when he died by accidental drowning (December 1967). There followed a Byzantine contest for his office, won by John Gorton. But Gorton failed to provide solid, consistent leadership, and early in 1971 the Liberal Party replaced him with William McMahon. In

Labor Party in power again

December 1972 a new Labor government under Gough Whitlam took power and entered upon a program of social, economic, and political reform.

The Labor program included a large increase in appropriations for social welfare, including a national health service, urban development, a doubling of aid to education, and the removal of investment incentives for the mining industry. Opposition to Whitlam's program was effective in the Senate as early as the following May in its defeat of his electoral reform bill; in December 1973 the nation defeated a referendum that would have permitted federal price and wage controls. Increased spending meant increased inflation, and this became a campaign issue in April 1974 when Whitlam dissolved Parliament in the hope of winning a majority in the Senate. He succeeded in eliminating the DLP from the Senate and increasing his numbers so that he could pass his program through a joint sitting of Parliament. A sudden increase in unemployment in July 1974, a 16 percent loss in state elections in Queensland in December, and a financial scandal in the Cabinet in early 1975 plagued Whitlam's Labor government.

New leadership came to the opposition in March 1975 with the election of Malcolm Fraser to head the Liberal Party. By October 1975 Labor strength in the Senate had been reduced by two states having filled vacancies with members from the opposition parties (an unconventional practice). Fraser took advantage of the situation to refuse to pass necessary appropriations bills, thereby forcing another election. Whitlam refused to submit to this pressure, so the Whitlam-appointed governor general, Sir John Kerr, dismissed the prime minister—the first time a representative of the crown had dismissed an elected prime minister in 200 years. With assurances from Fraser that the appropriations bills would be passed, he was made caretaker prime minister, and a new election for both houses was set for December 13. A bitter campaign ensued, with inflation and unemployment as the primary issues. The results were sweeping majorities for the Liberal–Country party coalition. With the loss of more than 30 seats, Labor was reduced to a weak opposition, and the equilibrium of the Menzies era of the 1950s and early '60s was returned.

Foreign affairs. The Labor governments of the 1940s, and especially Labor's minister H.V. Evatt, hoped that Australia would play an independent, substantial part in creating a stable international situation. Thus it was active in support of the early United Nations, Evatt himself serving a term as its president in 1947. Support of Indonesian independence was another important policy of those years, and there was advance toward building a civil service that was expert in international affairs.

Support of the United Nations

These positive trends continued with the Liberal victory of 1949. It was the next year, for example, that Australia signed the Colombo Pact, under which substantial aid, especially through education, was sent to Asian countries. Australia remained an active supporter of the United Nations and other international agencies. Menzies, especially, was a great admirer of Britain and its empire-commonwealth.

Meanwhile, the movement of world politics soured the optimism of the 1940s. Turbulence in Asia, especially a Communist victory in China, gave a new edge to Australia's traditional fear of attack from the north. In response, the country moved increasingly close to the United States, which alone offered the possibility of replacing Britain as a protector. Thus Australians fought along with United Nations troops in Korea; in 1951 the country accepted the U.S. view of Japan as a bulwark against Communism, while simultaneously the ANZUS (Australia, New Zealand, United States) Pact promised U.S. help in case of attack; and in 1954 Australia joined the Southeast Asia Treaty Organization (SEATO), which extended U.S. responsibility throughout the area. The U.S. built significant missile bases throughout Australia, and Australia's increasingly large defense expenditure bought much U.S. equipment. Australia also supported the U.S. commitment in Vietnam, and it sent troops (including conscripts) to fight there, but these troops were withdrawn between August 1971 and February 1972.

With the return of Labor to power in December 1972, a much greater emphasis was put on relating Australia to Asia. The People's Republic of China was recognized immediately and North Vietnam in February 1973. Whitlam visited Indonesia in February and India in June, at which time a joint statement was made that the Indian Ocean "should be free from international tensions, great-Power rivalry, and military escalation." A visit to China followed in November. Closer relations with the new Labor government of New Zealand began at the outset, and ties to Britain were lessened, including reduced military support in Singapore and Malaysia. Early in 1974 Whitlam toured six nations in Southeast Asia, again emphasizing his government's interest in the Eastern Hemisphere. At the end of the year Whitlam visited seven of the nine members of the European Economic Community, declaring that Australian prime ministers had neglected this important area, second only to Japan as Australia's trading partner. He also visited Yugoslavia and the Soviet Union.

Papua New Guinea, composed of the Australian external territory of Papua and the Australian-administered UN Trust Territory of New Guinea, became a self-governing state in December 1973. After several postponements, it achieved complete independence in September 1975. Papua New Guinea was the largest recipient of Australian aid.

BIBLIOGRAPHY

Short histories and commentaries: C.M.H. CLARK, *A Short History of Australia* (1963); G. GREENWOOD (ed.), *Australia: A Social and Political History* (1955, reprinted 1966).

References: Australian Dictionary of Biography (1966–); A.H. CHISHOLM (ed.), *The Australian Encyclopaedia,* 10 vol. (1958).

Cultural surveys: R. COVELL, *Australia's Music* (1967); H.M. GREEN, *A History of Australian Literature, Pure and Applied*, 2 vol. (1961); R. HUGHES, *The Art of Australia*, rev. ed. (1970).

Economic histories: J. GRIFFIN (ed.), *Essays in Economic History of Australia*, (1969); A.G.L. SHAW and H.D. NICOLSON, *Growth and Development in Australia* (1964).

Histories of specific periods: (to 1830): C.M.H. CLARK, *A History of Australia*, vol. 1 and 2 (1962–68); E.M. O'BRIEN, *The Foundation of Australia (1786–1800)*, 2nd ed. (1952); C.A. SHARP, *The Discovery of Australia* (1963); A.G.L. SHAW, *Convicts and the Colonies* (1966). *(1830–60)*: P. BURROUGHS, *Britain and Australia, 1831–1855* (1967); M.L. KIDDLE, *Men of Yesterday: A Social History of the Western District of Victoria 1834–1890* (1961); A.C.V. MELBOURNE, *Early Constitutional Development in Australia*, 2nd ed. by R.B. JOYCE (1963); G.H. NADEL, *Australia's Colonial Culture* (1957); S.H. ROBERTS, *The Squatting Age in Australia, 1835–1847* (1964); A.G. SERLE, *The Golden Age . . . Victoria, 1851–1861* (1963). *(1860–1900)*: N.G. BUTLIN, *Investment in Australian Economic Development, 1861–1900* (1964); G.L. BUXTON, *The Riverina, 1861–1891* (1967); M. CANNON, *The Land Boomers* (1966), on speculation in Melbourne; R.A. GOLLAN, *Radical and Working Class Politics . . . 1850–1910* (1960); J.A. LA NAUZE, *Alfred Deakin*, 2 vol. (1965); J.M. TREGENZA, *Professor of Democrary . . . C.H. Pearson, 1830–1894* (1968); D.B. WATERSON, *Squatter, Selector, and Storekeeper: A History of the Darling Downes, 1859–93* (1968). *(1901–45)*: F. ALEXANDER, *Australia Since Federation* (1967); C.H. FORSTER, *Industrial Development in Australia, 1920–1930* (1964); B.D. GRAHAM, *The Formation of the Australian Country Parties* (1966); E. SCOTT, *Australia During the War*, 2nd ed. (1936); G. SAWER, *Australian Federal Politics and Law*, 2 vol. (1956–63); I.A.H. TURNER, *Industrial Labour and Politics, 1900–1921* (1965). *(1945 to the present)*: T.B. MILLAR, *Australia's Defence Policies, 1945–65* (1967); R. MURRAY, *The Split: Australian Labor in the Fifties* (1970); K. TENNANT, *Evatt: Politics and Justice* (1970); K. WEST, *Power in the Liberal Party* (1966).

(Ml.R.)

Australian Aboriginal Cultures

Before the arrival of Europeans in Australia, Aborigines occupied the entire continent, including Tasmania. The usual estimate of their numbers at that time is about 300,000. Population density was highest in more fertile riverine and coastal areas; in arid zones, the exigencies of gaining a livelihood made distribution over much larger territories necessary. These broad environmental differences had direct sociocultural implications for mobility and for control over natural resources. Aborigines in the drier inland moved camp more frequently, had a narrower margin of survival in bad seasons, and relied more directly on ritual supplication as a necessary aid to everyday living.

There were approximately 500 tribes and subtribes—tribe being a convenient label for a constellation of interacting groups the members of which acknowledge certain features in common. Most tribes had specific names for themselves; other names, sometimes less flattering, were applied to them by their neighbours. They occupied recognized territories and were distinguished by characteristic features in culture, language, or dialect.

Early writers on the subject spoke of "nations," usually referring to a group of tribes identifying themselves by the same word for "man." But, although there was interaction between neighbours, there was no direct social communication across the continent, and the Aborigines were certainly not a nation in the modern sense: they had no overarching political system and no overall name.

TRADITIONAL SOCIOCULTURAL PATTERNS

All Aborigines, coastal and inland, were directly dependent on their natural environment. They were semi-nomadic, wandering across limited stretches of the country, hunting, and collecting food. This basic economic circumstance had several concomitants: (1) food-gathering groups were fairly small and usually scattered, and all members of a tribe rarely came together, even for ritual occasions; (2) cooperation was essential for survival; (3) over and above particular techniques and skills, Aborigines believed that something else was required for survival, namely religion.

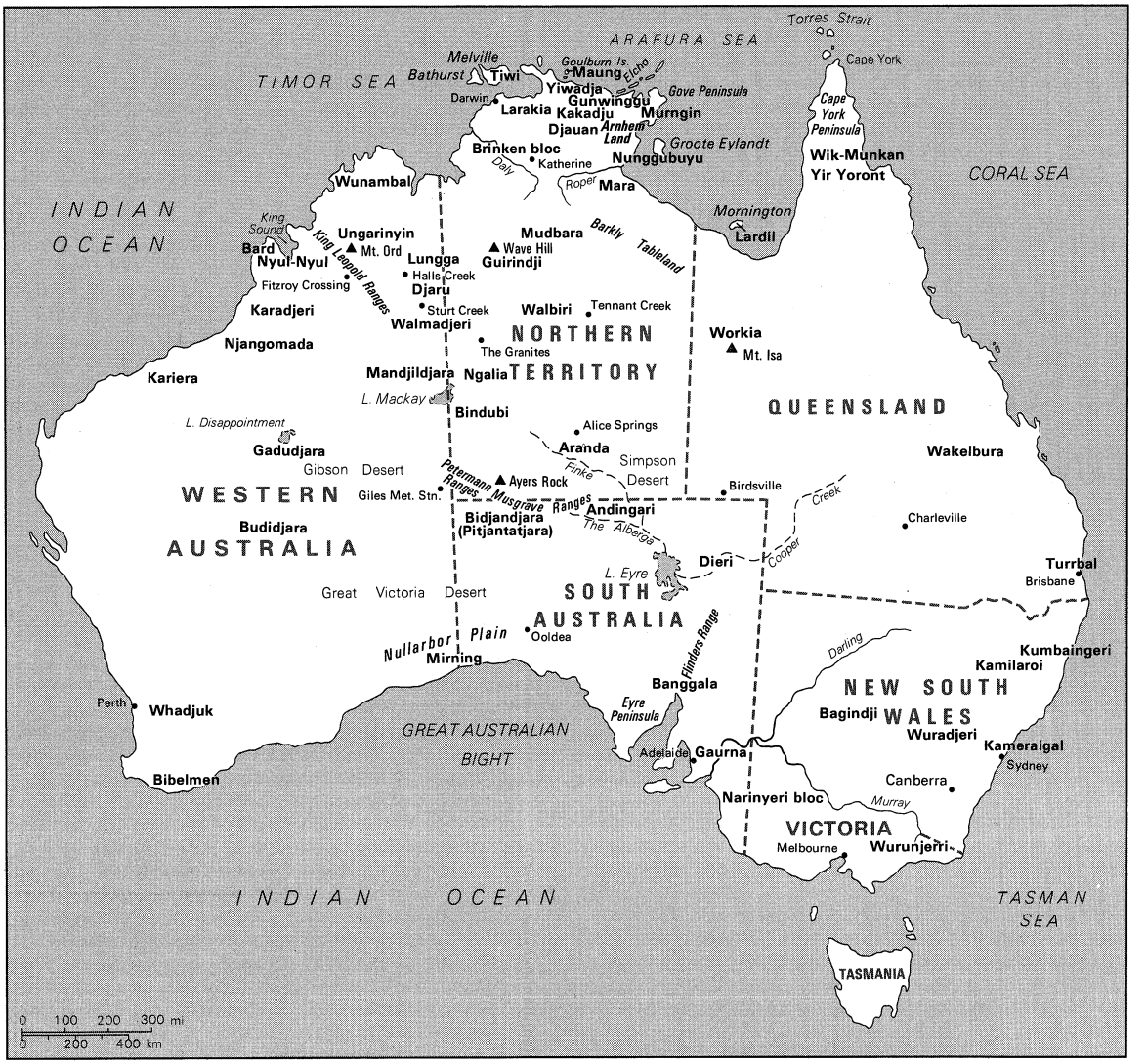
Aboriginal cosmogony

The Aborigines' view of their place in the world was summarized in their concept of "the dreaming," or "eternal dreamtime." This does not refer to dreams in an ordinary sense. In one of its meanings the dreaming was the formative or creative period at the dawn of time, when mythic beings shaped the land, brought various species into life, and established human life and culture. Although these mythic beings died or were transformed, they lived on eternally in spirit. They left tangible evidence of their physical presence on earth and were actually identified with or manifested through particular species and natural elements. The Aborigines also believed that these spirits were manifested through and in man, that the mythic creatures of the dreaming were not basically different from man, and that man, supernatural beings, and natural species were interdependent and mutually sustaining. Hence, an Aboriginal person was never isolated; he saw himself as acting and interacting with others, and the bonds of kinship were extended outward, embracing the nonhuman and nonempirical world.

Social groups and categories. The main issues in Aboriginal social life were religion and economics; these were seen as interdependent and were reflected in the two kinds of social units to which all Aborigines traditionally belonged. One of these social units was the local descent group, always patrilineal (descent through the male line), exogamous (marrying outside the group), and associated with a particular stretch of territory—the "estate." Its adult males were responsible for the upkeep of the traditional campsites and for the appropriate ritual —that is, for renewing and sustaining the land. Ownership of land was nontransferable: its members held land in trust collectively through an unwritten charter deriving from the dreaming. The second, larger kind of unit was the horde or band, the group concerned with the ordinary practical business of getting a living. This was a land-occupying and land-utilizing group, exploiting all the available natural resources. It comprised members of several families—that is, of more than one local descent group; and the range over which they normally travelled included several estates. It fluctuated in size and composition; in good seasons, a larger number of people could camp together.

In addition to membership in a local descent group and horde, an individual identified himself by reference to larger social groups: moieties, semimoieties, sections, subsections, and clans. A moiety is one of two basic complementary parts into which a tribe divides itself. Sections and subsections are further subdivisions, splitting the tribe into four and then eight parts. Such social divisions among the Aborigines governed marriage and other contracts. A person in one moiety, for instance, could marry only a person in the other moiety. When there were section or subsection systems, everyone was incorporated in such a system from before birth. The main advantage was that it represented a widely understood grid for categorizing everyone a person was likely to meet. With this went a shorthand statement of basic rules or conventions, which was a simple guide to everyday behaviour, especially useful for classifying distant kin and strangers. Close, genealogically traceable kinship virtually always took precedence over nominal kinship when priorities were at issue.

Aboriginal Australia was not socially stratified in the sense of having more or less fixed classes or strata arranged hierarchically on the basis of descent or acquired status. Status was clearly marked only within the religious sphere; females, for example, were mostly excluded from an executive role in secret-sacred ritual, and areas of privilege were further defined by graded acceptance of youths and adults passing through rites of learning. Essentially, however, Aboriginal society was "open"—there were no social barriers operating against a man becoming a leader in religious matters. Such achievement depended primarily on his own efforts, his kin's, and his observance of ritual obligations. Men acquired prestige through knowledge of religious practices and expertise in directing ritual. In matters of economics, entrepreneurial leaders did have considerable status in one or two areas—



Original distribution of the larger Aboriginal tribes.

among the Tiwi and Murngin, for example—but this was derived not so much from commercial activities as from their ability to manipulate certain kin, to arrange polygynous marriages for themselves and remunerative betrothals for their children, and to establish a network of alliances.

Kinship, marriage, and the family. The Aborigines were deeply conscious of social relations, especially of kinship. All Aboriginal kinship systems were basically classificatory—that is, a limited number of kin terms was extended to cover all known persons: terms for lineal relatives, such as "father," also referred to collateral relatives, such as uncles. This did not mean that a person could not recognize his actual father, mother, son, or other relative. The terms simply indicated the kind of behaviour that was appropriate—filial, brotherly or sisterly, fatherly, or whatever.

Function and significance of kin terms

Kin terms were behavioural signals, indicating, for example, the expectation of sexual access, restraint, avoidance, or responsibilities. Affines (relatives by marriage) were often classified with consanguineous (blood) relatives, although qualifying terms might be used. Certain terms indicated potential spouses or affines. A husband and wife were ideally, before marriage, always related to each other as kin, either actual or classificatory. Some relationships were more prominent than others. The brother–sister relationship, for example, was one of the most emotionally charged and often marked by some form of avoidance. The most outstanding avoidance relationship was between a man and his actual or potential mother-in-law—not just his wife's mother but all women and girls who on the basis of kinship and the recognition of preferred potential spouses, might become his mother-in-law.

Reciprocity was a fundamental rule in Aboriginal kinship systems and also in marriage. Marriage was not simply a relationship between two persons; it linked two families or groups of kin, which, even before the union was confirmed and most certainly afterward, had mutual obligations and responsibilities. Generally, throughout Aboriginal Australia those who received a wife had to make repayment either at the time of marriage or at some future time. In the simplest form of reciprocity, men exchanged sisters; and women, brothers. Such exchanges took place between different moieties, clans, or local descent groups or between certain types of kin. Most kinship-and-marriage systems provided for the possible replacement of spouses and for parent surrogates. At any one time, a man or woman had available a number of persons, who, because of their kinship, were termed spouses. Access to them depended on several factors, but, as far as premarital and extramarital liaisons were concerned, it was conventionally in this direction that partners should be sought.

Infant betrothal was common. If arranged before the birth of one or both of the prospective spouses, it was a tentative arrangement subject to later ratification, mainly through continued gift-giving to the girl's parents. In some Aboriginal societies, parents of marriageable girls played one man against another, although this was always a potentially dangerous game. Also, there might be a considerable age discrepancy between an affianced

pair. Generally, a long-standing betrothal, cemented by gift-giving and the rendering of services, had a good chance of surviving and fostering a genuine attachment between a couple.

For a marriage to be recognized it was usually enough that a couple should live together publicly and assume certain responsibilities in relation to each other and toward their respective families; but it might be considered binding only after a child was born. All persons were expected to marry. A girl's marriage should be settled before she reached puberty, and, ideally, a husband should be older than his wife.

Apart from formal betrothal, there were other ways of contracting marriages, such as elopement, capture during feuding or fighting, and redistribution of widows through the levirate (compulsory marriage of a widow to her deceased husband's brother) or patriate (distribution of a father's widows to his sons, unless they were actual or close mothers). Elopement was often supported by love magic, which emphasized romantic love, as well as by the oblique or direct approval of extramarital relations.

Polygynous marriage

Although most men had only one wife at a time, polygyny was considered both legitimate and "good." The average number of wives in polygynous unions was two or three. The maximum, in the Western Desert, was five or six; among the Tiwi, 29; among the Murngin, 20 to 25, with many men having 10 to 12. In such circumstances, women had a scarcity value. Having more than one wife was usually a matter of personal inclination, but economic considerations were important; so were prestige and political advantage. Some women pressed their husbands to take an additional wife (or wives), since this meant more food coming into the family circle and more baby-sitters. To terminate a marriage, a woman might try elopement. A man could bestow an unsatisfactory wife on someone else or divorce her. A formal declaration or some symbolic gesture on his part might be all that was necessary. In broad terms, a husband had more rights over his wife than she had over him. But, taking into account the overall relations between men and women, their separate and complementary spheres of activity in marriage and in other aspects of social living, the status of women in Aboriginal society was not depressed.

Socialization. A child's spirit was held to come from the dreaming to animate a fetus. In some cases, this was believed to occur through an action of a mythic being who might or might not be reincarnated in the child. Even when Aborigines acknowledged a physical bond between parents and child, the most important issue for them was the spiritual heritage.

In his early years a child's focus was on his actual parents, and especially on his mother, but there were others close at hand to care for him. Weaning occurred at about two or three years of age but occasionally not until five or six for a youngest child. Through observation of camp life around him and informal instruction, a child built up his sociocultural perspective, learning through participation. At the same time, he became familiar with his natural environment. Small children often went food collecting with their mothers and other women; as girls grew older, they continued to do so, but boys were thrown more on their own resources. Parents were, on the whole, very indulgent. Infanticide, even in arid areas, was much rarer than has been suggested.

Children learned quite early with whom they were allowed to play and whom they must avoid; brothers and sisters, for instance, or "mothers-in-law" and "sons-in-law," would not normally play together. Betrothal put special restraints on a girl, and, at or even before puberty, she normally went to live with her husband, assuming the status of a married woman.

With initiation, a boy's life changed drastically. His formal instruction as a potential adult began, and he was prepared for his entry into religious ritual. His future was henceforth in the hands of older men and ritual leaders who exercised authority in his community. But he was not among strangers: the relatives who played an active role in his initiation would also have significant roles in his adult life. A boy's age at the first rite varied: in the Western Desert it was about 16; in the Kimberleys, about 12; in northeastern Arnhem Land, six to eight; and among the Aranda 10 to 12 or even older. Generally, once he had reached puberty and facial hair had begun to show, he was ready for the initial rituals.

Initiation in Aboriginal Australia was a symbolic reenactment of death in order to achieve new life as an adult. As a novice left his camp, the women would wail and other noises would be made, symbolizing the voice of a mythic being who was said to swallow the novice and later vomit him forth into a new life. The initiation rites themselves were a focal point in discipline and training; they included songs and rituals having an educational purpose. All boys were initiated, and traditionally there were no exceptions.

Initiation rites

Circumcision was one of the most important rites over the greater part of Australia. Subincision (incisure of the urethra) was especially significant in its association with secret-sacred ritual. Other rites, according to the area, included piercing of the nasal septum, tooth pulling (in New South Wales this was central in initiation), and the blood rite, which involved bloodletting from arm vein or a penis incisure—the blood being used for anointing or sipping (red ochre was used as a substitute for blood in some cases). Hair removal, cicatrization (scarring), and playing with fire were also fairly widespread practices. Among the Aranda, fingernails were torn out and heads gashed open and bitten. All such rites were usually substantiated by mythology.

For girls, puberty was marked by either total or partial seclusion and by food taboos (also applied to male novices). Afterward, they were decorated and ritually purified. Ritual defloration and hymen cutting or both were widspread (in the Western Desert, for instance), while on Groote Eylandt the labia majora were removed.

Boys, after circumcision, became increasingly involved in adult activities. Although they were not free to marry immediately, even if they had reached puberty, they might do so after undergoing certain rites, such as subincision. Initiation was a prelude to the religious activity in which all men participated. It meant, also, learning a wide range of things directly concerned with the practical aspects of social living. Adulthood brought increased status but added responsibilties. A vast store of information had to be handed down from one generation to the next. Initiation served as a medium for this, providing a basis of knowledge upon which an adult could build. This was a process that continued through life and that was especially marked in men's religious activity.

For Aborigines birth and death were an open-ended continuum: a spiritual religious power emerged from the dreaming, was harnessed and utilized through initiation (as symbolic death–rebirth) and through subsequent religious ritual, and finally, on death, went back into the dreaming. Life and death were not seen as being diametrically opposed: the dreaming provided a thread of life, even in physical death.

Social control. Pressures toward conformity began and were formalized during initiation. There were rules and standards, right and wrong ways of doing things, and some allowance for variation. Sanctions upheld the accepted codes and were either positive or negative or a combination of both. Probably the strongest were fear of supernatural punishment and fear of sorcery.

Traditionally, most dissension arose over women, religious matters, and death. Women had trouble with husbands, eloped, and engaged in unsanctioned extramarital liaisons. Such behaviour could mean serious fighting, involving relatives of the parties concerned. Infringement of sacred law was less direct in its social repercussions but was nevertheless regarded as the most serious of all. In many cases, an ordinary or accidental death had wide ramifications, particularly if there were accusations of sorcery. An inquest was held, and, through divination, a supposed "murderer" was found. Punitive measures might or might not be taken against him.

The maintenance of law and order was quite narrowly

localized. Authority was limited and qualified by kinship claims. Precedents were sought in order to guide or influence actions resulting from a breach, and in all tribes there were approved procedures for maintaining the peace. There were no judicial bodies as such, though on the lower Murray River a formal council, or *tendi*, of clan headmen and elders did arbitrate disagreements between adjacent tribes. Generally, simple informal meetings of elders and men of importance dealt with grievances and other matters. There was also settlement by ordeal—the most outstanding example of this sort being the *magarada*, or *maneiag*, of Arnhem Land. During a ritualized meeting, the accused ran the gauntlet of his accusers, who threw spears at him; guilt was proved if the accused was wounded in the thigh.

Maintenance of law and order

Although it is inaccurate to speak of a gerontocracy in Aboriginal Australia, men of importance were easily distinguished. They were usually "elders," who had this status not necessarily because of their age or grey hair but because of their religious position and personal energy. This gave them considerable control over the affairs of others in matters not directly related to religion.

Economic organization. Dependence on the environment and on its resources was reflected in seminomadic living. The Aborigines had to be intimately acquainted with all the country within their range of movement. At the same time, their economic system was not conceptually separate from their religious system. Both before and after initiation, graphic stories (or myths) served as guides to specific localities, detailing what people could expect to find there. In other words, information used in obtaining a living was preserved and transmitted in a religious context. Religious ritual, furthermore, was believed to ensure the maintenance of the natural species and the proper fluctuation of the seasons.

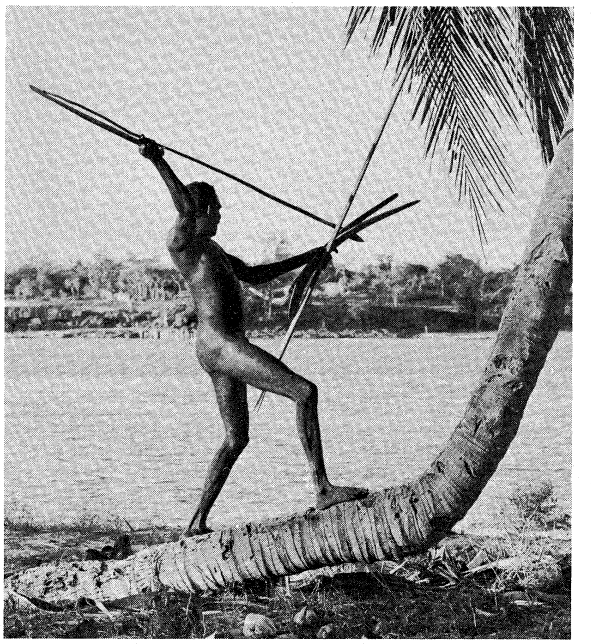
Before modification of their economic organization as a result of contact with Europeans, large groups of people did not remain in the same camp over a prolonged period. There were two basic patterns. Generally, in fertile regions of Aboriginal Australia, time-honoured camping areas were recognized, always in close proximity to water and usually with mythological associations; they were places where people always camped at certain times of the year. Camps were bases from which people made forays into the surrounding bush for food, returning in the late afternoon or spending a few days away. The second pattern involved a much larger territory in arid or desert areas across which Aborigines moved from waterhole to waterhole along well-defined tracks in small family groups. The whole camp moved and rarely established bases. Only in good seasons and at sizeable permanent waters was it possible for a large number of people to remain for an extended period.

These two patterns were reflected in domestic arrangements. In the north, people made bark shelters, and during the monsoonal rains used caves and stilted huts as protection against flood, mosquitoes, and sandflies. In the desert, windbreaks—bough shelters or saplings covered with brush or bark—were common. During fine weather most Aborigines preferred to sleep in the open, with a windbreak; when it was too cold, dogs helped to provide warmth. Fires were kept alight, and, when moving from one place to another, or even when hunting, people carried live fire sticks.

Aborigines, with rare exceptions (southwest of Darwin and northeastern Arnhem Land, for example), had no form of cultivation and no domestic animals, except the tamed dingo. They preferred to go naked, except for small pubic coverings or decorations. In certain areas, fur cloaks were a protection against cold.

Tools and artifacts

Outside the sphere of religion, material objects were minimal. Grinding stones and platters were left at certain camps, ready for use. Men carried spears and spear throwers and, in some areas, boomerangs. There were bark canoes and rafts and dugout log canoes, some with pandanus-mat sails. Women's digging sticks could double as fighting weapons. Their large deep wooden dishes held seeds, vegetables, or water, even babies. In some areas, painted bark baskets, plaited pandanus bags, and net bags served the same purposes. Rarer objects were the kangaroo-skin waterbags of the arid central areas and the skull drinking vessels of the Coorong in South Australia. Implements included a large selection of stone tools, wedges, bone needles, bobbins, and sharkskin files.



Aboriginal spearfishing near Darwin, Northern Territory.
Douglass Baglin

Generally, men were hunters, women food collectors; and, unless they were on the move, they carried out these tasks separately. A woman's economic pursuits were focussed on her own family and children. A man's obligations were wider. Providing for his own family was often of secondary importance: ideally, his ritual duties and his network of reciprocal obligations took priority. He was expected to provide meat for his family; but women bore the main responsibility, supplying vegetable foods and small game or shellfish. Although this was primarily a subsistence economy, even life in the desert was not reduced to a constant struggle for the bare essentials. It nevertheless called for skill, tenacity, and continuing social involvement. A person could not "opt out" and survive. Reciprocity was basic to economic life, and all adults were caught up in a web of rights and duties.

As far as trading was concerned, goods passed along defined routes from one group to another in an intricate crisscross patterning over the continent. Boomerangs, for example, went in one direction, red ochre in another; pearl shells from the Kimberleys found their way, gradually, to the Great Australian Bight; and central Australian shields appeared in the Canning Stock Route area.

In traditional Aboriginal society acquisition and gain were important, but the idea of making a profit was almost irrelevant. What really counted was the social relationship inherent in the exchange. The more organized trading partnerships entailed a debtor-creditor association; but prestige still hinged on retaining an equitable balance in the exchanges, on "solvency," and on not dropping out of the network. Private property was recognized, the claims of others upon an item not affecting an individual's right to exclusive use of it until such time as he decided to relinquish it. Traditionally, however, land was inalienable, being held in trust by members of various local descent groups.

BELIEF AND AESTHETIC VALUES

Religion. The Aborigines were concerned that the seasons should come and go in an orderly, predictable way; that human life should continue; and that they themselves should go on living in much the same way as they had always done. Religious beliefs and values gave tradition its power and force in influencing life in the present.

Totemism

Both before and since the French sociologist Émile Durkheim's study of *The Elementary Forms of the Religious Life* (1915), much has been written about Australian totemism. The concept rests on the belief that human beings are an integral part of nature, like all other living things. Totemism has been defined as a representation of the universe, seen as a moral and social order; as a philosophy that regards man and nature as one corporate whole; or as a set of symbols forming a conventional expression of the value system of a society. Such symbols provided intermediate links, both personal and social, between man and the mythic beings. Many of the mythic beings in Australia were "totemic" in the sense of exemplifying in their own persons, in their outward form, the common life-force pervading particular species. Others, originating in human or near-human form, at the end of their wanderings entered some physiographic feature or were metamorphosed as hills or rocks or turned into various creatures or plants. Totemism is another illustration that a sharp dichotomy between sacred and mundane cannot be drawn in Aboriginal Australia, where religion was directly relevant to everyday living.

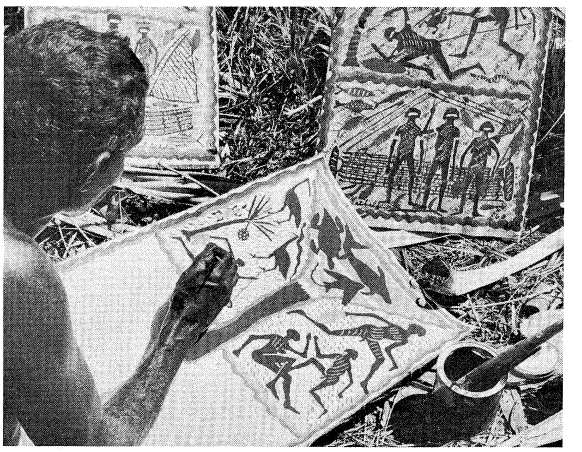
Among most Aboriginal tribes, the main ritual roles in the big religious sequences were reserved for men. Ordinarily, in such cases, women were forbidden to enter ritual grounds, and what went on there was secret and sacred. But women usually had complementary actions to perform or observed special taboos. In other rituals both men and women participated. Among some tribes, women had their own secret-sacred rites. Generally, any large ritual affair concerned a whole community and involved all its members in some measure.

Aesthetics. Sacred ritual provided immense scope for aesthetic expression, especially in dramatic performances with stylized posturing and complicated dance movements. Less intense but sometimes almost as elaborate were the nonsacred ceremonies (or corroborees) designed for entertainment and relaxation. Songs ranged in style from the succinct verses or couplets of central Australia and the Western Desert, which were made up of three, four, or more words repeated in linked sequences, to the more elaborate songs of northeastern Arnhem Land, which were long verses building up complex word pictures through symbolic allusion and imagery. There was no poetry in terms of spoken verse, but there were chants, some of them outstandingly beautiful. The majority of secret-sacred songs comprised mythic cycles, each containing several hundreds of verses. There was also a wide repertoire of songs on everyday events, such as the "gossip" songs of western Arnhem Land, composed by songmen with the aid of spirits. Instrumental music in the north was provided by the didgeridoo (drone-trumpet) and by clapping sticks. In southern and central regions, boomerangs or clubs were rhythmically beaten together or pounded on the ground; in southeastern Australia, women used skin beating pads. Tunes and rhythms varied greatly from area to area.

Songs, chants, and instrumental music

Oral literature was rich. In addition to sacred mythology, there were ordinary stories and tales, either historically true or presumed to be true. Some existed in several versions, depending on the situation in which they were told and on the individual background of the storyteller.

Each cultural area had its own distinctive style of art. *Tjurunga* (sacred object) art, consisting of incised patterns on flat stones or wooden boards, was representative of a large area of Australia, although centralized in Aranda territory. In central Australia, body decoration and elaborate headdresses on ritual occasions, using feather down, blood, and ochres, were especially striking. Everywhere, sacred ritual provided the incentive for making a large variety of objects—mostly impermanent, because the act of making them was itself one of the appropriate rites. In western Arnhem Land, *maraiin* objects—realistic and stylized carved representations of various natural species—were made. The *rangga*, or ceremonial poles, of eastern Arnhem Land, many of durable hardwood, bore ochre designs and long pendants of feathered twine. For mortuary rituals, the Tiwi made large wooden grave posts, and shaped and decorated receptacles for bones were common in eastern Arnhem Land. Also common were carved wooden figures of mythic beings and of contemporary persons, some used in sacred ritual, others as memorial posts for the dead.



Aboriginal artist working on a bark painting, Queensland.
Douglass Baglin

Paintings in ochre on sheets of bark were indigenous to Arnhem Land, although examples could be found in the Kimberleys and in southeastern Australia. They were used mostly on the initiation ground for the instruction of novices. In western Arnhem Land, naturalistic patterns showing figures against an open background were the norm; there was also a unique kind of "X-ray" art that depicted the internal organs of animals and human beings. Also found in various parts of Australia were cave and rock paintings or engravings.

For an outstanding artist or songman, the rewards lay mainly in prestige and in religious advancement. He was not absolved from earning a living in the ordinary way and shouldering routine responsibilities.

ABORIGINES IN AUSTRALIAN SOCIETY

The heritage of early alien contact. Traditionally, the Australian Aborigines had little knowledge of other peoples and were ill-equipped to withstand the rapid inroads of an alien and culturally different people.

Early contact with Indonesians

Two qualifications are necessary here. First, the Aborigines excelled in adjusting themselves to their natural environment. Secondly, in the north they came into contact with Indonesians at an early date (at least two to four centuries ago). The visitors came as sea-going traders to the Arnhem Land coast and made a powerful impact on local art, music, ritual, and material culture. On Cape York Peninsula, the influence of New Guinea and the Torres Strait Islands was evident in the adoption of masked ritual dancing and the use of the drum.

But the coming of Europeans, in 1788, was entirely different. Initially, many Aborigines were willing to welcome them, even regarding them as returning spirits of the dead or as manifestations of the mythic beings. European settlement soon expanded, however, making inroads into tribal territory and interfering with natural resources, excluding the Aborigines from their sacred and other areas and drawing them—often forcibly, always uncomprehendingly—into the life of the developing colony. Communication was minimal. Clashes marked virtually all situations in which Aborigines and Europeans pursued conflicting interests. In the period of "pacification by force," up to the 1880s, a large number of Aborigines were killed. Others were driven into the bush or remained in small pockets subject to the "civilizing" influence of missions, or were left to fend for themselves in the fringe settlements of cities and towns; still others remained in camps or pastoral and cattle stations to become the nucleus of a labour force. Diseases took their toll. The Colonial Office in London had instructed that the Aborigines' rights should be safeguarded and that they were to be accorded the benefits of British subjects. The actual situation was different.

The Aborigines reacted with spasmodic guerrilla warfare and stock killing or with passive resistance. But they soon learned that the only alternative was to adapt, at least to some extent. The result was pauperization. Gradually, missionaries and government welfare agents began to have some effect, and questions of humane treatment came to have a more practical meaning. But in outlying areas, maltreatment and violence lingered on into the early 1940s. Further, wherever European settlement was intensive, miscegenation took place, and part-Aborigines came to replace the full-blooded population. Their traditional life ceased to exist as a living reality over much of the southwestern, southeastern, and middle eastern areas of the continent. In the central and northern regions traditional life remained, even on some pastoral, mission, and government stations, although in a modified form. In more remote areas it was still possible for Aborigines to live approximately in the way they had before but with notable modifications, particularly in the sphere of law and order. It was for some time believed that the Aborigines would eventually die out, and reserves were established in the late 1920s and early 1930s to serve as a buffer between them and Europeans. But many were attracted to the fringe settlements, where they formed tribally and linguistically mixed communities. This meant the emergence of a new form of living, structurally linked to the wider Australian society.

Developments since World War II. From the early 1940s, government policies have been changing. But growing support for the goal of assimilation (in effect, Europeanization) did not include adequate programs for achieving it. Many Aborigines were resistant, skeptical, and disillusioned—the aftermath of neglect and the not-too-distant memory of the traumatic events of early contact. In the early 1950s, some sought withdrawal and a magico-religious escape in the *kurangara* (Kimberleys) or in a revival of quasi-Aboriginal religion (as in the north coast of New South Wales). The Pindan Cooperative at Port Hedland in Western Australia is a further example of a movement concerned with maintaining group cohesion on a quasi-traditional basis.

No Aborigines exist who have not had some contact with modern Australian society. By 1960 only about 7,500 Aborigines still kept even a modified traditional orientation. The great majority of the total population of some 45,000 "full bloods" and 78,000 mixed bloods (1971) were living in southern cities and country towns—more European both in physical appearance and in manner of living. Generally, within recent years, the emphasis has been on drawing Aborigines more closely into the wider Australian society and, in the process, erasing unfavourable forms of discrimination. All Aborigines are now Australian citizens, eligible to vote, to receive social service benefits, and to drink liquor. Facilities for primary and secondary education and for technical training have been improved, and more people of Aboriginal descent are taking advantage of this. Nevertheless, many remain poorly trained and educated, caught in the lowest socio-economic level of Australian society. Prospects are brighter for the younger generation, but social acceptance by many white Australians is still limited.

Population figures

Two major developments are significant. A Commonwealth (Federal) conference in 1965 nominally redefined assimilation policies by giving Aborigines a choice between an Aboriginal and an Australian–European orientation. The trend toward social and cultural alteration has actually proceeded so far that, in most areas, such choice is unreal. But it has left the door open in some areas where Aborigines still have an informed appreciation of the value of their traditions. The other development has been an increasing uniformity between state and Commonwealth policies, partly a result of the Federal Office of Aboriginal Affairs, which serves in an overview capacity and provides economic aid.

The last decade has seen the emergence of more articulate part-Aboriginal groups in the south, who insist on integration rather than assimilation—that is, on retaining Aboriginal identity as a unique status symbol marking them off from other Australians. This movement toward pan-Aboriginality has implications for all people of Aboriginal descent. In the north, the focus has been on questions of land ownership and control, including compensation (and not just royalties) for and a share in the mineral exploitation that is occurring on Aboriginal reserves. The Commonwealth has done little about this and in some cases has been aligned with mining bodies that oppose such Aboriginal claims. Generally, there is wide dissatisfaction with the progress of socio-economic development. A likely consequence of this is a reinforcement of the movements mentioned above toward maintaining what remains of the traditional life, reflecting a belief that the last significant aspects of the Aborigines' social and cultural identification should not be dissipated.

BIBLIOGRAPHY. The best general accounts of traditional Aboriginal life are A.P. ELKIN, *The Australian Aborigines* (1964); R.M. and C.H. BERNDT, *The World of the First Australians*, 3rd ed. (1969); F.D. MCCARTHY, *Australia's Aborigines: Their Life and Culture* (1957); R.M. and C.H. BERNDT (eds.), *Aboriginal Man in Australia* (1965); and R.M. BERNDT (ed.), *Australian Aboriginal Anthropology* (1970). H. SHEILS (ed.), *Australian Aboriginal Studies* (1963), also gives a general coverage focussed on research problems. Most of these volumes include sections on change; but see especially M. REAY (ed.), *Aborigines Now* (1964); I.G. SHARP and C.M. TATZ (eds.), *Aborigines in the Economy* (1966); W.E.H. STANNER, *After the Dreaming* (1969); D.E. HUTCHINSON (ed.), *Aboriginal Progress: A New Era?* (1969); and, related to Aboriginal policy and practice, C.D. ROWLEY, *The Destruction of Aboriginal Society* (1970), *Outcasts in White Australia* (1971) and *The Remote Aborigines* (1971). A large series of monographs on various aspects of Aboriginal life has been produced by the Australian Institute of Aboriginal Studies.

Separate and detailed studies of Aboriginal societies are by W.L. WARNER, *A Black Civilization* (1937); C.W.M. HART and A.R. PILLING, *The Tiwi of North Australia* (1960); M.J. MEGGITT, *Desert People: A Study of the Walbiri Aborigines of Central Australia* (1962); L.R. HIATT, *Kinship and Conflict: A Study of an Aboriginal Community in Northern Arnhem* (1965); and R.M. and C.H. BERNDT, *Man, Land and Myth in North Australia: The Gunwinggu People* (1970).

Among the classic studies of the pre-anthropological era are those by R.B. SMYTH, *The Aborigines of Victoria*, 2 vol. (1878); J.D. WOODS (ed.), *The Native Tribes of South Australia, Comprising the Narrinyeri* (1879); E.M. CURR, *The Australian Race*, 4 vol. (1886–87); W.B. SPENCER and F.J. GILLEN, *The Native Tribes of Central Australia* (1899, reprinted 1938); A.W. HOWITT, *The Native Tribes of South-East Australia* (1904); W.B. SPENCER, *The Native Tribes of the Northern Territory of Australia* (1914); B. MALINOWSKI, *The Family Among the Australian Aborigines* (1913); and H. BASEDOW, *The Australian Aboriginal* (1925).

The basis of professional anthropological work was established by A.R. RADCLIFFE-BROWN—see his *Social Organization of Australian Tribes* (1931) and *Structure and Function in Primitive Society* (1952); and A.P. ELKIN, *Studies in Australian Totemism* (1933) and *Aboriginal Men of High Degree* (1946); in the contact sphere, A.P. ELKIN'S article on "Reaction and Interaction: A Food Gathering People and European Settlement in Australia," *Am. Anthrop.*, 53:164–186 (1951), is significant. In regard to studies on Aboriginal women, see P.M. KABERRY, *Aboriginal Woman, Sacred and Profane* (1939); and C.H. BERNDT, *Women's Changing Ceremonies in Northern Australia* (1950); also F. GALE (ed.), *Women's Role in Aboriginal Society* (1971). On Aboriginal art, see F.D. MCCARTHY, *Australian Aboriginal Decorative Art*, 5th ed. (1958) and *Australian Aboriginal Rock Art* (1958); C.P. MOUNTFORD, *The Tiwi: Their Art, Myth, and Ceremony* (1958), and (ed.), *Art, Myth and Symbolism* (1956); and R.M. BERNDT (ed.), *Australian Aboriginal Art* (1964).

Controversies have stimulated discussion on certain aspects of traditional Aboriginal life. One arose from W.L. WARNER'S and A.R. RADCLIFFE-BROWN'S interpretation of the so-called "Murngin" kinship system. See also J.A. BARNES, *Inquest on the Murngin* (1967); and R.M. BERNDT in F.L.K. HSU (ed.), *Kinship and Culture* (1971). The concept of the local group vis-à-vis the horde, also arising from Radcliffe-Brown's earlier contentions, has been discussed by R.M. BERNDT in *Oceania*, 30:81–107 (1959–60); W.E.H. STANNER, *Oceania*, 37:1–26 (1965); and L.R. HIATT, *Oceania*, 32:267–286 (1962), and 37: 81–92 (1966); as well as by J.B. BIRDSELL, with a commentary by students of Aboriginal affairs, in *Cur. Anthrop.*, 11:115–131, 138–142 (1970). The issues of law and order in the absence of clearly defined political authority, the relevance of kinship in this respect, and the place of tribal elders are the subject of conflicting views. R.L. SHARP in *Systems of Political*

Control and Bureaucracy in Human Societies, ed. by V.F. RAY (1958); M.J. MEGGITT, *Indigenous Forms of Government Among the Australian Aborigines* (1962); and L. HIATT, "Social Control in Central Arnhem Land," *South Pacific*, 10: 182–192 (1959), have taken one approach against A.P. ELKIN and R.M. BERNDT in *Aboriginal Man in Australia* (1965); and T.G.H. STREHLOW in *Australian Aboriginal Anthropology* (1970). However, the most long-standing controversy, deriving from Durkheim, has centred on totemism and has recently been revived by C. LEVI-STRAUSS in *Le Totémisme aujourd'hui* (1962; Eng. trans., 1963).

(R.M.B.)

Australian Aboriginal Languages

There are approximately 260 Australian Aboriginal languages. This group of genetically interrelated tongues embraces the entire Australian continent as well as the western islands of Torres Strait, but apparently excludes Tasmania. The languages are characterized by great similarities in their sound systems and considerable agreement in grammar but often by markedly few similarities in vocabulary. Intelligibility between neighbouring forms of speech is common, and dialect chains stretching over amazing distances occur, though the two extremes of such a chain seem to be quite distinct languages.

Secret languages and special vocabularies

Every tribe speaks at least a distinct dialect, but bilingualism and multilingualism are common in many areas. Many individual languages have parallel forms, characterized by special vocabularies and sometimes by special sounds that are used in cultural avoidance situations (*e.g.*, to mothers-in-law) or as secret languages among initiated men on certain occasions.

No genetic link is known to exist between the Australian languages and any outside language. It is believed that languages ancestral to the present-day ones were introduced into Australia by peoples that crossed Arnhem Land in northern Australia many millennia ago. With the apparent exception of the influence of Papuan languages on the languages of the Cape York Peninsula, the Australian languages remained free from outside influence until the arrival of European settlers late in the 18th century.

The great majority of the Australian languages were nearing extinction in the fourth quarter of the 20th century, with about 50 or more extinct, predominantly in the east, south, and west of the continent. Speakers of languages believed extinct for decades are, however, occasionally discovered. Most languages have very few surviving speakers; still-vigorous languages have, for the most part, only a few hundred speakers each, though Mabuiag, the language of the western Torres Strait islands, and the Western Desert language have 7,000 and 4,000 speakers, respectively. About 47,000 Aborigines may still have some knowledge of an Australian language, but accurate figures of the speakers of individual languages are almost impossible to obtain.

Extensive research on the Australian languages has been carried out since 1960, largely through the Australian Institute of Aboriginal Studies in Canberra. The results of this and earlier research have shown the Australian languages to be interrelated and have made it possible to explain their structural differences in terms of a typological development from a simple to a complex structure. In addition, a considerable amount of detailed information on the grammar of numerous languages has been recorded.

Classification. Earlier classifications regarded the northern and northwestern languages, which are structurally rather different from the southern languages, as genetically distinct from the latter. (The interrelationship of the southern languages was recognized quite early.) Later, various classifications based on language type were established, and these demonstrated the ultimate unity of all Australian languages. The most recent classification is based on the degree of lexical (vocabulary) interrelationship between the languages; it subdivides the languages into 28 families, of which 27 are located in the north and northwest, covering about one-eighth of the continent, and a single family is found occupying the remaining seven-eighths of Australia. This skewed picture may be the result of the thorough spreading of a language form referred to as Common Australian (dated at about 5,000–6,000 years ago) from somewhere in northwestern Australia through most of the continent except the north and northwest regions. This diffusion appears to have coincided with that of an archaeologically recognized cultural revolution. The spreading of this Common Australian language may have brought about greater linguistic uniformity in much of Australia. Most of the 28 families are subdividable into groups and often into subgroups of individual languages. In the following list of language families these abbreviations are used: G = group or groups; SG = subgroups; L = language or languages.

Genetic relationship of all Australian languages

Family	
Tiwian (1L)	Djingili-Wambayan (2G, 3L)
Yiwadjan (4G, 2SG, 5L)	Karawan (2G, 2L)
Kakadjuan (1L)	Minkinan (1L)
Mangerian (2G, 2L)	Larakian (2G, 2L)
Gunavidjian (1L)	Kungarakanyan (1L)
Naragan (1L)	Warraian (1L)
Gunwingguan (6G, 3SG, 11L)	Daly (3G, 4SG, 10L)
Bureran (2G, 2L)	Murinbatan (1L)
Nunggubuyuan (1L)	Djamindjungan (4L)
Andilyaugwan (1L)	Djeragan (2G, 5L)
Maran (2G, 2SG, 3L)	Bunaban (2G, 2L)
Mangaraian (1L)	Wororan (3G, 12L)
Ngewinan (1L)	Nyul-Nyulan (4L)
Yanyulan (1L)	Pama-Nyungan (41G, 50SG, 177L)

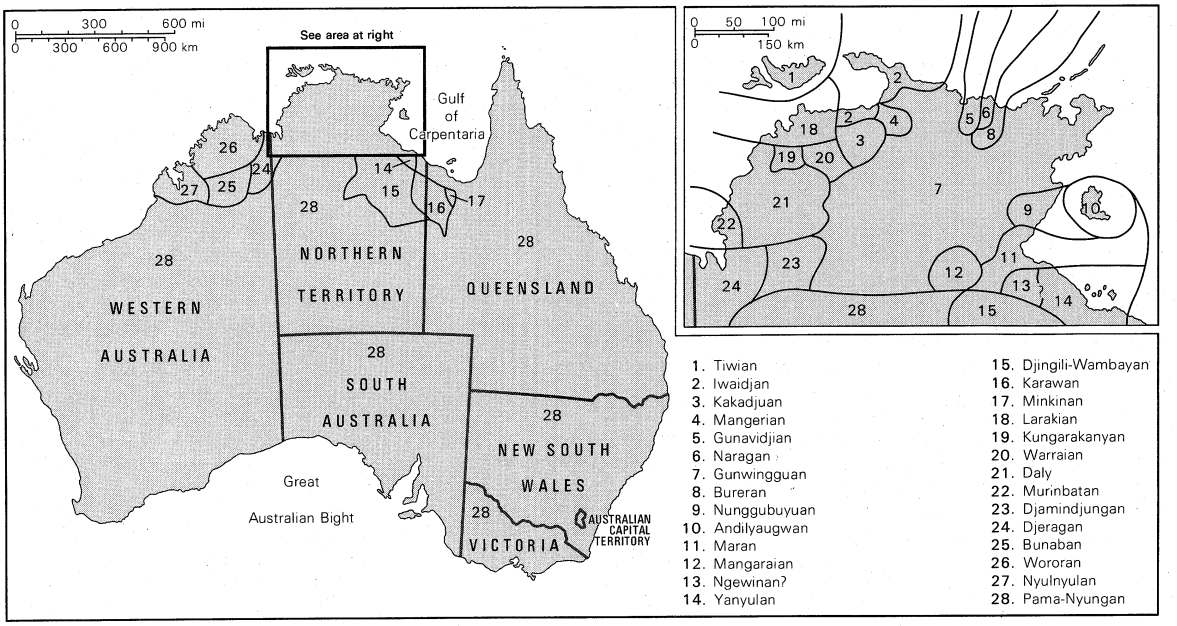
Some aspects of this classification are only tentative. The locations of the families are shown on the accompanying map.

In most areas in which Australian languages are still in daily use, individual languages have gained prominence over others and have become lingua francas (common languages), as a result either of their use as mission languages or of social and cultural factors active among the Aborigines themselves. One of the dialects of the Western Desert language, Bidjandjara (Pitjantjatjara), has become the Aboriginal lingua franca over a sizable portion of the western half of the continent.

Bidjandjara

Grammar. The Australian languages generally show considerable grammatical complexity. Affixes (word parts added initially, internally, or terminally) play an important role; prefixes used initially and suffixes used terminally are found in northern and northwestern Australia, and suffixes, for the most part, are used elsewhere. A peculiar feature of many languages is the suffixing of markers indicating the subject and object of the verb to the first word of the sentence, irrespective of what it is, or to special particles not connected with the verb. An example from Wanman of the Pama-Nyungan family is the use of the suffixes -ŋa and -ŋku in *paralʸi-ŋa-ŋku tʸiŋka-ŋa*, literally, "boomerang-I-you make-past," or "I made a boomerang for you." Another widespread feature is an ergative or agentive suffix, attached to nouns and pronouns, that indicates the actor of an action referred to by a transitive verb. In the Dungidjau language of the Pama-Nyungan family, *tʸa:n-tu pukinʸ-nʸa pumi* is directly translated as "*man*-ergative *dog*-object *hit*-past." Here *-tu* is the ergative suffix; attached to *tʸa:n* ("man"), *-tu* signifies that "man" is the actor of the verb, thus rendering the meaning "the man hit the dog." Great freedom of word order is a feature of many Australian languages. A number of languages, mainly northern ones, have gender and noun class systems, with adjectives, numerals, and demonstratives showing special forms for each of the classes of nouns and often, also, for their number. For example, in Andilyaugwa of the Andilyaugwan family, "where is (located)" is expressed as *ŋi-ŋampa na-mpilʸa* when referring to one man, but as *wunala-ŋampa wu-pilʸa* when referring to two men, *wura-ŋampa na-mpilʸa* in regard to three or more men, *ta-ŋampa iŋa mpilʸa* to one woman, and *ma-ŋampa numa-mpilʸa* to a ship.

Phonology. The sound systems of the Australian languages are extremely similar. Most of them share from four to six different points of articulation for stop consonants (made with complete stoppage of the breath from



Distribution of the Australian Aboriginal languages.
Adapted from S.A. Wurm, *Languages of Australia and Tasmania* (1971); Mouton & Co.

the lungs) and nasal consonants (made with the airstream passing through the nose), with many languages having such consonants produced with the tip of the tongue placed between the teeth (interdental consonants), indicated as *t̪*, *n̪*, or curled up against the hard palate (retroflexed consonants), written as *ṭ*, *ṇ*. The series of consonants may thus include labial consonants (*p*, *m*), interdental consonants, alveolar consonants (*t*, *n*), retroflexed consonants, palatalized consonants (t^y, n^y), and velars (*k*, and ŋ as the *ng* in "sing"). In addition, most Australian languages show no distinction between voiced and voiceless stops (such as voiced *b* and voiceless *p* in English), no fricative consonants (*e.g.*, *f*, *v*, *s*, z), only three vowels (*a*, *i*, *u*), but two or three distinct *r* sounds.

Vocabulary. In spite of the great vocabulary differences among the Australian languages, a number of common words are encountered in a great many languages all over the continent. These are believed to constitute a Common Australian element. In the vocabulary of a given language, various classes of words, such as nouns, verbs, and others, are clearly distinguishable and definable, and word formation is through the use of affixes. Australian languages have contributed to Australian English mainly animal and plant names and objects in nature—kangaroo, wallaby, kookaburra, budgerigar, galah, coolibah tree, billabong.

Common Australian words

Collections of mythological and other text materials have been and still are being made in a number of languages. Native literacy in Australian languages is limited but is on the increase as a result of members of the Summer Institute of Linguistics (an association of Protestant missionaries that specializes in studying primitive languages) who, like the people from the Methodist, Anglican, and Australian Inland missions, use specially adapted versions of English alphabets to write the languages.

BIBLIOGRAPHY. A concise, up-to-date discussion of all aspects of the study of Australian Aboriginal languages, with extensive bibliography, is provided in S.A. WURM, *Languages of Australia and Tasmania* (1971); extensive information may be found in the contributions by A. CAPELL, G.N. O'GRADY, and S.A. WURM in T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 8, *Linguistics in Oceania* (1971); and in G.N. O'GRADY and C.F. and F.M. VOEGELIN, *Languages of the World: Indo-Pacific Fascicle Six*, vol. 8, no. 2, *Anthropological Linguistics* (1966). Numerous articles and monographs on Australian languages appear in the serials: "Australian Aboriginal Studies," Canberra; "Pacific Linguistics," Canberra; "Oceania," Sydney; "Oceania Linguistic Monographs," Sydney.

(S.A.W.)

Australian External Territories

Apart from claims in Antarctica, the External Territories of the Commonwealth of Australia are made up entirely of islands, spanning almost one-third of the Southern Hemisphere and contending with the greatest possible range of climate and physical environment. They comprise the Australian Territory of Papua, occupying the southeastern section of the tropical island of New Guinea, immediately to Australia's north, with related archipelagos; innumerable small reefs, cays, and atolls between Queensland's Great Barrier Reef and longitude 157°10′ E; and several remote and diverse islands in the Pacific and Indian Oceans.

These latter oceanic outposts, far beyond the boundaries of the shelves that fringe the continental land masses, occupy the tips of submerged mountain ranges, many of volcanic origin. Those in the tropics often support fringing coral reefs, or atolls, where the sea level has risen sufficiently relative to the basic rock. With the exception of the mainland and islands of the Territory of Papua, none of the Australian External Territories was inhabited when they were first annexed by Great Britain, and subsequent development and exploitation has been carried out by migrant and recruited labour. The rugged terrain of New Guinea—the highest peaks in the Territory of Papua are frequently over 10,000 feet (3,500 metres)—was occupied by indigenous peoples long before any European contact. Dusky, woolly-haired Melanesians of the coastal and island regions and Papuans, the highly variable greater part of the inland population, with some Negrito traits in the central uplands, form the basic demographic elements. Polynesian and Micronesian incursions, from the Pacific generally, have further diversified these groups; and there is an almost inconceivably wide range of physical characteristics and of languages, with a rich, colourful, and varied accompaniment of ornaments, weapons, and other artifacts. The Australian Territory of Papua is bounded on the north by a trust territory (see NEW GUINEA, TRUST TERRITORY OF), which is currently administered by Australia under a United Nations Trusteeship Agreement, through the Papua and New Guinea Act (1949–68) of the Commonwealth of Australia. The Territory of Papua is bounded on the west by the Indonesian province of West Irian. The western boundary, settled by an agreement between Britain and Holland in 1895, lies along the 141st east meridian of longitude, except in the southern portion, where a small bulge to the west incorporates the most westerly point of the Fly River. South of this incursion, the boundary runs along the

line of 141°47′54″ E. The northern boundary roughly follows the complex mountain divide running across New Guinea (see NEW GUINEA, MOUNTAIN RANGES OF). The 86,100 square miles (223,000 square kilometres) of the Papuan territory contained, by the early 1970s, a population of approximately 670,000, most of whom still made their living by a combination of subsistence farming, hunting, or fishing.

Territorial composition

In detail, the composition of the remainder of the Australian External Territories is as follows: Norfolk Island, in latitude 29°02′ S, longitude 167°57′ E, and with an area of about 13.5 square miles (34 square kilometres), is a fertile and beautiful island famous for its indigenous pine. Associated with the earliest European settlement of Australia, its major industry today is tourism. Australia's newest territory (created by Act of the Commonwealth on Sept. 30, 1969) is also found in the Pacific Ocean. The Coral Sea Islands Territory, scattered over 400,000 square miles of tropical waters, is uninhabited except by the observers of the Commonwealth meteorological station on Willis Islets, latitude 16° S, longitude 150° E (established 1921).

The Cocos (Keeling) Islands, consisting of two groups of 27 islands with a total area of 5.5 square miles (14.2 square kilometres) are located in the Indian Ocean in latitude 12°5′ S and longitude 96°53′ E. By their transfer from Britain to the Commonwealth of Australia, under the Cocos (Keeling) Islands Act (1955), they ceased to be part of the Republic of Singapore. Christmas Island, 52 square miles (135 square kilometres) in area, in latitude 10°25′ S and longitude 105°40′ E, is a source of high-grade rock phosphate. Like the Cocos (Keeling) Islands, it was also transferred to Australia from Britain, under the Christmas Island Act (1958), having been previously governed and administered as part of the Republic of Singapore. The claim comprising Australian Antarctic Territory consists of all islands and territories, other than Adélie Land, situated south of the 60th parallel of south latitude and between the 45th and 160th eastern meridians. This territory, over which the British government formerly asserted the crown's sovereign rights, was transferred to Australia under the Australian Antarctic Territory Acceptance Act (1933; see ANTARCTICA). Heard Island, latitude 53° S, longitude 73°23′ E, and approximately 215 square miles (557 square kilometres) in area, is perpetually ice-capped; the McDonald Islands Act of the Commonwealth (1953–63) defines Australia's claim.

The northern districts of Papua New Guinea are bound by the Equator; Australian Antarctic Territory reaches to the South Pole. Heard Island, farthest west of all Australian territories except the Enderby Land section of Antarctica, is 2,500 miles southwest of Fremantle, Western Australia; Norfolk Island, most easterly, is located 1,035 miles east-northeast of Sydney, New South Wales. To contend with this diversity, radio networks provide channels for administrative and commercial interests and for an extensive meteorological service. Physical contact between the Commonwealth and its territories varies from daily jet air services between the Australian capitals and Papua New Guinea to the annual relief of its Antarctic stations by vessels strengthened for ice.

Administrative responsibilities

The Commonwealth Department of External Territories provides administrative services for the territories of Papua and New Guinea, Norfolk Island, the Cocos (Keeling) Islands, Christmas Island, and the Coral Sea Islands, except in matters of defense and civil aviation, for which separate ministries exist. For the inhabited territories, an administrator or official representative is appointed by the governor general of Australia to direct the government of the territory on behalf of the department. The department is concerned with matters relating to the United Nations Trusteeship Agreement for New Guinea and for other international aspects of Australia's territories. It recruits personnel for the territorial public services and, through the Australian School of Pacific Administration, in Sydney, it trains staff appointed to or serving in the various administrations. The department is pledged by the Commonwealth to "advance the people of Papua and New Guinea to self-determination." With increasing numbers of local government councils and a majority of indigenes in the House of Assembly and in the public service, considerable progress has been made toward political awareness and self-government. During the course of the parliament of 1972–76, internal self-government has been recommended by the UN Visiting Mission of 1971. Full political independence by 1980 has been forecast.

Although the former Trust Territory of Nauru gained independence as a republic on Jan. 31, 1968, the British Phosphate Commissioners, who manage and supervise the mining, shipment, and sale of phosphates from the island deposits, are responsible to the Minister for Australia's External Territories. The Commonwealth Department of Supply administers Australia's Antarctic interests and the enactments defined by the Antarctic Treaty Act (1960), the Australian Antarctic Territory Acceptance Act (1933), and the acts of 1953–63 relating to Australian Antarctic Territory and to Heard and the McDonald Islands. The Department of Supply, through its Antarctic Division, is responsible for the logistics of the annual relief operations of Australia's southern scientific stations (for which such departments as that of the Interior and of National Development supply certain scientific and technical personnel) and for their continuous general function and administration.

THE AUSTRALIAN TERRITORY OF PAPUA

Papua comprises the southeastern mainland of New Guinea with the islands of the Gulf of Papua and, east of this territory, to longitude 154°14′ E, between latitudes 8° S and 12° S, a considerable archipelago including the coral Trobriands, the Woodlark, Laughlan, D'Entrecasteaux, and Conflict groups, with the Louisiade Archipelago, and Samarai.

The northern boundary of Papua spans the New Guinea highlands to latitude 8° S, beyond which, for 300 miles, the territory occupies the whole southeastern cape centred by the Owen Stanley Range (highest peak Mount Victoria, 13,363 feet [4,073 metres]). Severe earthquakes, occasional volcanic activity, and continuing coastal uplift or subsidence testify to the geological instability of the region. The counterparts of the jagged and abrupt mountains are deep and precipitous gorges with swift and turbulent streams, which, apart from the Fly River in the west, are unnavigable for any great distance by cargo vessels. The Fly, an important river, is navigable for 500 miles by vessels of eight-foot draft. West of the Fly Delta is the low and stable Oriomo Plateau, formed of limestone and alluvial deposits.

Climate of Papua

Because Papua lies wholly within the tropics, there is little seasonal change of atmospheric temperature or humidity, both of which are uniformly high throughout the year. The mean maxima and minima of temperature rarely fall outside the range 70°–90° F (21°–32° C), the daily variation being 10°–15° F (6°–8° C), while the range of relative humidity is approximately 65 to 85 percent. A general lowering of temperatures occurs in the highlands, but the permanent snow line is not reached in Australian territory.

In all parts of New Guinea rain may fall throughout the year, for the northwest monsoon is active from December to March and the moisture-laden southeast trade winds blow from May to October, with brief, calmer, but unsettled, weeks during the transition periods. Though most places receive more than 100 inches, the disposition of mountains and coast causes remarkable variations in rainfall. The monsoon rains regularly exceed those of the trades except in the islands and at those places where rising ground forms a direct barrier to the winter winds from the southeast.

Such an area surrounds Kikori, on the Gulf of Papua, receiving 200–250 inches (510–635 centimetres) of rain annually. By contrast, Port Moresby, the administrative headquarters of Papua New Guinea, located only 250 miles away, receives little rain from the trades because the Owen Stanley Range runs parallel with their course. As a result of the rain-shadow effect of the central cordil-


AUSTRALIAN EXTERNAL TERRITORIES
Size of symbol indicates relative size of town • ◉ ■
Elevations in metres
© Rand McNally & Co.
A-599600-257
400 800 1200 1600 km
0 400 800 1200 mi
INDONESIA
AUSTRALIA
PAPUA NEW GUINEA
NEW ZEALAND
INDIAN OCEAN
PACIFIC OCEAN
TASMAN SEA
GREAT SANDY DESERT
GIBSON DESERT
GREAT VICTORIA DESERT
GREAT AUSTRALIAN BIGHT
TANAMI DESERT
SIMPSON DESERT
MACDONNELL RANGES
GREAT DIVIDING RANGE
Tropic of Capricorn
Equator
International Date Line
Perth
Adelaide
Melbourne
Sydney
Canberra
Brisbane
Darwin
Alice Springs
Hobart
Wellington
Auckland
Christchurch
Port Moresby
Noumea
Suva
Honiara
Port Vila
Apia
Nuku'alofa
Rarotonga
Papeete
Kingston
Avarua
Alofi
Fakaofo
Funafuti
Tarawa
Majuro
Kwajalein
Yaren
Palikir
Koror
Agana
Saipan
Hagatna
Pago Pago
Jarvis Island (U.S.)
Canton and Enderbury Islands (U.S.-U.K.)
Phoenix Islands (Claimed by U.S.)
Gilbert Islands (U.K.)
Nauru
Ocean Island
Tuvalu (Ellice Islands) (U.K.)
Solomon Islands (U.K.)
New Hebrides (Fr.-U.K.)
New Caledonia (Fr.)
Loyalty Islands
Wallis and Futuna (Fr.)
Western Samoa
American Samoa
Niue (N.Z.)
Cook Islands (N.Z.)
Tonga
Fiji
Norfolk Island (Austl.)
Lord Howe Island (Austl.)
Kermadec Islands (N.Z.)
Chatham Island (N.Z.)
Bounty Islands (N.Z.)
Antipodes Islands (N.Z.)
Auckland Islands
Campbell Island (N.Z.)
Macquarie Island (Austl.)
Heard and McDonald Islands (Austl.)
Ashmore and Cartier Islands (Austl.)
Cocos Islands (Austl.)
Christmas Island (Austl.)
Coral Sea Islands (Austl.)
Ile Nouvelle Amsterdam (F.S.A.T.)
Ile Saint-Paul (F.S.A.T.)
Iles Kerguelen (F.S.A.T.)
Mount Kosciusko 7316
Mount Cook 12349
Mount Victoria 13363
Mount Wilhelm 14793
Mount Ossa 5617
Ruapehu 9797
Round Mountain 1394
Mount Woodroffe 1440
Lake Eyre -16
Great Barrier Reef
Gulf of Carpentaria
Gulf of Papua
Torres Strait
Bass Strait
Cook Strait
Bay of Plenty
Shark Bay
Coral Sea
Solomon Sea
Arafura Sea
Timor Sea
Java Sea
Flores Sea
Banda Sea
Bismarck Sea
Joseph Bonaparte Gulf
Collier Bay
Great Australian Bight
Spencer Gulf
Darling Range
Flinders Ranges
Selwyn Range
Barkly Tableland
Kimberley Plateau
Darling
Murray
Lachlan
Thomson
Fraser Island
Kangaroo Island
Cape York Peninsula
Cape Leeuwin
Cape Naturaliste
Cape Nelson
Cape Everard
Cape Grim
Cape Farewell
North Cape
East Cape
West Cape
The Snares
Stewart Island
Flinders Island
Louisiade Archipelago
Santa Cruz Islands
Torres Islands
Banks Islands
Malekula
Vanua Levu
Viti Levu
Savaii
Upolu
Manua Is.
Rose Island
Suvorov Islands
Palmerston
Manahiki
Tokelau Islands (N.Z.)
Chesterfield Is.
Malaita
Guadalcanal
Choiseul
Santa Isabel
New Britain
New Ireland
Admiralty Islands
Manus Island
Bismarck Archipelago
Kepulauan Sula
Kepulauan Aru
Kepulauan Tanimbar
Kepulauan Batu
Pulau Siberut
Pulau Enggano
Billiton
Pulau Japen
Biak
Dampier Strait
Molucca Sea
Ceram Sea
Celebes Sea
Sulu Sea
Makassar Strait
Lombok Strait
Sunda Strait
Karimata Strait
Malacca Strait
Pulau Roti
Lesser Sunda Islands
Sumba
Sumbawa
Flores
Timor
Port Timor
Melville Island
Cape Arnhem
Cape Fourcroy
Cape Leveque
North West Cape
Sandy Cape
Round Mountain 1394
Mount Victoria 4073
Mount Wilhelm 4509
Puntjak Djaja 5030
Gunung Kerinci 3800
Gunung Semeru 3676
Gunung Tambora 2851
Gunung Rinjani 3726
Gunung Agung 3142
Gunung Merapi 2911
Gunung Slamet 3428
Gunung Ciremai 3078
Gunung Kelimutu 1690
Gunung Sibayak 2094
Gunung Leuser 3404
Gunung Kinabalu 4101
Gunung Bawakaraeng 2833
Gunung Latimojong 3478
Gunung Rantemario 3455
Gunung Mekongga 2790
Gunung Tambusisi 2421
Gunung Ambang 1795
Gunung Lokon 1580
Gunung Soputan 1784
Gunung Ruang 1725
Gunung Api 656
Gunung Banda Api 656
Gunung Gamalama 1715
Gunung Ibu 1325
Gunung Dukono 1229
Gunung Sibela 1730
Gunung Tidore 1730
Gunung Ternate 1721
Gunung Gamkonora 1635
Gunung Binaiya 3027
Gunung Karimui 2590
Gunung Wilhelm 4509
Gunung Hagen 3359
Gunung Giluwe 3992
Gunung Bosavi 2507
Gunung Victoria 4073
Gunung Obree 3720
Gunung Yelia 3380
Gunung Albert Edward 4035
Gunung Trikora 4730
Gunung Mandala 4702
Gunung Idenburg 4717
Gunung Doorman 4843
Gunung Carstensz 4884
Gunung Wilhelmina 4540
Gunung Sentani 2505
Gunung Kumawa 2937
Gunung Kobowre 2750
Gunung Nok 2565
Gunung Arfak 2955
Gunung Tamrau 2438
Gunung Fakfak 2950
Gunung Kowiai 2340
Gunung Kasi 2360
Gunung Kumawa 2937
Gunung Kobowre 2750
Gunung Nok 2565
Gunung Arfak 2955
Gunung Tamrau 2438
Gunung Fakfak 2950
Gunung Kowiai 2340
Gunung Kasi 2360

lera, it also receives a minimum from the monsoon; the annual aggregate is about 40 inches (100 centimetres).

Although the greater part of Papua is clothed in dense tropical rain forest, varying in type with altitude, most of the inland soils are of low fertility, thin, and washed out by continual rain. Cleared land rapidly becomes unproductive, except for small areas in the highland valleys and in regions where there has been an accumulation of alluvial or volcanic soils. Swamp forest and rank grass clothe much of the valley of the lower Fly River; parts of the coast and many of the islands are fringed by mangroves and palms—coconut, nipa, and sago palms—and, in lower rainfall areas, by savanna grasses and scrubs.

The people and their life

The indigenous peoples of Papua, Papuans of the west and of the interior and Melanesians of the islands and eastern mainland, numbering (1971 estimate) 671,000, in general exist in small, egalitarian communities of 300 or less, building huts of local materials—timber, grass, or palm fronds—and cultivating native gardens, frequently moved as soil fertility decreases, growing yams, taro, and sweet potatoes. Sago palms and, in some parts, bananas provide a staple in the diet. Hunting supplements such supplies, domestic pigs provide prestige as well as seasonal feasts, and, in coastal areas and along the rivers, fishing is important. Indigenous food plants, other than those already mentioned, include sugarcane, ginger, and several kinds of nuts.

There is great diversity of language, mostly mutually unintelligible. On the coast, Motu and pidgin English are widely understood. In the more than 850 schools, administration and mission (of many different Christian denominations), English is taught; there are 17 primary schools where the curriculum is Australian-oriented. By the start of the 1970s, about half of the children of school age were receiving some formal education. There were also limited facilities for secondary education. In 1965 the University of Papua and New Guinea was founded at Boroka, near Port Moresby, and within a year had faculties of arts, law, and science, with an enrollment of about 600 indigenous students.

The average Papuan man or woman is about five feet tall and weighs between 100 and 130 pounds. Because of the ravages of malaria, dysentery, and hookworm and other parasites, the proportion living beyond the age of 50 is still small. There are significant exceptions from the general pattern, and modern medicine is slowly but successfully combating disease and malnutrition; the infant mortality rate has been lowered consistently. In all districts of Papua, hospitals and medical services are provided both by the missions and the administration.

Everywhere in Papua, outside the main coastal town of Port Moresby and the few centres such as Popondetta with extended road communications, interest is still basically confined to the local community and its affairs. The aggregate length of vehicular roads had, by 1970, exceeded 690 miles, with a further 1,560 miles of roads and tracks usually allowing access. No region in Papua is more than 50 miles from an aircraft landing field, of which there are about 100 in the various districts.

Over the greater part of the country, however, magico-religious practices, taboos, and indulgences, based on the worship of ghosts and ancestral spirits, continue in some form, not infrequently blended with Christian derivatives. It is significant that in the most recent census more than nine-tenths of the population claimed membership in one of a dozen Christian religious denominations. From the beginnings of European administration, head-hunting, cannibalism, and vendetta have declined and in Papua today perhaps scarcely exist. The village sorcerer is still influential in many parts where the affairs of crops, hunting, and society are believed affected by social or antisocial magic. Otherwise, power is held by the men of the community in consultation, with leadership, single or multiple, generally based on skills of various kinds, such as weapon making and gardening, and on personality and possessions.

The Papuan animal life is largely related to that of Australia, marsupials being common and the largest of these being the tree kangaroo. The spiny anteater is present, as are numerous bats and rodents. Reptiles include numerous snakes, many venomous, and both estuarine and marine crocodiles and species of both turtles and tortoises. Birds are colourful, with both Australian and Malaysian affinities. The gorgeous bird of paradise and the large, flightless cassowary are distinctive of the region. There are myriad insects, spiders, scorpions, centipedes, and mites, including the typhus and malarial hosts against which precautions and vigilance are required.

Since the affairs of Papua first concerned Australia in 1906, the district organization has spread, gradually, through the disparate communal and geographical structure. This is made up of thousands of individual villages with virtually no common interests, hundreds of distinct languages, and a dozen markedly different environments. The district system has ultimately provided a system of communications; a means of dispersing knowledge and policy; and legal, medical, educational, and agricultural services. This made possible the beginnings of native government and, eventually, the formation of local government councils. The latter, established in 1950, had, by the 1970s, increased to more than 50 in Papua (with a further 90 in the Trust Territory of New Guinea), representing about 85 percent of the population.

Administration of Papua New Guinea

The Department of District Administration of the Public Service of Papua New Guinea provides for six Papuan districts, from west to east, Western, Southern Highlands, Gulf, Central, Northern, and Milne Bay—all except the Southern Highlands with access to the sea—and 12 in the trust territory. Part of the district of Chimbu (mainly in the trust territory) enters Papua east of the Southern Highlands. Each district is administered by a district commissioner, a deputy district commissioner, several assistant district officers and patrol officers, and medical, agricultural, and educational officers, resident in the chief local centres or the subdistricts. The Public Service in general, with 16 departments, having headquarters in Port Moresby, is responsible for implementing government policy for the political, social, and economic development of Papua and the Trust Territory of New Guinea. A majority of its employees, assisting in all departments, are natives. Through their local government councils, the indigenous peoples are becoming politically aware and are taking more part in community management; working with the district administration; and assisting with the construction of roads, bridges, aircraft facilities, meeting halls, schools, and other public buildings.

Representation of the indigenous population of Papua New Guinea on a national basis commenced with the provision, in 1951, for three representatives to be elected to the Legislative Council. The number was increased to 12 elected members in 1960. In 1964 the Papua and New Guinea House of Assembly replaced the Legislative Council, the present constitution (1968) allowing for 84 elected members and 10 official members appointed by the governor general of Australia, on the nomination of the Administrator. The 1972–76 (third) House of Assembly contained 4 official members. The body, whose chief minister, Michael Somare, was New Guinean, has power to make ordinances for the peace and good government of the territory and is advisory to the Administrator's Executive Council, which also includes a majority of elected members nominated by the House of Assembly (with the concurrence of the administrator), who must assent to all ordinances before they take effect. The Australian Commonwealth's grant to revenues at present equals that raised internally. The judiciary of Papua New Guinea consists of the Supreme Court, with a chief justice and other judges, and district, local, and children's courts under the jurisdiction of stipendary and resident magistrates, an increasing number of whom are indigenes.

Virtually all of the commercial agricultural production of Papua, as distinct from the subsistence crops of villagers, is obtained from plantations, mainly near the coast or on the islands and occupying less than 1 percent of the land. Slightly more forest land is being exploited for timber. Copra is produced from all districts except

the Southern Highlands, which, with the mountainous areas of the Central District, grow high-quality coffee. Other than copra, rubber is the principal crop, and some cacao plantings are promising. The Southern Highlands support small crops of pyrethrum (a plant whose pungent root is used in medicine); a growing number of native-owned cattle are also pastured in high grasslands.

Economic development

About 16,600 tons of copra were exported annually by the 1970s, together with 5,756 tons of rubber; other exports included sawed timber, cocoa beans, fish and crustaceans, and shells. The value of present imports is about five times that of export goods. Mineral production in Papua, by contrast with that of the Trust Territory of New Guinea, has been trivial, but the considerable quantities of copper ore reported to exist in the inaccessible Star Mountains in the far northwest may be exploited.

Hydroelectric power stations are gradually being established. Port Moresby, Popondetta, Daru, Mendi, and Samarai possess hydroelectric or thermoelectric stations with a total installed capacity (1971) of 35.9 megawatts.

NORFOLK ISLAND

Norfolk Island, with a total area of about 13.3 square miles (34.5 square kilometres), is volcanic in origin and has a mean altitude of 350 feet; two high points, Mount Pitt and Mount Bates, just exceed 1,000 feet. The climate is pleasant and equable, with an annual rainfall of 52 inches (132 centimetres) and a mean annual temperature of 61° F (16° C). Although most of the island has been cleared for cropping and pasture, the once-dominant Norfolk Island pines are still a notable feature of the landscape. The beautiful island is famed for the ruins dating from its cruel convict era, and for the "islanders" who are descendants of the mutineers of H.M.S. "Bounty." The latter, in 1789, cast off the captain of the vessel, Lieut. William Bligh, in a provisioned boat, which ultimately reached Timor; while they themselves, with some Tahitian women, first settled on lonely Pitcairn Island. After many vicissitudes, the majority of their descendants were resettled on Norfolk Island in 1856, taking over the evacuated convict establishment. Among the contemporary islanders, there are many residual traits and customs, including some of Tahitian origin.

Descendants of the mutineers

The 20th-century population (including immigrants from the mainland) first exceeded 1,200 as recently as 1970; but tourists, numbering about 10,000 annually, have created a great change in the island's economy. There is considerable competition for the visitors' custom and at times a local scarcity, formerly unknown, of island produce. Norfolk Island is served by regular air services from Sydney and by less frequent shipping from Sydney, Noumea, and New Zealand ports.

An administrator, appointed by the Australian governor general, is at the head of a small civil service in Kingston, the main settlement. He is also chairman of the Norfolk Island Council, an advisory body consisting of eight elected local representatives. The legal system includes a Supreme Court with visiting mainland judges, and a Court of Petty Sessions under the jurisdiction of local magistrates.

Tourism, some fresh-fruit produce, and the sale of bean, palm, lemon, pine, and flower seeds, with some hides, underlie the economy; while the important items of revenue are the philatelic sale of postage stamps, customs duty, and the sale of liquor, which is a government monopoly. Australian currency is legal tender.

CORAL SEA ISLANDS

Australia's Coral Sea Islands Territory includes the Flinders Reefs, Herald Cays, Holmes Reefs, Moore Reefs, Bougainville Reef, Ospray Reef, Willis Islets, and others in the northwest of the Coral Sea, and a southeasterly group embracing Frederick Reef, Saumarez Reef, and Cato Island; all except Willis Islets are uninhabited, but they are occasionally visited by scientists, fishermen, and prospectors for oil or minerals.

British naval vessels—H.M.S. "Cato" (1803), H.M.S. "Frederick" (1812), and H.M.S. "Herald" (1854–60)—discovered and surveyed these low coral islands, all of which are either topping volcanic foundations or the dissected remnants of submarine plateaus.

The automatic weather station (1966) on Cato Island supplements information from the manned Australian Bureau of Meteorology station (1921) on Willis Islets.

Since 1967 extensive surveying of the numerous and variable cays, and of the low reefs and islands has been undertaken by the Australian Department of National Development.

CHRISTMAS ISLAND

At several points exceeding 1,100 feet in altitude, Christmas Island, 52 square miles (135 square kilometres) in area, is the summit of an oceanic mountain surrounded by depths of 6,000 feet. The abyssal Java Trench (24,500 feet at its deepest) is to the northeast. The rock is mainly limestone stratified with old volcanic flows, with valuable deposits of phosphate of lime near Flying Fish Cove and on the Southern Plateau. The island is 815 miles from Singapore, only 224 miles from Java Head, and 1,630 miles from Fremantle, Western Australia. Its climate is salubrious and warm (70°–90° F [21°–32° C]), with little seasonal variation but a tropical rainy season between November and April. Rich forest clothes the island, and both flora and fauna are varied and numerous. The latter includes birds, small, harmless reptiles, crustaceans, insects, spiders, scorpions, and centipedes.

The island was first sighted by Capt. William Mynors, of the British East India Company, on Christmas Day in 1643. It was annexed by Great Britain in 1888. A lease granted to the local families of George Clunies-Ross and Sir John Murray was transferred six years later to the Christmas Island Phosphate Company. These interests were acquired by the Australian and New Zealand governments (1948). In 1958, when the island became an Australian territory, a new agreement ratified the existing legislature; the British Phosphate Commission, on behalf of both governments, quarried and shipped the valuable deposits. The existence of some 200,000,000 tons of phosphate of various grades has been confirmed; present production exceeds 1,000,000 tons annually.

Phosphate deposits on Christmas Island

The population is 3,600, consisting mainly of recruited Chinese and Malay labourers from Malaysia, Singapore, and the Cocos (Keeling) Islands, working for the commission, with Australian or British management personnel, and the island administration.

An administrator of Christmas Island is appointed by the Australian governor general, with a staff for secretariat, education, police, radio, and harbour duties. The judiciary consists of a Supreme Court, District Court, and Magistrate's Court. The functions of the latter, empowered to try minor offences, are exercised by resident magistrates, usually both Chinese and European. Revenue is gained from royalties paid by the British Phosphate Commission, with small contributions from customs duty, duty on liquor, and from the philatelic sale of postage stamps.

Although, since settlement and exploitation of the island's resources, some of the unique original animal life of Christmas Island has become extinct, the mass movement of millions of red land crabs, which annually deposit their eggs along the coast and return to the inland jungles, still provides a remarkable phenomenon.

COCOS (KEELING) ISLANDS

The Cocos (Keeling) Islands, 2,290 miles almost due west of Darwin, comprise 27 small coral exposures in two atolls, the northern and much smaller North Keeling Island lying about 15 miles from the larger circular group of West Island, South Island, Home Island, and numerous smaller isles. The climate is governed largely by the southeast trade winds and is warm, with temperatures consistently between 70° and 89° F (21° and 32° C), average relative humidity 75 percent, and rainfall about 70 inches. The population (1971 estimate) was 671: 143 Europeans and most of the rest descendants of the original Muslims of Malayan origin brought to the islands in the early 19th century. Between 1948 and 1951, an extensive emigration movement, aided by the management of

the Clunies-Ross Estate and the government of Singapore, reduced the islands' population by more than 1,600 Cocos Islanders, who were mainly settled on estates in Sabah, Malaysia (then North Borneo).

An official representative of the Australian government is resident, and various instrumentalities of the Commonwealth provide essential services in aviation, communications, meteorology, and in legal and fiscal matters. The indenture of 1886, granting all land in the territory to the Clunies-Ross family, subject to the crown's resumption of areas necessary for public purposes, was still valid in the early 1970s.

HEARD AND THE MCDONALD ISLANDS

Antarctic research on Heard Island

Heard Island was occupied continuously from 1947 to 1954 by the Australian National Antarctic Research Expeditions (ANARE), the scientific program then being transferred to stations on the Antarctic mainland, at Mawson, and then Davis. The deserted station on Heard is visited occasionally by relief expeditions bound for the more southern bases and has been used by U.S. scientists and technicians concerned with satellite tracking and other space programs. The island was first sighted in 1833, and many teams exploited the island's elephant seals and penguins for their oil during the second half of the 19th century.

The summit of the volcanic Big Ben (9,000 feet [2,700 metres]) was reached by an Australian expedition led by W. Deacock in 1964–65.

The McDonald Islands have not been occupied. Like the lower parts of Heard Island, they are sparsely vegetated with cushion plants, rough tussock grass, mosses, and lichens; and they are visited by large numbers of elephant seals, leopard seals, and penguins, chiefly the gentoo, the rock hopper, and the macaroni. Several petrels, albatrosses, and skuas that breed on Heard Island probably also inhabit the McDonald Islands.

FUTURE STATUS OF THE ISLANDS

Until virtually the end of the 20th century, under the terms of the international Antarctic Treaty, territorial claims and ambitions in the Antarctic are effectively dormant, and the concept of a United Nations or other international control of the whole continent in perpetuity shapes itself.

The other Australian territory that, it would seem, must eventually change its status is the Territory of Papua, as part of the jointly administered Papua New Guinea, the two entities being seldom differentiated in terms of their future by Australian, if not by Afro-Asian, opinion.

In recent years there has been increasing pressure from the United Nations General Assembly's trusteeship committee to speed the processes leading to the self-determination of the Trust Territory of New Guinea.

Within the Territories of Papua and New Guinea and in the local government councils themselves, there is an increasing native realization that the course toward nationhood connotes progress in economics, industry, and communications inimical to an age-old group cultural and social context. Paradoxically, constitutional development committees, led by indigenes, have reported widely held opposition to many of the inevitable changes contingent on the transition from tribal decentralization to the national unity, and there have been proposals of a necessary interim of self-government before complete independence is achieved.

BIBLIOGRAPHY

General: COMMONWEALTH BUREAU OF CENSUS AND STATISTICS, *Official Year Book of the Commonwealth of Australia* (annual); DEPT. OF TERRITORIES, COMMONWEALTH OF AUSTRALIA, *External Territories* (1970); A.H. CHISHOLM (ed.), *The Australian Encyclopaedia* (1965); R. ROSE, *Australia's Island Territories* (1967); C.R.H. TAYLOR, *A Pacific Bibliography: Printed Matter Relating to the Native Peoples of Polynesia, Melanesia, and Micronesia*, 2nd ed. (1965). See also the annual reports of the Australian Department of Territories.

Australian Territory of Papua (geographical and general works): J. ANDREWS *et al.*, *New Guinea and Australia* (1958); J. ANDREWS, *New Guinea* (1957); F. CLUNE, *Prowling through Papua* (1942); B. ESSAI, *Papua and New Guinea* (1961); J.N. JENNINGS and J.A. MABBUTT (eds.), *Landform Studies from Australia and New Guinea* (1967); R. JOYCE, *New Guinea: A Brief Account of Travels, Mainly in British New Guinea and Papua* (1960); D.A.M. LEA and P.G. IRWIN, *New Guinea: The Territory and its People* (1967); J.K. MCCARTHY, *Patrol into Yesterday: My New Guinea Years* (1963); A. MALLARD, *A Traveller's Guide to Papua-New Guinea* (1969); A.G. PRICE, *The Challenge of New Guinea: Australian Aid to Papuan Progress* (1966); O. RUHEN, *Mountains in the Clouds* (1963); C. SIMPSON, *Plumes and Arrows: Inside New Guinea* (1962); G. SOUTER, *New Guinea: The Last Unknown* (1963); G.W.L. TOWNSEND, *District Officer from Untamed New Guinea to Lake Success* (1968); R.G. WARD and D.A.M. LEA (eds.), *An Atlas of Papua and New Guinea* (1970); O.E. WHITE, *Parliament of a Thousand Tribes* (1965); M. WILLIAMS, *Stone Age Island: Seven Years in New Guinea* (1964). Current information may be found in the *Handbook of Papua and New Guinea* (1969); and in the *Current Affairs Bulletins* (1952–) and in *New Guinea Research Bulletins* (Aust. Nat. University, Canberra).

Historical and political works: D.G. BETTISON *et al.*, *The Independence of Papua-New Guinea: What are the Pre-requisites?* (1962); T. BEVAN, *Toil, Travel, and Discovery in British New Guinea* (1890); D.C. GORDON, *The Australian Frontier in New Guinea, 1870–1885* (1951); P. HASLUCK, *Australian Policy in Papua and New Guinea* (1956); L. LETT, *Sir Hubert Murray of Papua* (1950); H. NELSON, *Papua New Guinea: Black Unity or Black Chaos?* (1972); D. STEPHEN, *A History of Political Parties in Papua New Guinea* (1972); R.W. ROBSON, *Queen Emma: The Samoan-American Girl Who Founded an Empire in Nineteenth Century New Guinea* (1965).

Anthropological works: M. MEAD, *Growing Up in New Guinea: A Comparative Study of Primitive Education* (1930); C. SELIGMANN, *Melanesians of British New Guinea* (1910); C. SIMPSON, *Adam with Arrows* (1953).

Norfolk Island: P.J. MARKS, *Norfolk Island and the Bounty Mutiny* (1935); J.J. SPRUSON, *Norfolk Island: Outline of its History from 1788 to 1884* (1885); U. WHITE and R. SRIBER, *Norfolk Island Sketchbook* (1968).

Christmas Island: C.W. ANDREWS, *A Monograph of Christmas Island* (1900); J.S. HUGHES, *Kings of the Cocos* (1950).

Cocos (Keeling) Islands: CHARLES DARWIN, *The Structure and Distribution of Coral Reefs* (1842); J.S. HUGHES, *Kings of the Cocos* (1950).

Heard Island and the McDonald Islands: P.G. LAW and T. BURSTALL, *Heard Island* (1953); M.C. DOWNES *et al.*, *The Birds of Heard Island* (1959); P. TEMPLE, *The Sea and the Snow: The South Indian Ocean Expedition to Heard Island* (1966).

Australian Antarctic Territory: J.K. DAVIS, *High Latitude* (1962); P.G. LAW and J.M. BECHERVAISE, *Anare: Australia's Antarctic Outposts* (1957); D. MAWSON, *The Home of the Blizzard: Being the Story of the Australasian Antarctic Expedition, 1911–1914* (1915); A.G. PRICE, *The Winning of Australian Antarctica* (1962); R.A. SWAN, *Australia in the Antarctic* (1961).

(J.M.B.)

Australopithecus

Australopithecus is a name applied to fossilized remains of manlike higher primates known primarily from Africa. The remains range in time from late Pliocene to middle Pleistocene, a period of at least 3,000,000 years, probably more. Because of their anatomical structure and because stone implements have been found associated with some of the specimens, the australopithecines are generally believed to be closely related to man.

Discovery of the australopithecines

The first known australopithecine was found at Taung, South Africa, in 1924 and taken to Raymond Dart, then professor of anatomy at the University of the Witwatersrand, Johannesburg. Dart identified the fossil as an incomplete juvenile skull of a fossil primate not previously recognized. He named it *Australopithecus africanus* ("southern ape of Africa") and suggested that it was a form of ape transitional between typical apes and man. The scarcity of fossils of earlier forms of man at that time, coupled with the interpretive difficulties because the specimen is immature (skulls of young apes appear more manlike than those of adults), led to the rejection of Dart's view. There the matter rested for a decade, with the Taung skull unaccepted as a prehuman type.

In 1936 R. Broom, one of the few people who had accepted Dart's belief as to the possible relationship between the Taung skull and man, was shown fossil re-

mains discovered in mining operations at Sterkfontein, near Johannesburg. These he classified tentatively as australopithecine and began the search for more australopithecine fossils. Between 1936 and 1941, and again after World War II, many more fossil australopithecines were found by Broom and others in South Africa, eastern Africa, and Asia (see Table). Not all of these fossils were immediately identified as australopithecine, and in the case of one of these, *Gigantopithecus blacki,* it is now assumed that *Gigantopithecus* is not an australopithecine but may be ancestral to the australopithecine line. All other australopithecine fossils are generally accepted as belonging to the genera *Australopithecus* or *Paranthropus.*

Taxonomy of the australopithecines

Opinions differ greatly with respect to both the extent of the differentiation within the australopithecine group and whether any or all of the known forms are directly ancestral to modern man. The resolution of the latter problem will depend upon how the former is viewed. Most students incline to the view that all known australopithecines, placed in one genus, *Australopithecus,* were essentially similar in anatomy, behaviour, and ecology. This view is represented by such reputable scholars as Dart, who described the first australopithecine in 1925, and Sir Wilfrid E. Le Gros Clark. In contrast, since 1954 J. T. Robinson has argued that the australopithecines include two forms, which differ anatomically, behaviorally, and ecologically: *Paranthropus,* robust and apelike, and *Australopithecus,* gracile and manlike. In this view the latter is directly ancestral to early true man, *Homo erectus,* and is itself an offshoot of the *Paranthropus* lineage. Latterly this view has received increased support and many students accept that at least some differentiation—into two species—was present, but the majority still prefer to place these species in a single genus *Australopithecus.* The controversy over the role of australopithecines in the ancestry of man concerns primarily whether *Paranthropus* was directly in that ancestry or not; most students accept that *Australopithecus* was. Even here opinions differ; for example, P.V. Tobias and J. Napier believe, as did L.S.B. Leakey, that another lineage existed (*Homo habilis*) that was directly ancestral to true man, while *Paranthropus* and *Australopithecus* were closely related, but not directly ancestral, to it.

There is general agreement that the study of australopithecines is the study of the emergence of man; at the very least that it throws much light on the transition from apes to man. Much present understanding of this area of study, however, is tentative and speculative. This is partly because fossil evidence can at best represent a once-living population only incompletely, partly because of gaps in the known fossil record, and partly because samples are too small to reveal the range of variation in many significant features. Without knowledge of such variation the significance of differences is difficult to assess.

INTERPRETATION OF THE EMPIRICAL EVIDENCE

Interpreting the australopithecine material is difficult because the only direct evidence of the hominids themselves

Discoveries of Australopithecus and Related Fossils

year	site	name originally assigned to fossil	specimens*	discoverer(s)	present taxonomic status†
1924	Taung, Rep. S. Afr.	*Australopithecus africanus*	1	R. Dart	same
1935	Hong Kong	*Gigantopithecus blacki;* then *Giganthropus* (1945, F. Weidenreich, taxonomically invalid name)‡	5 teeth	G.H.R. von Koenigswald	same
1936	Sterkfontein, Rep. S. Afr.	*A. transvaalensis,* then *Plesianthropus transvaalensis*	20	R. Broom	*A. africanus*
1938	Kromdraai, Rep. S. Afr.	*Paranthropus robustus*	several specimens single adult/1 specimen juvenile	R. Broom	same, or *A. robustus*
1939	Lake Eyasi, Tanzania	*Meganthropus africanus* (Weinert, 1950)‡	2	L. Kohl-Larsen	usually *A. Africanus*
1939, 1941	Sangiran, Java	*Meganthropus palaeojavanicus;* and	1	G.H.R. von Koenigswald	both *Meganthropus*
		Pithecanthropus dubius (G.H.R. von Koenigswald, F. Weidenreich, 1945)‡	1		
1947–49	Sterkfontein, S. Afr.	*Plesianthropus transvaalensis*	80	R. Broom, J.T. Robinson	*A. africanus*
1947–62	Makapansgat, S. Afr.	*A. prometheus*	40	R. Dart	*A. africanus*
1948–53	Swartkrans, S. Afr.	*Paranthropus crassidens;* and	160	R. Broom, J.T. Robinson	*P. robustus* or *A. robustus;*
		Telanthropus capensis	several specimens/2 individuals		*Homo erectus*
1953	Sangiran, Java	*Meganthropus palaeojavanicus*	1	Marks	same
1956	Kwangsi Province, South China	*Gigantopithecus blacki*	3/1,000 teeth	W.C. Pei	same
1959–64	Olduvai Gorge, Tanzania	*Zinjanthropus boisei* (1959)‡	13	M. Leakey, L.S.B. Leakey	*A. robustus* or *P. robustus;*
	Lake Natron, Tanzania	"Chellean Man"; and *Homo habilis* (1964, Leakey, P.V. Tobias, J.R. Napier)‡	1 4 (?)		*Homo erectus;* *A. africanus* and *H. erectus*
1965	Lake Rudolf, Kenya	*A. africanus*	2	B. Patterson	same
1967	Omo, Ethiopia	*Paraustralopithecus aethiopicus*	1 (?)	C. Arambourg, R. Coppens	*A. robustus* or *P. robustus*
1968	Omo, Ethiopia	*A. robustus;* and	30 teeth	F.C. Howell	same
		A. africanus	2		same
1969	N. India	*Gigantopithecus bilaspurensis*	1	E.L. Simons, S.R.K. Chopra	same

*Only teeth are listed separately. Several specimens may belong to the same individual; a specimen is a bone fragment or bone. †Gracile forms all *A. africanus;* robust forms either *A. robustus* or *P. robustus,* depending on school of thought. *Homo erectus* is man, not man-ape. *Meganthropus* and *Gigantopithecus* are apes. ‡Date and author of name, if different from date of discovery and discoverer.

concerns parts of the skeleton and, in some cases, some aspects of material culture (occasional stone or bone tools). Much information needed to understand the lives of the early hominids—functional, behavioral, and ecological aspects—must be deduced from the narrow segment of the life and times of these creatures represented by the empirical evidence. Fortunately, the evidence also contains some information about climate and associated animals. An added difficulty is that the soundest available criterion of human status is the capacity to use conceptual thought; but this capacity cannot be deduced from the anatomical evidence available. Hence it is possible to properly classify as a hominid a fossil exhibiting accepted anatomical criteria of hominid status without its necessarily having been capable of conceptual thought. This dilemma appears at present to be insurmountable; cultural evidence has helped a little, but has not resolved it.

Physical characteristics of the australopithecines

Objective traits. The major characteristics of the australopithecine phase of hominid evolution are small brain size, within the size range of modern apes; dentition, manlike with canines not projecting significantly beyond adjacent teeth, tooth row compact and without diastemata (gaps between canine and premolars, common in animals with large, cutting canines); presence of anatomical adaptation for at least a moderately erect bipedal posture. The anatomical features associated with bipedal posture include low, relatively horizontal nuchal plane of the occiput (the area of the back of the head where the neck muscles attach; see Figure 1); pronounced lumbar curvature of the spine; reduced height of the iliac blade of the pelvis (see Figure 2); clearly defined greater sciatic notch produced by backward expansion of the iliac blade; presence of a zone of strengthened bone running upward from the acetabulum (cup-shaped joint for the thigh bone), associated with specialization of the gluteus medius and minimus muscles for lateral balance control; and proportionately increased width of the sacrum (base of the spine).

From W.E. Le Gros Clark, *The Fossil Evidence for Human Evolution* (1964); The University of Chicago Press

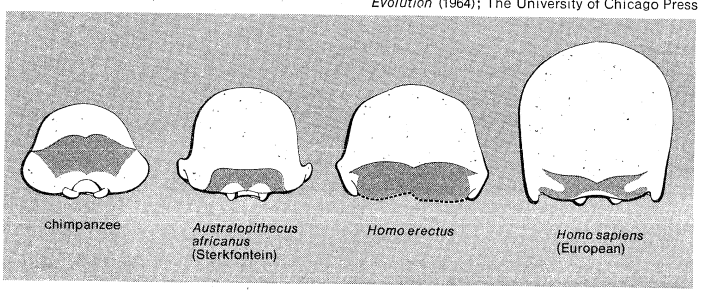


Figure 1: Occipital view of skulls with nuchal plane shaded.

These hominids had thus not departed noticeably from the ape grade of organization in brain volume but had departed entirely from the ape grade and become strongly manlike with respect to the dentition and with respect to erect posture. That they had these features in common does not mean that they were all alike. Most

The robust and gracile distinction

workers accept that two distinct forms are known, one robust (*Paranthropus*) and the other gracile (*Australopithecus*).

Considering anatomical characters, the division into robust and gracile forms is fairly clear-cut. A few major features will serve to illustrate the point. Although endocranial volume was apparently much the same in the two, in the gracile form the braincase rose relatively high above the upper orbital (eye sockets) margins; convexity of the frontal region produced a well-defined forehead; compound cranial crests (bony ridges along the top of the head to anchor heavy jaw muscles) were apparently rare, and the mastoid region below and behind the ear projected laterally to a moderate extent only. By contrast, in the robust form the braincase was proportionately low, within the ape range of variation and well away from that of the gracile form and of later human forms; the frontal region was flat and apelike, not convex; cranial crests (usually sagittal only) were normally present, and the mastoid region was well developed and laterally protuberant.

Similarly, although the dentitions of the two forms are more manlike than apelike, they differ significantly. That of the gracile form closely resembles that of *Homo erectus* (see HOMINIDAE) morphologically and in proportions along the tooth row. In contrast, the robust form has large cheek teeth, consistent with its robustness, but canines that are absolutely smaller than those of the gracile form, of *H. erectus,* and even of some races of modern man, and proportionally smaller than all of them. This curious proportionate smallness of canines—and to some extent of incisors also—occurs in the specimens from South and East Africa (*Paranthropus*) and Java (*Meganthropus*) throughout the several million years in which the robust form is known. It was thus a significant, stable, long-term characteristic of the robust lineage. The normal presence of a sagittal crest (ridge along the top of the head from front to back) resulted from the large size of the temporal muscles attached to a robust jaw; in larger individuals these produced partial compound crests in conjunction with the nuchal muscles.

Both forms have the essential pelvic, lumbar, and nuchal features associated with erect posture, but, again, they are not identical. The gracile form had the short ischium (see Figure 2) and long lower limb of modern man and a remarkably modern foot. The robust form, in contrast, had a long, apelike ischium apparently associated with a proportionately short lower limb. Very scanty evidence suggests a mobile and flexible foot intermediate in character between ape and human feet.

From W.E. Le Gros Clark, *The Fossil Evidence for Human Evolution* (1964); The University of Chicago Press

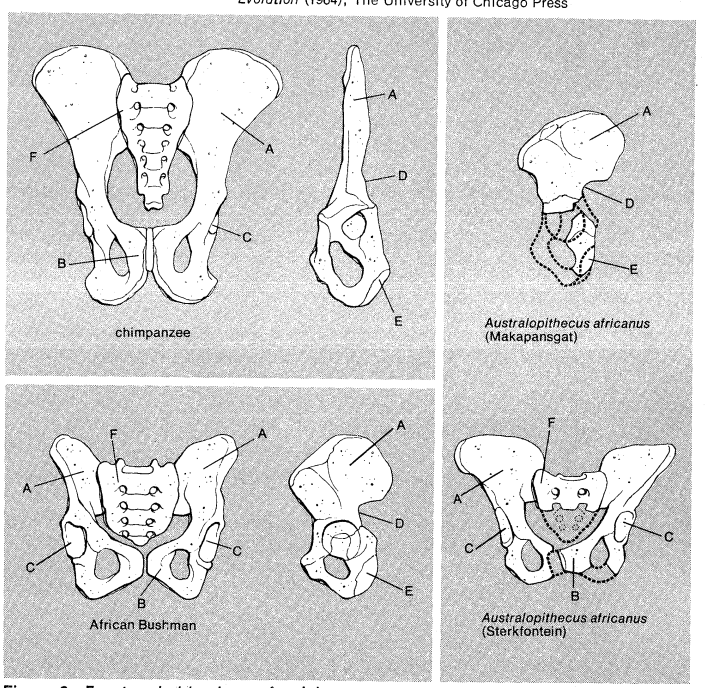


Figure 2: *Front and side views of pelvis.*
(A) Iliac blade. (B) Pubic bone. (C) Acetabulum (hip joint). (D) Sciatic notch. (E) Ischium. (F) Sacrum.

On available evidence the two forms are distinguishable on any known part of the skeleton. Larger samples, by providing better knowledge of the normal range of variation among australopithecines, might show the degree of difference between the forms to be less important. The features discussed, however, are ones that do not vary widely within species of living higher primates and since they concern major adaptations, they are hardly likely to vary widely within the same population. The smallness of the samples is partially offset by this fact.

Behavioral and ecological inferences. *Posture and gait.* Both types of australopithecine regularly used erect posture. This conclusion follows from such anatomical features as the low area of attachment for neck muscles

on the back of the head (Figure 1), well-developed lumbar curvature, wide sacrum, and low iliac blade strongly expanded backward and more laterally placed than is that of apes. The strengthened buttress of bone passing upward from the hip joint indicates a fully established and regularly used lateral balance control mechanism. Moreover, the long lower limb and short ischium of the gracile form strongly suggest that it was able to stride and that the propulsive mechanism worked as it does in modern man. The robust form, however, with a power-oriented propulsive mechanism and mobile foot, was not as manlike as the gracile form and in this respect evidently represents a compromise between ape and man. Possibly it spent some time climbing trees, which would explain the need for shorter and more power-oriented lower limbs; when on the ground it probably used erect posture but almost certainly did not have a fully developed striding gait.

Tool use. There is no clear evidence that either form used or made tools in the earliest levels from which they are known. Since investigation of the earliest forms is very new, further work may change the evidence on this point. Stone tools do occur in the lower Pleistocene (about 2,000,000 years ago); *e.g.*, in the excavation known as Bed I at Olduvai Gorge in Tanzania. Good evidence indicates that bones and other objects were used as tools; one may speculate that such activity preceded stone tool making, though there is no clear evidence of this. There is no evidence that at that time hominids more advanced than australopithecines existed. The presumption is thus that either or both australopithecine forms had developed the capacity to use and make tools. It is difficult to say which group made the tools. On three grounds it is more likely to have been the gracile form: (1) this form is anatomically much more manlike than the robust form; (2) in Beds I and II at Olduvai the artifacts improve with time during a period when the gracile australopithecine appears to become more manlike, possibly even altering into a true *H. erectus*, which is already present by the end of Bed II time; but there is no evidence of similar change in the robust form; (3) cultural capacity represents so powerful an adaptive mechanism that it is improbable that two hominids as anatomically different as the australopithecines could have been developing it simultaneously and yet have lived in the same region over at least 1,000,000 years. If only one form had culture, it is reasonable to believe that it was the more manlike one of the two.

Use of tools

Hunting and diet. For much of his history man appears to have been a hunter-gatherer, as some technologically less advanced peoples are today. But nonhuman, higher primates are virtually exclusively herbivores of some sort, following a way of life widely different from that of man. Presumably this will also have been true of the apelike ancestors from which man came; hence at some stage the herbivorous ecology will have given way to an omnivorous hunter-gatherer ecology based on culture. It is commonly assumed that the australopithecines had already made this transition and were primitive hunters. The presence of stone tools in the record during the latter half of their known history supports this conclusion, since it is unlikely that herbivores would have developed an active stone technology. The argument that the plentiful associated animal bones prove the same point is less secure. The South African specimens occur in ancient limestone caverns that are commonly regarded as having been dwelling sites and the associated fossils as food remains. During the period of accumulation, however, the caverns were deep underground and could not have been occupation sites; instead they were traps for material lying about on the surface. Moreover, the accumulation rate was so slow that one bone falling in per decade or even one per century would account for the accumulation. Australopithecine activity may have been involved; but so was that of many other animals.

Evidence of hunting technology

The hunter-gatherer type of ecology is reasonably associated with the gracile form. It was erect and efficiently bipedal, had a masticatory apparatus quite similar to that of *H. erectus* and later forms of man, was able to live in arid plains conditions, and evidently used bones and other objects as tools and made tools of stone in its later history.

The robust form was anatomically more primitive than the gracile form, had a masticatory apparatus somewhat unlike that of *H. erectus* and later forms of man, was not as efficiently bipedal as the gracile form, probably needed a wetter environment, coexisted with both the gracile form and later with *H. erectus*, and as a herbivore is less likely to have been actively culture-bearing than the gracile form. These considerations make it doubtful that it was a hunter-gatherer and are more in keeping with the suggestion that it had not seriously embarked upon the transition to a manlike way of life.

Geologic age and geographic distribution. Until quite recently all known australopithecines were of late lower Pleistocene or early middle Pleistocene age (1,000,000–500,000 years old). The rich South African sites are difficult to date; the only usable technique so far is faunal comparison. By this method, the sites yielding the gracile form, *Australopithecus africanus,* are all of late lower Pleistocene age, Taung and the lower strata of Sterkfontein probably being slightly older than Makapansgat, while the stone-tool-yielding middle levels at Sterkfontein probably was the latest in this sequence. Swartkrans is apparently only a little later, from the end of the lower or beginning of the middle Pleistocene. Still later, of middle Pleistocene age, is Kromdraai. Swartkrans and Kromdraai yielded the robust form, while *H. erectus* is also present at Swartkrans. The gracile form sites all reflect arid climates, with the latest, the Sterkfontein middle levels, beginning to trend to moister conditions. Swartkrans was from somewhat wetter times and Kromdraai wetter still.

Dating the australopithecines

Relative dating methods suggest that the Bed I and II sequence at Olduvai covers much the same time period as do the South African sites. Fortunately, the Olduvai beds include volcanic material that provides suitable minerals for the use of the potassium-argon radiometric dating method. This method is not without its difficulties and much remains to be settled about the Olduvai dates. The available evidence, however, indicates that the age ranges from approximately 2,000,000 years at the base of Bed I to about 500,000 years at the top of Bed II. This dating is part of the evidence that has recently led to the acceptance of an age of 2,000,000 or 2,500,000 years for the Pleistocene, rather than 1,000,000. The age of the Sangiran specimens is more controversial but probably falls near the later end of the Olduvai Bed I and II sequence.

The newer East African and Ethiopian material has not been studied sufficiently for the dating to be well established. Some of the Omo specimens have potassium-argon dates of between 3,000,000 and 4,000,000 years, and there is some suggestion of dates in Kenya reaching back even further. It would be premature to accept these dates now, but they do indicate the probability that the earlier material is of late Pliocene age. Since both the robust and the gracile forms appear to be represented at these early times in East Africa and Ethiopia the presence of the gracile form well before the robust one in the South African sequence is not due to the former being the older of the two. Evidently the robust form did not occur there during the more arid climatic time but moved in from the north when the climate ameliorated.

The australopithecines are found in eastern and southern Africa and in southern Asia. The gracile form thus far is known from Africa only; it is well represented in both East and South Africa. According to present evidence, it existed in Africa as early as the robust form, but in both South and East Africa it apparently disappeared from the record somewhat earlier than the robust form. Its disappearance seems to coincide approximately with the appearance of *H. erectus.* The robust form apparently had a wider distribution; it is well represented in South and East Africa, and many, though not all, workers accept that *Meganthropus* from Java is also a representative of the robust form, although others consider it a representative of *H. erectus.* In Sangiran, Java, at Swart-

krans in South Africa, and in East Africa good evidence indicates that the robust form and *H. erectus* simultaneously occupied the same region. In East Africa, there is evidence of simultaneous presence of the robust and the gracile australopithecines. There is as yet no sound evidence of the gracile form and *H. erectus* living together. On present evidence, then, the robust form had a wide distribution in the Old World; the gracile form was apparently confined to Africa.

CONTROVERSIES CONCERNING CLASSIFICATION

As already noted, the view that two forms are represented in the early hominid material is now commonly accepted: the gracile *Australopithecus africanus*, and the more robust form usually referred to as *Australopithecus robustus*, but as *Paranthropus robustus* by some. Followers of this two-stream hypothesis differ concerning the magnitude of the difference between them. The majority believe that the differences are relatively slight and can best be represented by treating the two as species of one genus.

Some workers (*e.g.*, Le Gros Clark and Dart) believe that even species separation may be unwarranted; the two forms are seen as variants of the same species much as two races of modern man are variants of the species *Homo sapiens*. This hypothesis has recently appeared in extreme form as the view that the robust and gracile forms are merely males and females of the same species. This latter hypothesis need hardly be taken seriously in view of the South African evidence. It is difficult to accept that during the thousands or hundreds of thousands of years that Sterkfontein was accumulating, only females were present or only female remains were preserved, and that for similar periods in Swartkrans time only males were present or preserved. Moreover, the nature of the variation present in these samples suggests that each contained two sexes. Finally, if it is true that the gracile form disappears from the record before the robust form, the implication would be that males survived long after females were extinct.

The racial-variant hypothesis is probably almost as difficult to support. Animal subspecies or races are local geographic variants of a species; by definition, subspecies of the same species are not sympatric (*i.e.*, occupying the same territory), except in exceptional circumstances such as the overlapping ends of a ringlike distribution. Modern man is also unusual because relatively advanced means of travel can bring widely different races together quickly, hence a number of racial variants can be found together. But history shows, and theory predicts, that interbreeding soon begins to break down and submerge the differences between them. It is very improbable that the simultaneous occurrence in one region of different racial variants of man could continue over a period of millions of years, as in the case of the early hominids. The existence over millions of years of a sharp distinction between the two australopithecine forms implies that they were genetically isolated from each other and hence that they were at least distinct species.

In the two-stream hypothesis, the species seems to be the lowest taxonomic level at which the two forms may reasonably be placed, and this is the present majority view. The alternate to this view of the two-stream hypothesis is that the anatomical, behavioral, and ecological differences between the two are so pronounced that assigning separate generic status is preferable. This interpretation was originally proposed by J.T. Robinson, continues to be defended by him, and is supported by some other workers. This view interprets the masticatory apparatus differences and cultural evidence as meaning that *Paranthropus* was an essentially cultureless herbivore but *Australopithecus* was a culture-bearing, hunting omnivore. It interprets the anatomical evidence on posture and locomotion as indicating that *Paranthropus* was specialized for a combination of a moderately efficient, erect, bipedal posture on the ground with some arboreal climber activity, while *Australopithecus* was a highly efficient, ground-dwelling biped. Interpreted in this manner, the levels of organization of the two are very different, *Paranthropus* being a somewhat hominized ape, *Australopithecus* a primitive man with evolutionary potential vastly superior to that of *Paranthropus*. This view recognizes a common taxonomic viewpoint that the genus defines a clear-cut way of life and that the species within the genus represent specializations within that way of life.

An objection raised to this view is that differences as great as those observed between *Paranthropus* and *Australopithecus* occur within a single genus of living pongids (anthropoid apes) or within the modern species *H. sapiens*. Dental, cranial, and pelvic characters, however, tend to indicate that this belief is mistaken.

Finally, some workers believe that more than two basic lineages occur in the known australopithecine material. This view is represented, for example, by that of Leakey and Tobias, who accepted a single genus, *Australopithecus*, for all hominids regarded by them as australopithecines but that three different forms of equal status were involved. These are *A. africanus*, *A. robustus*, and *A. boisei* (Olduvai) in Tobias' view but subgenerically distinct in Leakey's view as *A.* (*Australopithecus*) *africanus*, *A.* (*Paranthropus*) *robustus*, and *A.* (*Zinjanthropus*) *boisei*. Both excluded the Olduvai Bed I and II material named *Homo habilis*, believing this to represent another independent lineage. They disagreed, however, on which specimens belong in *H. habilis* and on what its affinities are. Tobias labels only the Bed I material *H. habilis* and believes it to be directly ancestral to the Bed II material, which he believes is *H. erectus*, ancestral to *H. sapiens*. Leakey, in contrast, believed that both the Bed I and Bed II material belongs to *H. habilis* and that this lineage was directly ancestral to *H. sapiens*. In his view, *H. erectus* is yet another independent lineage. This is the most multilineal of the current interpretations of the early African hominids. Most observers believe, however, that *H. habilis* is not separable from the *A. africanus* lineage and that *H. erectus* is a stage between *A. africanus* and *H. sapiens*.

This very brief review of opinion about australopithecine history and affinities shows how much uncertainty exists in the field. Much more detailed study and more material is needed to introduce more certainty into the picture. Because of the difficult nature of this field, however, which in some respects must remain speculative, considerable difference of opinion may always exist and absolute certainty will always be impossible.

BIBLIOGRAPHY. W.W. BISHOP and J.D. CLARK (eds.), *Background to Evolution in Africa* (1967), a valuable collection of papers providing background information concerning hominid evolution in Africa; M.H. DAY, *Guide to Fossil Man* (1965), a well-illustrated guide to the better specimens of fossil man; W.E. LE GROS CLARK, *The Fossil Evidence for Human Evolution*, 2nd ed. (1964), a fairly brief, authoritative discussion of the evidence for human evolution from the earliest hominids down to the present, and *Man-Apes or Ape-Men?* (1967), a work covering the *Australopithecus* phase of evolution, including some historical and anecdotal material; J.T. ROBINSON, "Adaptive Radiation in the Australopithecines and the Origin of Man," in F.C. HOWELL and F. BOURLIERE (eds.), *African Ecology and Human Evolution* (1963), an outline of the anatomical differences between the two major forms of early hominid with an attempt to interpret them in terms of differences in ecology and behaviour, and "Variation and the Taxonomy of the Early Hominids," in T. DOBZHANSKY, M.K. HECHT, and W.C. STEERE (eds.), *Evolutionary Biology*, vol. 1 (1967), a discussion of the importance of analyzing variation in interpreting early hominid fossils that provides evidence of two forms of early hominid; P.V. TOBIAS, *The Cranium and Maxillary Dentition of Australopithecus (Zinjanthropus) boisei*, vol. 2 of L.S.B. LEAKEY (ed.), *Olduvai Gorge* (1967), very detailed description of one good skull of an East African early hominid; C. STARR (ed.), *Anthropology Today* (1971), a well-illustrated discussion of the whole field of anthropology containing useful material on the *Australopithecus* phase of human evolution.

(J.T.R.)

Austria

Located in the centre of Europe, Austria, a federal republic with a scenic and largely mountainous terrain of 32,375 square miles (83,850 square kilometres), and a population exceeding 7,000,000 in the 1970s, is one

of the smaller nations of the world. In the decades following the collapse, in 1918, of the multinational Austro-Hungarian Empire of which it had been the heart, Austria experienced more than a quarter-century of social and economic turbulence and a Nazi dictatorship. Yet the establishment of permanent neutrality in 1955, associated with the withdrawal of the four-power troops that had occupied the country for a decade, enabled Austria to develop into a stable and socially progressive nation, with a flourishing cultural life reminiscent of its earlier days of international musical glory. Its social and economic institutions, too, have been characterized by new forms and a spirit of cooperation, and although political and social problems remain, they have not erupted with the intensity evidenced in other countries of the Continent.

Geographical position

Much of Austria's current status can be attributed to its geographical position: it is the centre of European traffic between east and west along the great Danubian trade route, and between north and south through the magnificent Alpine passes. Austria is bordered on the west by Switzerland, and together the countries form what has been characterized as the neutral core of Europe. The tiny Principality of Liechtenstein also forms a small enclave on the west. Hungary on the east; the Federal Republic of Germany (West Germany) and Czechoslovakia on the north; and Italy and Yugoslavia on the south, illustrate the variety of political and economic systems within which contemporary Austria is embedded (for further details see ALPS MOUNTAIN RANGES; ALPINE LAKES; AUSTRIA, HISTORY OF; and VIENNA).

THE LANDSCAPE

The natural landscape. *Relief features.* Mountains and forests give the Austrian landscape its character, although in the northeast of the country the Danube winds between the eastern edge of the Alps and the hills of Bohemia and Moravia, heading toward the Hungarian Plain. Vienna, the national capital, and one of the great cities of Europe, lies in the area where the Danube emerges from the mountains into the drier plains.

The landscape of the eastern Alps offers a complex geological and topographical pattern, with the highest elevation—the Grossglockner (12,457 feet [3,797 metres])—being reached toward the west. The Austrian Alps may be subdivided into a northern and southern limestone range, each of which is composed of rugged mountains, separated by a central range, softer in form and outline, and composed of crystalline rocks. North of this massive Alpine spur, which forms the physical backbone of the country, lies a hilly subalpine region, stretching in a zone between the northern Alps and the Danube, while to the north of that river lies a further, richly wooded, foothill area. The lowland area downstream from Vienna may be regarded as a western extension of the great Hungarian Plain.

Drainage. Austria is a land of lakes, many of them a legacy of Ice Age erosion, which scooped out mountain lakes in the central Alpine district, notably around Salzkammergut. The largest lakes—lying partly in the territory of neighbouring countries—are the Bodensee (Lake of Constance) in the west and the marshy Neusiedlersee (Neusiedler Lake) to the east.

Ninety-six percent of Austrian territory drains to the Danube River system. The main watershed between the Black Sea and the North Sea runs across northern Austria, in some places lying only 22 miles from the Danube, while to the west the watershed between the Danube and the river systems emptying into the Atlantic and the Mediterranean coincides with the western political boundary of Austria. In the south, the Julian and Karnische Alps, and, farther to the west, the main Alpine range, mark the watershed of the region draining into the Po River of northern Italy.

Climate. The wooded slopes of the Alps and the small portion of the plains of southeastern Europe are characterized by differing climatic zones: the wetter western regions of Austria have an Atlantic climate with a yearly rainfall of over 39.4 inches (1,000 millimetres), whereas the eastern regions, in particular those under the influence of the drier, more continental type of climate, have less.

In the lowlands and the hilly eastern regions, the median temperature ranges between 30.4° F (−0.9° C) in January and 68.6° F (20.3° C) in July. In those regions above 10,000 feet, by contrast, the temperature range is between 11.8° F (−11.3° C), with a snow cover of about 10 feet in January, and 35.8° F (2.1° C) in July, with about five feet of snow cover.

The prevailing wind direction is from the west, and the humidity, therefore, is highest in the west, diminishing toward the east.

Vegetation and animal life. Two-thirds of the total area of Austria is covered by woods and meadows, and the country is the most densely forested nation in central Europe. Spruce dominates the forests, with larch, beech, and oak also making a significant contribution. In the Alpine and foothill regions coniferous trees predominate, while leaf-bearing deciduous trees are more frequent in the warmer zones.

Wild animals, now protected by conservation laws, include the brown bear, the eagle, and all buzzard species, as well as falcons and owls, cranes, swans, and storks. Game hunting is restricted to certain periods of the year: deer and rabbits are the most frequent quarry. Austrian rivers nurture river and rainbow trout, grayling, hake, pike, perch, and carp.

Traditional regions. Western Austria, comprising the *Bundesländer* (states) of Vorarlberg, Tirol, and Salzburg, is characterized by high Alpine regions with majestic mountains and magnificent scenery. This high Alpine character also extends to the western part of Kärnten (Carinthia), the Salzkammergut of central Austria, to the Alpine blocks of Steiermark (Styria), and to the eastern rim of the Alps.

Alpine scenery

The outer fringes of the Alps dominate the northern portion of Oberösterreich (Upper Austria) with their subalpine characteristics, while the richly wooded Bohemian massif extends across the Czechoslovak border. This part of Austria is furrowed by many valleys that, for centuries, served as passageways leading to the east and southeast of Europe and even, in the case of medieval pilgrims and crusaders, to the Holy Land.

A hilly vineyard district extends into the *Bundesland* of Burgenland. This region, together with the eastern part of the state of Steiermark, lies at the gateway to the Hungarian Plain. Taken together, Burgenland, Steiermark, Kärnten, and Niederösterreich (Lower Austria) form a region, throughout which agriculture abounds. The city of Vienna and its peripheral suburbs and industrial settlements border both Alpine and lowland regions.

Human influence on the landscape. The pattern of rural settlement in Austria was shaped centuries ago by the exigencies of the Alpine environment, and new rural building is still influenced by these ancient traditions, particularly in the west and the centre of the country. Rural housing in the eastern parts of the nation, particularly in the lowlands, is, in contrast, dominated more by agricultural needs than by harsh weather conditions.

Forty-four percent of Austria's population in 1971 lived in cities and towns with more than 10,000 inhabitants: Austria is thus not only a mountainous but also a highly urbanized country. Over one-fifth of the total Austrian population lives in Vienna. Furthermore, the state capitals have, as a consequence of national economic and social development, grown somewhat in population since the end of World War II. Graz, Austria's second-biggest city with 249,000 inhabitants in 1971, is the gateway to the Balkans; Innsbruck is the rail and road centre through which national north–south traffic of western Austria passes; Salzburg is one of the best-known European centres of musical culture and Baroque architecture; the Linz of the 1970s has become an important industrial centre; and Klagenfurt lies astride routes giving access to both Italy and Yugoslavia.

Major Austrian cities

THE PEOPLE

Ethnic and linguistic heritage. The Austrian population had almost reached 7,500,000 by the early 1970s,

MAP INDEX

Political subdivisions

Burgenland.....47·30n 16·20e
Kärnten.....46·50n 13·50e
Niederösterreich.....48·20n 15·50e
Oberösterreich..48·15n 14·00e
Salzburg.....47·25n 13·15e
Steiermark.....47·10n 15·10e
Tirol.....47·15n 11·20e
Vienna.....48·13n 16·23e
Vorarlberg.....47·15n 9·55e

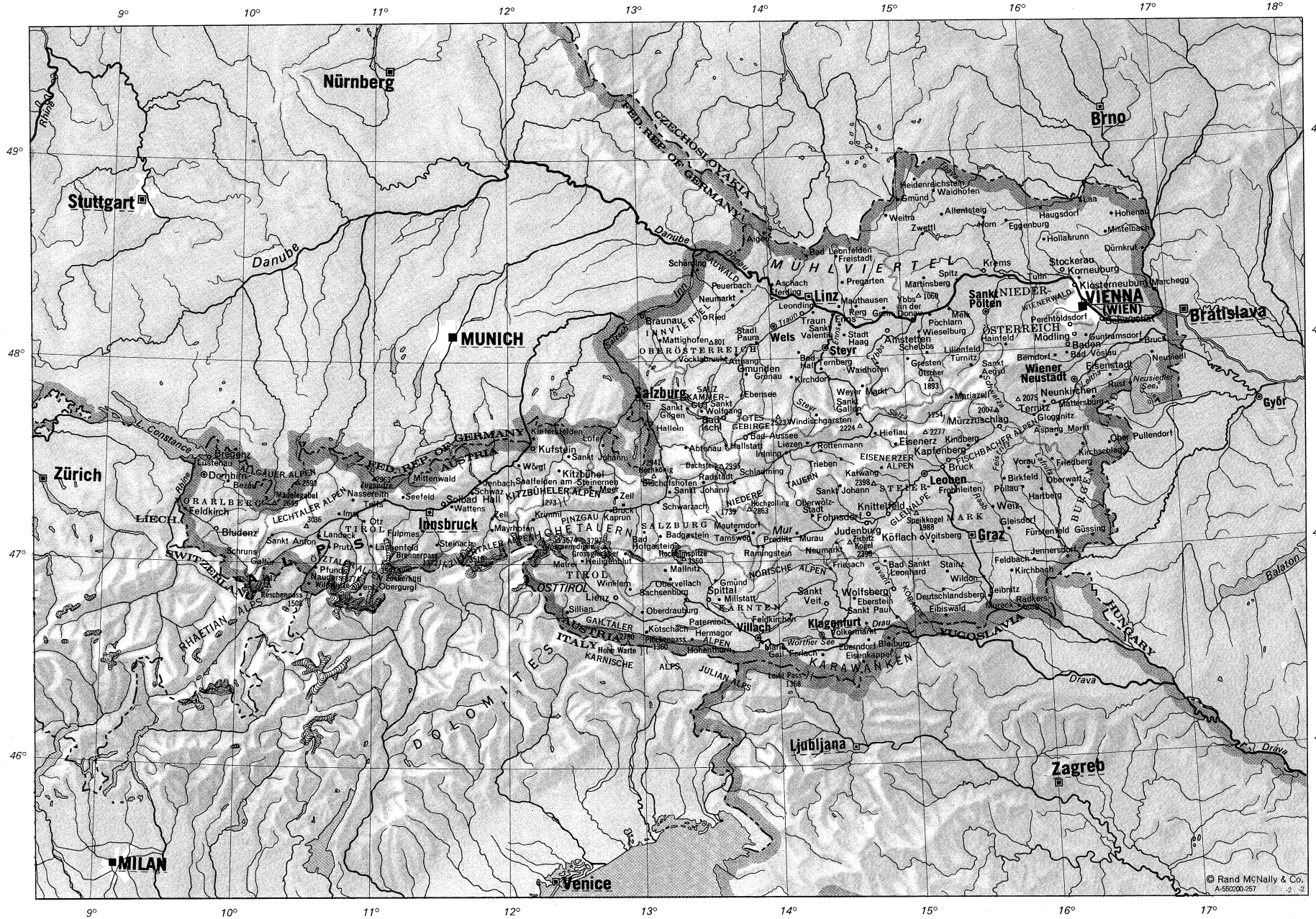
The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Abtenau.....47·33n 13·21e
Aigen [im Mühlkreis].....48·39n 13·58e
Allentsteig.....48·42n 15·20e
Amstetten.....48·07n 14·53e
Aschach [an der Donau].....48·22n 14·02e
Aspang Markt..47·33n 16·06e
Attnang.....48·01n 13·43e
Bad Aussee....47·36n 13·47e
Baden.....48·00n 16·14e
Badgastein.....47·07n 13·08e
Bad Hall.....48·02n 14·13e
Bad Hofgastein.47·10n 13·06e
Bad Ischl.....47·43n 13·37e
Bad Leonfelden.48·33n 14·19e
Bad Sankt Leonhard [im Lavanttal].....46·58n 14·48e
Bad Vöslau.....47·57n 16·16e
Berndorf.....47·57n 16·08e
Bezau.....47·23n 9·54e
Birkfeld.....47·21n 15·42e
Bischofshofen..47·25n 13·13e
Bleiburg.....46·35n 14·48e
Bludenz.....47·09n 9·49e
Braunau [am Inn].....48·15n 13·02e
Bregenz.....47·30n 9·46e
Bruck.....47·17n 12·49e
Bruck [an der Leitha].....47·57n 16·44e
Bruck [an der Mur].....47·25n 15·16e
Deutschlandsberg.....46·49n 15·13e
Dornbirn.....47·25n 9·44e
Dürnkrut.....48·28n 16·51e
Ebensee.....47·48n 13·46e
Eberndorf.....46·35n 14·38e
Eberstein.....46·48n 14·34e
Eferding.....48·18n 14·02e
Eggenburg.....48·39n 15·50e
Eibiswald.....46·41n 15·15e
Eisenerz.....47·33n 14·53e
Eisenkappel....46·29n 14·36e
Eisenstadt.....47·51n 16·32e
Enns.....48·13n 14·29e
Feldbach.....46·57n 15·54e
Feldkirch.....47·14n 9·36e
Feldkirchen [in Kärnten].....46·43n 14·05e
Ferlach.....46·31n 14·18e
Fohnsdorf.....47·13n 14·41e
Freistadt.....48·31n 14·31e
Friedberg.....48·01n 13·15e
Friesach.....46·57n 14·24e
Frohnleiten.....47·16n 15·20e
Fulpmes.....47·10n 11·21e
Fürstenfeld.....47·03n 16·05e
Galtür.....46·58n 10·11e
Gänserndorf....48·20n 16·43e
Gleisdorf.....47·06n 15·44e
Gloggnitz.....47·40n 15·57e
Gmünd.....48·47n 15·00e
Gmünd.....46·54n 13·32e
Gmunden.....47·55n 13·48e
Götzis.....47·20n 9·38e
Graz.....47·05n 15·27e
Grein.....48·14n 14·51e
Gresten.....48·00n 15·02e
Grünau [im Almtal].....47·51n 13·57e
Guntramsdorf...48·03n 16·19e
Güssing.....47·04n 16·20e
Hainburg an der Donau.....48·09n 16·57e
Hainfeld.....48·02n 15·46e
Hallein.....47·41n 13·06e
Hallstatt.....47·33n 13·39e
Hartberg.....47·17n 15·59e
Haugsdorf.....48·42n 16·05e
Heidenreichstein.....48·52n 15·07e
Heiligenblut....47·02n 12·50e
Hermagor.....46·37n 13·22e
Hieflau.....47·36n 14·44e
Hohenau.....48·36n 16·55e
Hohenthurn....46·33n 13·40e
Hollabrunn.....48·34n 16·05e
Horn.....48·39n 15·39e
Imst.....47·14n 10·44e
Innsbruck.....47·16n 11·24e
Irdning.....47·33n 14·01e
Jenbach.....47·24n 11·47e
Jennersdorf.....46·57n 16·08e
Judenburg.....47·10n 14·40e
Kalwang.....47·26n 14·46e
Kapfenberg.....47·26n 15·18e
Kaprun.....47·16n 12·46e
Kindberg.....47·31n 15·27e
Kirchbach [in Steiermark]....46·54n 15·44e
Kirchdorf [an der Krems]....47·56n 14·14e
Kirchschlag [in der Buckligen Welt].....47·31n 16·18e
Kitzbühel.....47·27n 12·23e
Klagenfurt.....46·38n 14·18e
Klosterneuburg.48·18n 16·20e
Knittelfeld.....47·14n 14·50e
Köflach.....47·04n 15·05e
Korneuburg....48·21n 16·20e
Kötschach.....46·40n 13·00e
Krems [an der Donau].....48·25n 15·36e
Krimml.....47·13n 12·11e
Kufstein.....47·35n 12·10e
Laa [an der Thaya].....48·43n 16·23e
Landeck.....47·08n 10·34e
Längenfeld.....47·04n 10·58e
Leibnitz.....46·48n 15·32e
Leoben.....47·23n 15·06e
Leonding.....48·16n 14·15e
Lienz.....46·50n 12·47e
Lilienfeld.....48·03n 15·36e
Linz.....48·10n 14·18e
Lofer.....47·35n 12·41e
Lustenau.....47·26n 9·39e
Mallnitz.....46·59n 13·10e
Marchegg.....48·17n 16·55e
Maria Gail.....46·36n 13·52e
Mariazell.....47·47n 15·19e
Martinsberg....48·22n 15·09e
Matrei [in Osttirol].....47·00n 12·32e
Mattersburg....47·44n 16·25e
Mattighofen....48·06n 13·09e
Mauterndorf....47·08n 13·40e
Mauthausen....48·14n 14·32e
Mayrhofen.....47·10n 11·52e
Melk.....48·14n 15·20e
Millstatt.....46·48n 13·35e
Mistelbach [an der Zaya].....48·34n 16·35e
Mittersill.....47·16n 12·29e
Mödling.....48·05n 16·17e
Murau.....47·07n 14·10e
Mureck.....46·42n 15·36e
Mürzzuschlag ..47·36n 15·41e
Nassereith.....47·19n 10·50e
Nauders.....46·53n 10·30e
Neumarkt [in Steiermark]....47·05n 14·26e
Neunkirchen ...47·43n 16·05e
Neusiedl [am See].....47·57n 16·51e
Oberdrauburg ..46·45n 12·58e
Obergurgl.....46·52n 11·01e
Oberpullendorf .47·31n 16·31e
Obervellach....46·56n 13·12e
Oberwart.....47·17n 16·13e
Oberwölz-Stadt.47·13n 14·17e
Ötz.....47·12n 10·54e
Paternion.....46·43n 13·38e
Perchtoldsdorf..48·07n 16·17e
Perg.....48·15n 14·37e
Peuerbach.....48·21n 13·56e
Pfunds.....46·58n 10·33e
Pöchlarn.....48·12n 15·13e
Pöllau.....47·18n 15·51e
Predlitz.....47·04n 13·55e
Pregarten.....48·21n 14·22e
Prutz.....47·05n 10·40e
Radkersburg ...46·41n 15·59e
Radstadt.....47·23n 13·27e
Ramingstein....47·04n 13·50e
Ried [im Innkreis].....48·13n 13·30e
Rottenmann....47·31n 14·22e
Rust.....47·48n 16·41e
Saalfelden am Steinernen Meer.....47·23n 12·38e
Sachsenburg...46·50n 13·21e
Salzburg.....47·48n 13·02e
Sankt Aegyd [am Neuwalde]47·52n 15·35e
Sankt Anton [am Arlberg]..47·08n 10·16e
Sankt Gallen....47·41n 14·37e
Sankt Gilgen....47·46n 13·22e
Sankt Johann [am Tauern]...47·22n 14·29e
Sankt Johann [im Pongau]...47·21n 13·12e
Sankt Johann [in Tirol].....47·31n 12·26e
Sankt Paul [im Lavanttal].....46·42n 14·52e
Sankt Pölten....48·12n 15·37e
Sankt Valentin..48·10n 14·32e
Sankt Veit [an der Glan].....46·46n 14·21e
Sankt Wolfgang [im Salzkammergut].....47·44n 13·27e
Schärding.....48·27n 13·26e
Scheibbs.....48·00n 15·10e
Schladming....47·23n 13·41e
Schruns.....47·04n 9·55e
Schwarzach [im Pongau].....47·19n 13·09e
Schwechat.....48·08n 16·29e
Seefeld [in Tirol].....47·20n 11·11e
Sillian.....46·45n 12·25e
Solbad Hall [in Tirol].....47·17n 11·31e
Spittal [an der Drau].....46·48n 13·30e
Spitz.....48·22n 15·25e
Stadl Paura.....48·05n 13·53e
Stadt Haag.....48·06n 14·34e
Stainz.....46·54n 15·16e
Steinach.....47·05n 11·28e
Steyr.....48·03n 14·25e
Stockerau.....48·23n 16·13e
Tamsweg.....47·08n 13·48e
Telfs.....47·10n 11·22e
Ternberg.....47·58n 14·22e
Ternitz.....47·44n 16·03e
Traun.....48·13n 14·14e
Trieben.....47·29n 14·30e
Tulln.....48·19n 16·10e
Türnitz.....47·57n 15·30e
Vent.....46·52n 10·56e
Vienna (Wien)...48·13n 16·23e
Villach.....46·36n 13·50e
Vöcklabruck....48·01n 13·39e
Voitsberg.....47·03n 15·10e
Völkermarkt.....46·39n 14·38e
Vorau.....47·25n 15·54e
Waidhofen [an der Thaya].....48·49n 15·18e
Waidhofen [an der Ybbs].....47·58n 14·47e
Wattens.....47·17n 11·36e
Weitra.....48·42n 14·54e
Weiz.....47·13n 15·37e
Wels.....48·10n 14·02e
Weyer Markt....47·52n 14·41e
Wiener Neustadt.....47·49n 16·15e
Wieselburg.....48·08n 15·09e
Wildon.....46·53n 15·31e
Windischgarsten.....47·44n 14·20e
Winklern.....46·52n 12·52e
Wolfsberg.....46·51n 14·51e
Wörgl.....47·29n 12·04e
Ybbs an der Donau.....48·11n 15·05e
Zell [am See]...47·19n 12·47e
Zell [am Ziller]..47·14n 11·53e
Zwettl.....48·37n 15·10e

Physical features and points of interest

Allgäuer Alpen, *mountains*.....47·20n 10·25e
Alps, *mountains*.....47·00n 10·30e
Bodensee, *lake*..47·35n 9·25e
Brennerpass, *pass*.....47·00n 11·30e
Dachstein, *mountains*.....47·29n 13·36e
Danube (Donau), *river*..48·10n 17·03w
Drau, *river*.....46·37n 14·58e
Eisenerzer Alpen, *mountains*.....47·28n 14·45e
Enns, *river*.....48·14n 14·32e
Feistritz, *river*...47·01n 16·08e
Fischbacher Alpen, *mountains*.....47·28n 15·30e
Gailtaler Alpen, *mountains*.....46·42n 13·00e
Gleinalpe, *mountains*.....47·15n 15·03e
Grossglockner, *mountain*.....47·04n 12·42e
Grossvenediger, *mountain*.....47·06n 12·21e
Hochalmspitze, *peak*.....47·01n 13·19e
Hochgolling, *mountain*.....47·16n 13·45e
Hochkönig, *mountain*.....47·25n 13·04e
Hohe Tauern, *mountains*.....47·10n 12·30e
Hohe Warte, *mountain*.....46·37n 12·53e
Inn, *river*.....47·43n 12·10e
Innviertel, *physical region*.48·10n 13·15e
Karawanken, *mountains*.....46·30n 14·25e
Kitzbüheler Alpen, *mountains*.....47·20n 12·20e
Koralpe, *mountains*.....46·50n 14·58e
Lafnitz, *river*....47·01n 16·15e
Lavant, *river*....46·38n 14·57e
Lechtaler Alpen, *mountains*.....47·15n 10·30e
Leitha, *river*.....47·57n 17·18e
Loibl Pass.....46·26n 14·16e
Mädelegabel, *mountain*.....47·18n 10·18e
March, *river*....48·10n 16·59e
Mühlviertel, *physical region*.48·25n 14·10e
Mur, *river*.....46·39n 16·02e
Neusiedler See, *lake*.....47·50n 16·46e
Niedere Tauern, *mountains*....47·18n 14·00e
Norische Alpen, *mountains*....46·55n 14·05e
Osttirol, *historic region*.....46·55n 12·30e
Ötscher, *mountains*.....47·52n 15·12e
Ötztaler Alpen, *mountains*.....46·45n 10·55e
Pinzgau, *valley*..47·15n 12·40e
Plöckenpass, *pass*.....46·36n 12·58e
Raab, *river*.....46·57n 16·15e
Reschenpass, *pass*.....46·50n 10·30e
Rhätikon, *mountains*.....47·03n 9·40e
Rhine (Rhein), *river*.....47·29n 9·39e
Salza, *river*.....47·40n 14·43e
Salzkammergut, *physical region*.47·45n 13·30e
Sauwald, *forest*.....48·28n 13·40e
Schwarza, *river*.47·43n 16·13e
Silvretta, *mountains*.....46·50n 10·10e
Speikkogel, *mountain*.....47·14n 15·03e
Steyr, *river*.....48·03n 14·25e
Thaya, *river*.....48·35n 16·57e
Totes Gebirge, *mountains*.....47·42n 13·55e
Traun, *river*.....48·09n 14·01e
Wienerwald, *mountains*.....48·10n 16·00e
Wildspitze, *peak*.46·53n 10·52e
Worther See, *lake*.....46·37n 14·10e
Ybbs, *river*.....48·10n 15·06e
Zillertaler Alpen, *mountains*.....47·00n 11·55e
Zirbitz Kogel, *mountain*.....47·04n 14·34e
Zuckerhütl, *mountain*.....46·58n 11·09e
Zugspitze, *mountain*.....47·25n 10·59e


AUSTRIA
Size of symbol indicates relative size of town
Elevations in metres
0 25 50 100 150 km
0 25 50 100 mi
© Rand McNally & Co.
A-550200-257
Nürnberg
Stuttgart
MUNICH
Zürich
MILAN
Venice
Ljubljana
Zagreb
Brno
Bratislava
VIENNA (WIEN)
Linz
Salzburg
Innsbruck
Graz
Klagenfurt
Villach
Leoben
Wiener Neustadt
Sankt Pölten
Steyr
Wels
Kufstein
Lienz
Spittal
Judenburg
Knittelfeld
Zell
Mallnitz
Sillian
Oberdrauburg
Lofer
Hallein
Bad Ischl
Gmunden
Mattighofen
Braunau
Ried
Schärding
Aigen
Freistadt
Pregarten
Aschach
Peuerbach
Neumarkt
Mauthausen
Grein
Perg
Enns
Amstetten
Waidhofen
Melk
Krems
Stockerau
Korneuburg
Klosterneuburg
Gmünd
Weitra
Zwettl
Horn
Eggenburg
Haugsdorf
Hollabrunn
Mistelbach
Hohenau
Dürnkrut
Marchegg
Gänserndorf
Bruck
Neusiedl
Eisenstadt
Rust
Neusiedler See
Mattersburg
Neunkirchen
Gloggnitz
Mürzzuschlag
Kapfenberg
Bruck
Eisenerz
Hieflau
Admont
Liezen
Rottenmann
Irdning
Trieben
Leoben
Donawitz
Leoben
Vordernberg
Eisenerz
Kindberg
Aspang Markt
Ober Pullendorf
Kirchschlag
Friedberg
Birkfeld
Oberwart
Hartberg
Fürstenfeld
Güssing
Jennersdorf
Weiz
Gleisdorf
Köflach
Voitsberg
Deutschlandsberg
Leibnitz
Radkersburg
Mureck
Wildon
Leibnitz
Feldbach
Kirchbach
Bad Sankt Leonhard
Stainz
Friesach
Wolfsberg
Eberstein
Sankt Paul
Eibiswald
Völkermarkt
Klagenfurt
Feldkirchen
Villach
Hermagor
Kötschach
Paternion
Obervellach
Gmünd
Millstatt
Sachsenburg
Winklern
Matrei
Heiligenblut
Lienz
Sillian
Oberdrauburg
Kufstein
Wörgl
Kitzbühel
Saalfelden am Steinernen Meer
Zell
Mittenwald
Seefeld
Solbad Hall
Wattens
Schwaz
Jenbach
Mayrhofen
Zell
Krimml
Bruck
Kaprun
Zell
Schwarzach
Badgastein
Bad Hofgastein
Tamsweg
Mauterndorf
Predlitz
Murau
Ramingstein
Neumarkt
Sankt Johann
Radstadt
Schladming
Bischofshofen
Sankt Johann
Irdning
Hallstatt
Abtenau
Bad Aussee
Ebensee
Gmunden
Vöcklabruck
Attnang
Kirchdorf
Grein
Waidhofen
Steyr
Ternberg
Weyer Markt
Sankt Gallen
Windischgarsten
Grosses Priel
Hochkönig
Dachstein
Hochgolling
Hochfeind
Hochalmspitze
Großglockner
Ankogel
Hochalmspitze
Großvenediger
Hohe Tauern
Hohe Warte
Pasterze
Pfunds
Nauders
Reschenpass
Obergurgl
Sölden
Zwieselstein
Ötztaler Alpen
Brenner Pass
Matrei
Steinach
Landeck
Prutz
Fulpmes
Lanersfeld
Schruns
Gaschurn
Silvretta
Piz Buin
Bludenz
Dornbirn
Bregenz
Feldkirch
Nasserreith
Imst
Telfs
Mittenwald
Wendelstein
Zugspitze
Waxenstein
Hochfeiler
Hochgall
Hochfeind
Hochgolling

Austria, Area and Population

	area		population	
	sq mi	sq km	1961 census	1971 census*
Bundesländer (States)				
Burgenland	1,530.94	3,965.15	271,000	273,000
Kärnten (Carinthia)	3,680.66	9,532.92	495,000	526,000
Niederösterreich (Lower Austria)	7,401.65	19,170.28	1,374,000	1,412,000
Oberösterreich (Upper Austria)	4,625.03	11,978.84	1,132,000	1,224,000
Salzburg	2,762.11	7,153.88	347,000	400,000
Steiermark (Styria)	6,326.62	16,385.96	1,138,000	1,192,000
Tirol (Tyrol)	4,883.19	12,647.47	463,000	539,000
Vorarlberg	1,004.38	2,601.34	226,000	275,000
Wien (Vienna)	160.05	414.53	1,628,000	1,603,000
Total Austria	32,374.63	83,850.37	7,074,000	7,444,000

*Preliminary figures.
Source: Official government figures.

and 99 percent of this total was German-speaking. The 1 percent belonging to other nationalities was made up of about 30,000 Croats and 11,000 Magyars living in Burgenland; 24,000 Slovenes living in the southern part of Kärnten; and some 5,000 Czechs living in Vienna. In terms of religion, the population is 89 percent Roman Catholic; 6 percent are Protestant; 0.5 percent are so-called Old Catholics; some 0.2 percent are the remainder of a once much bigger Jewish community; and 3.8 percent profess no religion at all.

The westernmost states, Vorarlberg in particular, are inhabited by Alemanic groups whose dialect is similar to the Swiss-German dialect, while the dialect of the people in Salzburg and Oberösterreich, and also the eastern states, resembles that spoken in Bavaria. People in Steiermark and Kärnten speak a dialect which is clearly distinguishable from that of their western fellow citizens. In general, the German spoken in Austria has a softer, more drawling, and melodious sound than that of Germany.

Demographic trends. The birth rate varies within the Austrian states: the yearly national average of live births was 15.1 per thousand at the start of the 1970s, but in the strongly Catholic western states and rural areas birth rates are higher, reaching a peak of 19.2 in Vorarlberg. In Steiermark, the birth rates rank above the national average, whereas in Vienna the rate is only 10.9.

There has also been a distinct population shift from the east to the west of Austria: Vienna, Niederösterreich, and Burgenland have suffered losses in population, while the western states have gained, during the decades following World War II. The most remarkable of these changes has taken place in the capital itself: Vienna had 1,918,600 inhabitants in 1923 but only 1,603,408 according to the 1971 census.

Population shifts

The almost complete destruction of Vienna's Jewish population (about 500,000 in 1938) during the Nazi era was only partly compensated by an influx from rural areas and by refugees after World War II. Increased urbanization, associated with the industrialization of some of the western states, is the main cause of the population growth in those areas.

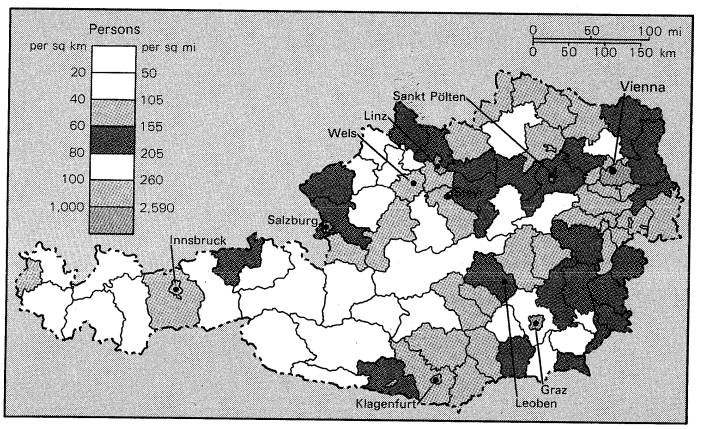
Emigration resulted in some losses in national population during the disastrous post-World War II years, although a stream of expellees from eastern European countries helped to make up some of the deficit. Most of these refugees were soon integrated into Austrian society. As a result of the turmoil in Hungary in 1956, some 180,000 Hungarians crossed the frontiers into Austria, creating a new refugee problem. By the end of the 1960s fewer than 9,000 remained in Austria, as most had migrated to other countries.

Life expectancy for Austrian citizens has been rising constantly since the beginning of the century. Around 1900 a boy born in Austria could expect to live about 39 years: by 1970, his life expectancy had risen to about 67 years. The life expectancy of women rose even higher, from 41 to about 73 years over the same period.

With a birth rate increasing as a result of growing prosperity and rising life expectancy, a continued increase in Austria's total population can be expected, unless economic conditions change for the worse. Further urbanization is also probable: in the decades since World War II the number of persons engaged in forestry and agriculture declined as more jobs became available in industry, and the trend is continuing. The associated depopulation of rural areas has been highest in Kärnten, Burgenland, and in certain districts of Niederösterreich. A decline in agricultural population has also been noted in Oberösterreich, where steel and aluminum have, from the 1950s onward, become important industries. The same process has taken place in the westernmost parts of Austria, where the textile industry has expanded. Tourism also attracts rural manpower. The loss to agricultural population has been smallest in the Tirol and Salzburg areas where there has been less industrialization.

Rural depopulation

In general, the future development of Austria's population will depend chiefly on further industrialization and the growth of tourism. Due to domestic labour shortages some workers have been imported from abroad, chiefly from Yugoslavia and, to a lesser degree, from Italy, Spain, and Turkey; but to a smaller extent than, for example, in West Germany.



Population density of Austria.

THE NATIONAL ECONOMY

The Austrian economy has been shaped by two factors: firstly, the nation occupies a mountainous country, only half the area of which is even potentially usable for food production; secondly, and as a result, the success of industrial production, exports, and trade is basic to the viability of the entire economy. The natural resources available within the country for industrial exploitations are, therefore, of considerable significance.

The extent and use of resources. *Mineral resources.* Austria is the world's leading producer of magnesite, a natural magnesium carbonate used extensively in the chemical and other industries, with Kärnten being the main centre of production. In recent decades, new methods of smelting and processing magnesite have been developed, and production, by the 1970s, exceeded 1,500,000 tons annually.

Iron ore deposits are found in the Erzberg in Steiermark—which produces 75 percent of Austria's total iron production—and in the Hüttenberg region in Kärnten. Resources have been estimated at 300,000,000 tons, and the ore has an iron content of some 30 percent. Iron production reached a record high of almost 4,000,000 tons annually by the start of the 1970s. During the same period, Austria's oil production, almost 3,000,000 tons of crude oil per year, ranked third in Europe (after Romania and West Germany), a remarkable achievement for an industry of comparatively recent origin (it was set up in 1927). Potential oil sources, chiefly located in Niederösterreich, are estimated at about 60,000,000 tons, with vast additional natural gas resources.

Developing oil production

Coal, found only in small quantities and mostly of the soft variety, is mined chiefly in Steiermark, Kärnten, and Burgenland. Over 200,000 tons of salt are also produced annually.

Hydroelectric and other power resources. Austria is one of Europe's foremost producers of hydroelectric power: the nation's total potential capacity is estimated to be at least 40,000,000,000 kilowatts per year. A large portion of the hydroelectric power resources were put to use only in the years following World War II. By the 1970s, a national network maintained a steady power-production capacity even in the winter months, when river levels are low.

Sources of national income. Industry and trade form the most important sources of national income. The other main contributors to the national economy are public service, construction, and manufacturing.

Since barely half of Austria's territory can be used for food production, an intense program of agriculture modernization was introduced after World War II. The number of tractors, for example, rose from 1,800 in 1939 to 260,000 thirty years later. The rural economy is dominated by smallholdings, although some large forest estates still remain. As a result of the modernization program, the yield of the most important agricultural products has risen considerably.

In the lowland regions of the country, the abundance of pastures and the production of feed has promoted cattle breeding. Dairy products are plentiful, resulting in an overproduction in terms of national need. With the exception of horses, Austria's livestock has grown in numbers over the period of agricultural modernization.

Austrian food consumption rose to an average daily intake of almost 3,000 calories by the 1970s. The modernized domestic food industry was, nevertheless, able to supply 83 percent of the national requirements.

Timber from the vast national forest cover is only partly processed within the frontiers of Austria, and untreated lumber figures prominently in the national export statistics. More than half of the forests are in the hands of smallholders, the rest partly in large, centuries-old estates and partly government-owned.

Manufacturing. Manufacturing has also strengthened its position in recent decades. Iron and steel production, in particular, increased greatly after World War II: pig iron production, for example, rose from under 900,000 tons in 1950 to almost 3,000,000 tons twenty years later, while raw steel output grew from 950,000 to almost 4,000,000 tons in the same period. Despite an abundance of domestic iron ore, scrap iron and iron ore had to be imported, due to increased steel production. A newly developed oxygen blast furnace, based on the so-called L.D.-process (called after the cities of Linz and Donawitz [now Leoben], where it was developed), succeeded in reducing the need for imported scrap iron, and is now used in many countries under license agreements.

The iron and steel industry

Aluminum has recently become one of the most important Austrian manufacturing industries, and the Ranshofen works, in Oberösterreich, was among the largest in Europe by the 1970s. Aluminum output has risen in a spectacular manner, from 1,100 tons in 1946 to over 90,000 tons in 1960, and almost 125,000 tons by 1970.

Other manufacturing includes the paper industry, based on the nation's extensive forest reserves: paper, one of Austria's largest foreign exchange earners, is exported to over 80 European and overseas countries. The chemical industry is dominated by two giant chemical plants—a nitrogen works in Linz and the Lenzing staple fibres factory. Plastics, a more modern industrial product, were being produced in 100 large, and over 300 smaller, plants by the 1970s.

Financial services. Monetary policy is jointly determined by the Ministry of Finance and the National Bank of Austria. Although the authorities grant a large measure of freedom to those individuals and institutions engaged in international transactions, as much as 38 percent of the total gross fixed investment in the country, including housing, is controlled by the federal government.

The Austrian National Bank operates under a Bank Law of 1955, as amended in 1969; this law changed both ownership and management of this important institution: the government now holds 50 percent of the shares, with the balance being allotted by the government to individuals or economic organizations.

Financial affairs are also executed through a wide range of other institutions: joint-stock commercial banks; private commercial banks; provincial mortgage organizations; saving banks; and agricultural, as well as industrial, credit cooperatives.

Foreign trade. Austria's foreign trade steadily increased in the decades following World War II.

Austria's main trading partners are her neighbours, West Germany and Switzerland, who between them account for almost half of Austria's imports and one-third of its exports. Trade with the European Economic Community (Common Market) as a whole is also important: at the start of the 1970s, its members provided 56 percent of Austria's imports and received 39.4 percent of its exports, whereas only 19.1 percent of Austria's imports and 25.3 percent of its exports came from dealings with countries of the European Free Trade Association (EFTA). Trade with eastern European countries amounted to an even smaller proportion: only 9.3 percent of Austria's imports and 12.9 percent of its exports. Timber, iron, and steel are the most important items in the export trade, accounting together for over a quarter of the total.

The Common Market's important role in Austria's foreign trade has prompted the national authorities to seek a "special arrangement" with that organization (full membership would not be compatible with Austria's policy of neutrality). Protracted negotiations were concluded by an agreement in July 1972.

The tourist trade is Austria's outstanding invisible export. The number of nights spent by foreign tourists in Austria rose from 6,000,000 in 1951–52 to over 50,000,000 some 20 years later, and the amount of foreign currency spent showed an even greater increase. The outlook for tourism continued to be healthy in the 1970s.

Management of the economy. The private sector, comprising roughly three-quarters of Austria's mixed economy, is concentrated in agriculture and food processing; forestry and timber; paper mills; textile and clothing; food and beverages; the retail and wholesale trade; and department stores. Private enterprise has, as yet, a relatively low degree of monopolistic concentration, although ties among the guildlike cooperatives (Berufsgenossenschaften und Innungen) are not without influence.

Nationalization of industry

In 1946, the Austrian parliament nationalized a large segment of Austrian industry which, at that time, was being held under Soviet control as alleged former German property. After the State Treaty of 1955 established Austrian neutrality and brought about the end of Soviet authority, all these enterprises passed into exclusive Austrian control. The 1946 nationalization covered the three biggest banks and some 70 larger industrial enterprises, chiefly in the fields of iron and steel, aluminum, and machinery. Subsequent reorganization reduced the number of nationalized enterprises to 19. By the 1970s, when Austrian industry as a whole employed over 850,000 persons, over 100,000 worked in nationalized undertakings, and the net value of their output reached about 25 percent of the total industrial production.

Most nationalized industries are organized as joint-stock corporations, and their central direction by the government has undergone several changes, for both political and managerial reasons. Until 1966, for example, a cabinet member supervised the nationalized industries, but after that date they were handed over to the government-owned Austrian Industrial Administration (Österreichische Industrie–Aktian–Gesellschaft—OIAG). Its board of directors, although representing the government, is independent as far as managerial decisions are concerned.

The tax income of the federal government amounted to over 80,000,000,000 schillings annually by the 1970s. The most important source of federal tax revenue was the income tax, which has a progressive scale reaching up to almost 70 percent of taxable income. Corporation taxes

take from 30 percent to 60 percent of corporation profits, and the capital gains tax is fixed at just under 20 percent. The second biggest source of tax income is a sales tax of 5.5 percent on all sales and transactions, with a lower rate of 2 percent for privileged wholesale trade, while a special trade tax also provides for a 2 percent payroll tax. In 1970 the government started preparations for a major tax reform.

Contemporary economic policies, problems, and prospects. Full employment, stability of prices and currency, a certain government influence on investment and economic growth, and the distribution of national income for conscious social goals, without impairing a free-market, consumer-oriented economy, were taken to be the main aims of Austrian economic policies by the 1970s. To avoid price increases, chiefly of imports from West Germany, the Austrian schilling was up-valued by 5 percent during the international currency crisis of May 1971.

Maintaining or expanding foreign trade, particularly with the Common Market countries; securing price stability in a generally inflationary era; and ending overproduction in certain branches of agriculture were some of Austria's problems in the same period. The booming economy, full employment, a growing tourist trade, and modernization of an expanding industry and agriculture were considered as hopeful and healthy signs in the early 1970s.

TRANSPORTATION

Austria's location in the centre of Europe determines its role in European road, rail, air, and river traffic, whether flow is from north to south or from east to west. Austria is consequently not only a tourist country but in many aspects also may be thought of as a freight clearinghouse, supplying the great trade routes running across the Alps and along the Danube River.

Roads. The road system, heavily damaged during World War II, was subsequently adapted to vastly increased traffic requirements. A federally built east-west highway was in operation between Vienna and Salzburg by the early 1960s. It will eventually lead to the western frontier. A north-south autobahn, or federal highway, ultimately to run from Vienna to the Italian frontier, has been started. The total Austrian road network excluding city streets, much of which leads over spectacular Alpine passes, covers some 20,000 miles. The total number of motor vehicles has shown a remarkable increase in recent decades, rising from some 285,000 units in 1950 to 2,290,000 some 20 years later.

Railroads. Forty-one percent of the Austrian railroad network and 381 bridges were destroyed during World War II and had to be repaired as well as modernized in subsequent years. By the early 1970s, nearly half of the rail network, carrying 80 percent or more of the traffic, had been converted to electric traction. Passenger and freight traffic has shown a proportionate increase. The Austrian railroads are state-owned, although, by a law of 1969, the railroad administration, which had formerly been part of the Transportation Ministry, became an independent commercial enterprise.

Importance of the Danube

Water transport. The Danube (*q.v.*) is the most important river connection between Germany and the Black Sea, and the federally owned Austrian Danube Steamship Company plays an important role in both freight and passenger traffic along this waterway. Although Austria is landlocked, Austrian shipyards, owned by the Steamship Company, build vessels not only for Austria but also for Bulgaria, Yugoslavia, Pakistan, Egypt, and Nigeria.

Air transport. Austrian Airlines, which began operations in March 1958, has since established service to 33 airports in 21 countries of Europe and the Middle East. Austria's main airport is at Wien-Schwechat, southeast of Vienna.

ADMINISTRATIVE AND SOCIAL CONDITIONS

The structure of the government. *Constitutional framework.* Under the constitution of 1920—with minor changes made in 1929—Austria is a ". . . democratic republic: its power derives from the people. . . ." A federal state, Austria consists of nine self-governing states: Burgenland, Kärnten, Niederösterreich, Oberösterreich, Salzburg, Steiermark, Tirol, Vorarlberg, and Wien (Vienna).

In 1934 the Austrian constitution was replaced by an authoritarian regime under chancellors Engelbert Dollfuss and Kurt von Schuschnigg. This, in turn, was eliminated by Hitler after Nazi Germany annexed Austria in 1938. With the liberation of Austria in 1945, the constitution of 1929 was revived, and subsequently became the foundation stone of constitutional and political life in the "Second Republic."

The federal president and the cabinet share the executive authority. The president is elected by popular vote for a term of six years. He acts as head of state, appoints the cabinet, and calls parliament into session. He can dissolve parliament during the four-year legislative period, unless it dissolves itself by law, and can order new elections. The president also acts as commander-in-chief of the armed forces.

The president appoints the federal chancellor and, at his suggestion, the other cabinet members. The cabinet cannot remain in office if it and its members do not enjoy the confidence of the majority of the National Council.

The Austrian parliament

The parliament consists of two houses: the National Council (Nationalrat), wielding the primary legislative power; and the Federal Council (Bundesrat), representing the states. The National Council, which, since 1920, had had 165 members, was expanded to 183 members under a law passed in November 1970 which took effect at a national election in October 1971. The National Council is elected by all citizens over the age of 20, and every citizen over the age of 26 is eligible to run for office. The distribution of parliamentary seats is based on a system of proportional representation.

The members of the Federal Council represent the states, and the state assemblies, or diets, elect the members by a proportional system based on the population of the state.

The legislative process originates in the National Council. Each bill—except for the budget, which is the sole prerogative of the National Council—must be approved by the Federal Council. The National Council is nevertheless empowered to override a Federal Council veto by a simple majority vote.

Each of the nine states is administered by a government headed by a governor elected by the legislative diet, elected by general and equal ballot.

Local municipalities elect their mayors and city councils. Vienna is a unique case: as both municipality and state, its mayor also functions as a governor.

The political process. The first popular election of a president, although provided for by the 1929 amendment to the constitution, did not take place until after the death of the first post-World War II president, Karl Renner (1870–1950), who had been unanimously elected by the National Assembly after the liberation of 1945 and the ensuing unsettled political circumstances. A succession of Socialist presidents, starting with the former mayor of Vienna, General Theodor Körner, has occupied the post since.

In the eight elections held from the immediate post-World War II period to the early 1970s, the stability of the two main Austrian political parties—the People's Party and the Socialist Party—is shown by their respective shares of the vote, which has fluctuated, in each case, around 45 percent of the total votes cast. The Liberal Party has polled an average of but 5 percent of the votes, and the Communist Party has polled an even smaller share. From 1945 until 1966 the two main parties maintained coalition governments with the chancellorships held by the People's Party. In 1966 the latter, with a slim majority of parliamentary seats, formed the government alone, but after the 1970 elections, in which the People's Party lost its majority, the Socialists, as the strongest party, formed a minority government as the first of the new decade. Its first budget was passed with the help of the votes of the Liberal Party. In the parliamentary election in October 1971 the Socialist Party won more than

50 percent of the popular vote, the first party to do so in Austrian history.

The major parties

The People's Party is the successor of the Christian Social Party, founded in the 1890s. It represents a combination of conservative forces and various social and economic groups forming semi-independent federations within the overall party. The divergencies of economic and social interests exhibited by these groups—the farmers, the businessmen, blue- and white-collar workers—necessitate policy compromises, which are not always easy to implement.

The Socialist Party, which followed a democratic Marxist program during the 1918–34 period, has adopted a more pragmatic approach since 1945. The new party program, adopted unanimously in 1958, stressed these new tendencies. Emphasis was placed on social problems; on government influence on an expanding and socially oriented economy; on full employment; and on a rising standard of living, with a more equal distribution of wealth. The Austrian Socialists are no longer an exclusive party of the working class: by the 1970s they had a much wider appeal.

The Liberal Party has undergone several changes in emphasis since being admitted into the political arena by the occupying powers in 1949, at which point it could be characterized as a third party with nationalistic tendencies. The initially prominent influence of former Nazis has diminished.

The Communist Party, always negligible in Austrian politics, reached its high point during the Soviet occupation of the eastern zone of the country, when it attained 5 percent of the popular vote. By 1970, it had lost representation in parliament, in all of the provincial diets, and most municipal councils.

The participation of the citizens. The Austrian constitution provides for two kinds of popular initiatives: one is the so-called popular demand (*Volksbegehren*) by which 400,000 vote-eligible citizens can petition parliament for approval of any bill; it can also be initiated, in the case of any bill or proposal, by a majority of the National Council. A total revision of the constitution must be approved by plebiscite.

Justice. The administration of justice is independent of Austrian legislative and administrative authorities. Judges are not subject to any government influence: they are appointed by the cabinet upon nomination by judicial panels, and can neither be dismissed nor transferred without their agreement.

Austria has three high courts: one sitting as the highest body of appeal in civil and criminal matters; a top administrative court to which citizens aggrieved by administrative decisions can appeal; and a constitutional court that decides on all constitutional matters, civil rights, and election disputes.

The armed forces. Under the State Treaty of 1955 Austria is permitted armed forces of 50,000 men for the defense of its territory. The armaments permitted Austria specifically exclude all atomic weapons and rockets. Only the four signatory powers—the United States, Great Britain, the Soviet Union, and France—can permit such armaments, and they must do so by a unanimous decision.

All male citizens between the ages of 18 and 50 are liable to military service, and conscripts are called up for a normal tour of duty of six months. Volunteers may serve for 12 or 15 months, or may enlist for careers up to 12 years. Expenses for the armed forces amount to about 4.1 percent of the national budget.

Services. *Educational services.* Austria's educational system is based on obligatory school attendance between the ages of 6 and 14, with a ninth school year provided under certain circumstances. A major reform of the school administrative structure, providing for a unitary school system with access to higher education, was in preparation in the early 1970s.

Intermediate schools include those preparing for university and other higher studies, teachers' and commercial colleges, and other specialized institutions.

Austrian universities

Austrian universities are among the oldest German-speaking universities of Europe: the University of Vienna, for example, was founded in 1365. Austria entered the 1970s with 16 universities, including two technical universities and academies for fine arts, music, and drama. Austria also has a huge adult-education system, with 350 centres and an additional 1,900 dependent local institutions.

Health and welfare services. Public health in Austria is the responsibility of the Health Department within the Ministry of Social Administration, which supervises 14 subsidiary institutes responsible for the prevention of infectious diseases and inspection of drugs and food. The provincial governments also have public-health centres, and each municipality must employ a public-health physician.

National-health insurance covers expenses of medical and hospital treatment: blue- and white-collar workers and salaried employees are protected in cases of sickness, disability, unemployment, maternity, and in their old age. There are also survivors' pensions. Pension systems for self-employed persons and farmers have also been established.

Housing. After the reconstruction of war-damaged houses, a building boom started in 1950. Between 1951 and 1961, 22 percent of all apartment dwellings were newly built, according to official statistics; from 1961–69 the figure was 16 percent. In the western states the increase was even higher.

Forty-four percent of all dwellings are one- or two-family houses. In the capital, only 6.6 percent of the population live in such houses; large apartment houses, many of them built by the city or with public assistance, predominate. Vienna, since the middle 1920s, has had a record of intense apartment building by the city, and some 20 percent of all dwellings are municipally owned. Nationally, two-thirds of all Austrian homes are privately owned; 26 percent of all apartments are one-room and 34 percent are two-room.

Police services. Austria's main police force is federal. By the 1970s it comprised almost 15,000 persons, about 10,000 of whom were in uniform, with a further 2,000 plainclothesmen. The police are under the jurisdiction of the federal minister of the interior.

Austria also has about 10,000 "gendarmes" responsible for public safety in rural areas and also attending to certain administrative functions.

Social conditions. *Wages and cost of living.* During the later years of the 1960s, the consumer price index in Austria rose by about 10 percent, while, during the same period, the net weekly income of workers rose by some 20 percent. Living standards have improved during the 1960s as a whole, but particularly since 1966, although Austria's wages still lag behind those of neighbouring industrial countries.

Cooperative wage-price policy

Government, management, and labour have followed a wage-price policy that has attempted to avoid social cleavages and strikes through cooperative participation in a joint wage-price commission. The representatives of the various segments of the economy, the chambers of commerce, of agriculture, and of labour, together with the federation of trade unions, the association of industry, and the farmers' federation, have tried, with the active cooperation of the government, to coordinate wage and price movements. It is within this framework that collective bargaining takes place, and agricultural prices are also negotiated by the wage-price commission without infringement of the market economy.

Social and economic divisions. The three important economic groups in Austria—labour, management, and the farmers—have similar structures: each has its own independent organization: the trade unions, the management association, and the farmers' federation. At the same time, laws provide for semi-official "chambers" for each grouping. This type of guild organization promotes cooperation in the governmental wage-price commission. Despite the divergent interests of the various groups, their cooperation has resulted in relative economic stability, and labour-management relations have remained unmarked by major crises.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. The contemporary cultural milieu of Austria has a rich heritage: in architecture and poetry this goes back to the Middle Ages, and in medicine and science it can be traced to the 18th and 19th centuries. Similarly, Vienna's art galleries are among the most famous in Europe because of their wealth of old Dutch masters. Austria's most highly recognized cultural contribution has been in the field of music, and this tradition still dominates its present cultural achievements. Austrian cultural life has exerted attraction beyond its borders. Such great musicians as Beethoven (a native of Bonn), Brahms (from Hamburg), Mahler (from Bohemia) and, to a certain extent, Richard Strauss (from Munich) are considered as much a part of Austria's musical life as the native Austrian giants Haydn, Mozart, and Schubert. The Vienna waltzes emanating from the Johann Strauss dynasty, as well as the Vienna operetta, whose presiding genius was Franz Léhar, round out the rich traditions of Viennese musical life. Austria has also been the birthplace of modern and especially 12-tone music.

Musical tradition

Although in literature Austria has often been considered a segment of German culture, writers such as Franz Grillparzer, Johann Nestroy, and Ferdinand Raimund in the postclassic era, and Hugo von Hofmannsthal and Arthur Schnitzler in the 20th century, have all developed special Austrian traits. Austrian culture was also strongly influenced by the various cultural elements of the multinational Austro-Hungarian Empire, with its Slavic and Magyar traces. They still exist, making the Austria of today a bridge between various parts of Europe and attracting performing artists from East and West.

Austrian cultural and scientific life had to recover from the severe wounds inflicted by the Nazi domination and the loss of intellectuals and artists to emigration and persecution after 1938. A ten-year occupation by four great powers promoted recovery of an independent cultural life only to a small degree. Austria has, nevertheless, managed—particularly since the later 1950s—to regain something of its past cultural posture.

The state of the various arts. Austrian expressionism in literature has gained world renown through the visionary novelist Franz Kafka, who hailed from Prague (where he was born in 1883) and who died near Vienna in 1924. As part of a general Kafka revival, one of his haunting novels, *Der Prozess* (*The Trial*), was made into an opera by Austria's best-known contemporary composer, Gottfried von Einem.

Georg Trakl, a promising Expressionist who was killed in the war in 1914; Franz Werfel, an Expressionist novelist and playwright (born in 1890, died as refugee in the United States in 1945); Karl Kraus, a critical lyricist and essayist; and Robert Musil, who wrote along Expressionist lines, have all played a considerable role in contemporary Austria. Hermann Broch, a novelist with symbolic tendencies (1886–1951, also died in the United States) is also recognized beyond Austria. Heimito von Doderer (1896–1966) is considered the most remarkable post-World War II novelist. Modern playwrights who are well known include Alexander Lernet-Holenia (born 1897) and Fritz Hochwälder (born 1911), who has created historical dramas that have been produced all over the world.

Austrian theatre and opera

The Austrian theatre and opera recovered quickly after the war. The Vienna Burgtheater, considered to be the best German-speaking theatre in the 19th century, may again be counted among the foremost German stages. It presents the classics, as well as modern English, American, and occasionally Russian and Czech, plays. A new trend in the Vienna theatre has developed in so-called cellar-theatres ("off-Broadway" stages) that favour the theatre of the grotesque.

The Vienna Staatsoper (State Opera), which was completely rebuilt after the war, ranks today with La Scala in Milan and the Hamburg and Munich operas, while the Vienna Philharmonic Orchestra has played in almost all the musical capitals of the world. Austrian theatre life is intense, with large and enthusiastic audiences and a bewildering succession of shows. The theatre has spread to small stages in the Vienna suburbs, while state capitals have also become theatre and light-opera centres, developing their own identities. Special performances for students aid an early participation in cultural life.

Modern music was born in Vienna, and Arnold Schoenberg, Alban Berg, and Anton von Webern are considered among its major creative founders.

Clemens Holzmeister, the best-known contemporary Austrian architect, was responsible for the two festival theatres in Salzburg, which combine a classic and modern theatre style. Artur Perotto, of Linz, has designed several striking skyscrapers in western Austria, while in Vienna, public apartment building has had a strong overall influence on architecture.

In sculpture, Fritz Wotruba is an internationally recognized figure. His style is often compared with that of the leading English sculptor, Henry Moore.

Austria's most famous contemporary painter is Oskar Kokoschka, truly a world figure: he and Alfred Kubin can be regarded as two of the foremost creators of modern painting in Austria. In addition, a young group of Surrealists has shown its works in international exhibitions.

Folk art and folk traditions, supported by provincial governments, have survived in western Austria, especially in Tirol.

The oldest Austrian academic research institution is the Academy of Science, whose traditions reach back to the end of the Middle Ages. More modern scientific foundations, notably the Körner Foundation and the Renner Foundation, support scientific research and other cultural endeavours: their main support comes from government sources. A new federal Ministry of Science was established in 1970 with the aim of reviving or increasing scientific and cultural activities.

Press and broadcasting. By the 1970s, the Austrian press had undergone far-reaching changes. These were particularly marked at the end of the 1960s, when the so-called Boulevard-press, offering light reading material and pictures, and occupying an intermediate status between the "quality" and "yellow" press, made extensive headway.

A second development has been the demise of the party-political press. Whereas the Socialist Party, with considerable financial effort, has been able to maintain its rather elderly *Arbeiter-Zeitung* (founded 1889), the People's Party was forced to cease publication of its organ in Vienna in November 1970.

The only Viennese paper with a conservative slant that tries to project a worldwide picture, *Die Presse*—namesake of one of the best-known papers prior to World War I—survives with the help of the federation of business.

Another change has been brought about by the shift of population and a measure of social life to the states: the latter have thus developed their own newspapers.

Growing popularity of television

The main reason for the changes that have taken place in the newspaper field has been the growing popularity of television. Austrian networks are state-owned but controlled by the Austrian Broadcasting Company, an entity which came into being in 1957, following a popular initiative leading to a legal reorganization. The company is established along the lines of an independent private enterprise, with a capital of 115,000,000 schillings, 112,800,000 of which are contributed by the federal authorities and 2,200,000 by the states. The Austrian radio network is made up of some 300 stations (including relay stations), with the most important being located in provincial capitals: they broadcast three different programs.

Regular television service in Austria started on January 1, 1957; in 1971, Austria had over 120 television stations. Commercials are strictly limited.

THE FUTURE

In a difficult period of slow adjustment from its status as the century-old centre of a multinational empire to a small neutral nation, Austria has experienced a crippling depression in the early 1930s; two civil wars in 1934; an authoritarian regime until 1938; a Nazi regime and

World War II; and a ten-year occupation by four great powers during the Cold War era. Yet Austria has emerged as a stable country whose neutrality is challenged neither by the super-powers, nor by its various neighbours.

This status has been, in part, the result of, and, also in part, the stimulus behind, recent favourable economic and social developments. Only since 1955 has Austria been able to develop, effectively and independently, its natural resources—chiefly oil and hydroelectric power—and also to make more efficient use of its iron ore and magnesite deposits. Agriculture, as a result of modernization, is now able to cater to most of the country's food requirements. Flexibility in adjusting foreign trade to the changed commercial and political environment, together with the great expansion of the tourist trade as a foreign-currency earner, have also contributed to the country's prosperity. All these factors, combined with a progressive social policy, have caused the resurgence of a new and distinctive Austrian national consciousness.

BIBLIOGRAPHY. FEDERAL PRESS SERVICE IN VIENNA, *Austria: Facts and Figures*, 8th ed. (1970); and CENTRAL AUSTRIAN STATISTICAL OFFICE, *Introducing Austria* (1968), are publications giving comprehensive statistical, as well as other, information on the various fields of economic, and, to a certain extent, cultural developments in Austria since 1945. PRESS SERVICE OF THE AUSTRIAN GOVERNMENT, *Austria: Land of Music* (n.d.); CULTURAL NEWS FROM AUSTRIA, *Austrian Short Story in the Twentieth Century* (n.d.), and *The Modern Austrian Drama* (n.d.), deal with various fields of Austrian cultural life. KURT SKALNIK, *Die Oesterreichische Presse* (1964), is a survey of the Austrian press, giving an objective description of the far-reaching changes in the make-up of the Austrian information media. JACQUES HANNAK (ed.), *Bestandaufnahme Oesterreich 1945–1963* (1963); C.T. GRAYSON, *Austria's International Position, 1938–53: The Re-establishment of an Independent Austria* (1953); RICHARD HISCOCKS, *The Rebirth of Austria* (1953); and ALLARD SVEN, *Diplomat in Wien* (1965), review Austria's immediate post-war development as well as the first period of independence following the conclusion of the State Treaty. HEINRICH SIEGLER, *Austria: Problems and Achievements Since 1945* (1967), provides a comprehensive documentary of Austria's political and diplomatic history after 1945. E.E. BAUMANN, *Crossroads of European Art: A Concise History of Art and Architecture in Austria* (1964), is an extensive history of Austria's cultural and art history. For statistical information, see *Oesterreichisches Jahrbuch 1969* (1970); *Wirtschaftsstatistisches Handbuch (Jahrbuch), 1968* (1969); and BUCHVERLAG SCHMIDINGER, *Austrian by Origin* (in six languages), 2 vol. (1969). See also the recent work by KARL R. STADLER, *Austria* (1971).

(O.L.)

Austria, History of

This article begins with a brief history of the area of the Austrian Republic from ancient times; it covers in greater detail the emergence of Austria as a German principality in medieval time, and the history of the Austrian-centred Habsburg state until 1918; it concludes with the history of the Austrian Republic. The article is divided into the following sections:

I. Austria to 1740
- Prehistory and Roman times
- Early Middle Ages
  - Germanic and Slavic settlement
  - The Babenberg period
- Late Middle Ages
  - The contest for the Babenberg heritage
  - The accession of the Habsburgs
  - The division of the Habsburg lands
  - The Burgundian and Spanish marriages
- Reformation and Counter-Reformation
  - The acquisition of Bohemia
  - The advance of Protestantism
  - Rudolf II and Matthias
  - The Bohemian rising and the victory of the Counter-Reformation
  - The struggle with Sweden and France
- Austria as a great power
  - The War of the Spanish Succession
  - The problem of the Austrian succession
  - New conflicts with Turkey and the Bourbons
  - Social, economic, and cultural trends in the Baroque age

II. The period 1740–1866
- From the accession of Maria Theresa to the Congress of Vienna
  - The war period, 1740–63
  - Foreign policy, 1763–92
  - The struggle with France, 1792–1815
  - Reforms and their reversal, 1740–1815
- The age of Metternich, 1815–48
  - Foreign and domestic policy
- Revolution and counter-revolution, 1848–59
  - The revolutions of 1848–49
  - The neoabsolutist era, 1849–60
  - Exclusion from Germany and Italy
- The transition to constitutional government, 1860–66

III. Austria-Hungary and the republics of Austria, since 1867
- Austria-Hungary, 1867–1918
  - The liberal ascendancy
  - National conflict and reform
  - Foreign policy, 1878–1908
  - The last years of peace
  - World War I
  - The end of the Habsburg Empire
- The First Republic and the Anschluss
  - The war's aftermath
  - Authoritarianism: Dollfuss and Schuschnigg
  - Anschluss and World War II
- The Second Republic
  - The Allied occupation
  - Restoration of sovereignty

I. Austria to 1740

PREHISTORY AND ROMAN TIMES

In the territories of the Austrian Republic the first traces of human settlement date back to the Early Paleolithic Period. The archaeological material becomes richer and more varied for subsequent periods, giving evidence of several distinct cultures succeeding one another or co-existing. The Austrian site of Hallstatt gave its name to the principal culture of the Early Iron Age (*c.* 800–450 BC). Celtic tribes invaded the eastern Alps around 400 BC and eventually founded the kingdom of Noricum, the first "state" on Austrian territory known by name. In the west, however, the ancient race of Raetians was able to maintain its seat. Then, attracted by the rich iron resources and the strategic importance of the region, the Romans began to assert themselves. After an initially peaceful penetration during the last two centuries BC, Roman troops finally occupied the country in *c.* 15 BC, and the lands as far as the Danube became part of the Roman Empire, being allotted to the Roman provinces of Raetia, Noricum, and Pannonia.

Occupation by Roman forces

The Romans opened up the country by an extensive system of roads. Among the Roman towns along the Danube, Carnuntum (near Hainburg) took precedence over Vindobona (Vienna), while Lauriacum (Lorch, near the confluence of the Enns and the Danube) belonged to a later period. Roman municipalities (*municipia*) also grew up at Brigantium (Bregenz), Juvavum (Salzburg), Ovilava (Wels), Virunum (near Klagenfurt), Teurnia (near Spittal an der Drau), and Flavia Solva (near Leibnitz). North of the Danube the Germanic tribes of the Naristi, Marcomanni, and Quadi settled. Their invasions in AD 166–180 arrested the peaceful development of the provinces, and even after their repulse by the emperor Marcus Aurelius the country could not regain its former prosperity. In the 3rd century the Roman frontier defenses began to be hard pressed by invasions from the Alemanni. Finally, in the 5th century, heavy attacks by the Huns and eastern Germans put an end to the Roman provincial defense system on the Danube.

There is archaeological evidence of a Christian cult in this area from the 4th century, and the biography of St. Severinus by Eugippius constitutes a unique literary source for the dramatic events of the second half of the 5th century. At that time several Germanic tribes (Rugii, Goths, Heruli, and later Langobardi) settled on Austrian territory. In the year 488 part of the harassed Norican population withdrew to Italy.

EARLY MIDDLE AGES

Germanic and Slavic settlement. After the departure of the Langobardi to Italy (568), further development

was determined by the Bavarians in a struggle with the Slavs, who were invading from the east, and by the Alemanni, who settled in what is now Vorarlberg. The Bavarians were under the political influence of the Franks, whereas the Slavs had Avar rulers. At the time of their greatest expansion the Slavs had penetrated as far as Styria, Carinthia, and eastern Tirol. After 624 the western Slavs rose against the Avars under the leadership of the Frankish merchant Samo, whose short-lived rule may also have extended over the territories of the eastern Alps. Around 700 the Bavarian lands again bordered on Avar territory, with the lower course of the Enns forming the approximate frontier. On the death of the Frankish king Dagobert I (639) the Bavarian dukes from the house of Agilolfing became virtually independent.

Christianity had survived only here and there among the remnants of the Roman population, when around 600 and then again around 700 Christian missionaries from the west became active, with the support of the Bavarian dukes. At the end of the 7th century St. Rupert, who came from the Rhine, founded the church of Salzburg. When they were threatened once more by the Avars, the Alpine Slavs (Karantaner) placed themselves (before 750) under the protection of the Bavarians, whose mission was extended to them. At the same time, Bavarian settlers penetrated into the valleys of Carinthia and Styria. Charlemagne, emperor of the neighbouring Franks, however, defeated the Bavarian duke Tassilo III, wiping out the Bavarian dukedom for a century. During the following years (791–796) Charlemagne also led a number of attacks against the Avars and destroyed their dominion. Surviving Avars were made to settle in the eastern part of Lower Austria between the rivers of Fischa and Leitha, where they soon disappeared from history, most probably mixing with the native population.

Frankish domination in the 8th and 9th centuries

As was the usual Frankish practice, border provinces (Marken, or marches) were instituted in the newly won southeastern territories. The Avar March on the Danube and Lower and Upper Pannonia and Karantania were to form a border fortification; but this soon became less effective because of frequent disagreements among the nobility. To that unrest was added a threat from the Bulgarians and from the rulers of "Great Moravia." Nevertheless, the process of Germanization and Christianization went on, in the course of which the churches of Salzburg and Passau came into conflict with the eastern mission led by the Slav apostles Cyril and Methodius. The Frankish kingdom richly endowed the church and nobility with new lands, which were settled by Bavarian and Frankish farmers.

In 881 the beginning of incursions by the Magyars led to a first clash near Vienna. By 906, they had destroyed greater Moravia, and, in 907, near Pressburg (Bratislava), the Magyars defeated a large Bavarian army that had tried to win back lost territory. Liutpold of Bavaria and Theotmar, the archbishop of Salzburg, were killed in battle. The Lower Austrian territories as far as the Enns River and Styria as far as the Koralpe fell under Magyar domination. Nevertheless, a certain continuity of German–Slav settlement was maintained, so that, after the victory of the German king Otto I (955) and the further repulse of the Magyars in the 960s, a fresh start could be made.

The Babenberg period. The first mention of a ruler in the regained territories east of the Enns is of Burchard, who probably was count (burgrave) of Regensburg. It appears that he lost his office as a result of his championship of Henry II the Quarrelsome, duke of Bavaria. In 976 his successor, Leopold I of the House of Babenberg, was installed in office. Under Leopold's rule the eastern frontier was extended to the Vienna Woods after a war with the Magyars. Under his successor, Henry I, the country around Vienna itself must have come into German hands. New marches were also created in what was later known as Carniola and Styria. Wars against Hungarians and Moravians took up the reign (1018–55) of Margrave (a count who ruled over a march) Adalbert. Parts of Lower Austria on both sides of the Danube were lost temporarily; after they were retaken, they became the so-called Neumark (New March), which for some time enjoyed independence—as did the Bohemian march to the north of the Babenberg territories. The position of the Babenbergs was at that time still a modest one; their territorial rights were no greater than those of other leading noble families. Their power within their own official sphere was further diminished by ecclesiastical immunities (Passau in particular, but also Salzburg, Regensburg, and Freising), with numerous monasteries owning large territories as well.

Austria was repeatedly drawn into the disputes of the investiture controversy in which the Pope and the Holy Roman Emperor fought for control of the church in Germany. In 1075 Margrave Ernest, who had regained the Neumark and the Bohemian March for his family, was killed in the Battle of the Unstrut, fighting on the side of the king (later emperor) Henry IV against the rebellious Saxons. Altmann, bishop of Passau, a leader of church reform and a champion of Pope Gregory VII, influenced the next Babenberg margrave, Leopold II, to abandon the cause of Henry IV. As a result, Henry roused the Bohemian duke Vratislav II against him, and in 1082 Leopold II was defeated near Mailberg and his territories north of the Danube devastated. The Babenbergs, however, managed to survive these setbacks. Meanwhile, the cause of church reform gained ground, with its centres in the newly founded monasteries of Göttweig, Lambach, and, in Styria, Admont.

Involvement in the papal-investiture controversy

Under Leopold III (1095–1136) the history of the Babenbergs reached its first culmination point. In the struggle between emperor and pope, Leopold avoided taking sides until a consensus had built up among the German princes that it was Emperor Henry IV who stood in the way of a final settlement. Then Leopold did not hesitate to side with Henry's rebellious son, Henry V (1106). For this he was rewarded with the hand of Henry V's sister Agnes, who had formerly been married to the Hohenstaufen Frederick I of Swabia. The intermarriage with the reigning dynasty not only increased Leopold's reputation but no doubt also brought him additional power. Leopold was even proposed as a candidate to the royal throne, but he declined. It was apparently his intention to concentrate on consolidating his position in Austria. He was the first Austrian margrave to describe himself as the holder of territorial principality (*principatus terrae*), and during his time Austrian common law is mentioned for the first time, another proof of the developing national consciousness.

Leopold's reputation with the clergy was high, and he was eventually canonized (1485). He gave generous endowments to religious communities, establishing the Cistercians at Heiligenkreuz, and he founded, or at least restored, the monastery of Klosterneuburg, which he gave to Augustinian canons. It was in Klosterneuburg as well that he built a residence in which he stayed even after he had acquired Vienna.

On the death of Leopold III, the Babenbergs were drawn into a conflict between the two leading dynasties of Germany, the Hohenstaufen and the Welfs—on the side of the Hohenstaufen because of their family ties. In 1139 the German king Conrad III bestowed Bavaria, which he had wrested from the Welfs, on his half-brother, Leopold IV. After the latter's untimely death, Henry II Jasomirgott succeeded to the rule of Austria and Bavaria.

Conflict of the Hohenstaufen, Welfs, and Babenbergs

Emperor Frederick I Barbarossa tried to put an end to the quarrel between the Welfs and the Hohenstaufen, and in the autumn of 1156 at Regensburg he arranged a compromise. Bavaria was restored to the Welf, Henry the Lion, duke of Saxony, while the Babenbergs were confirmed in their rule of Austria, which was made a duchy, and were given the "three countships," the actual location of which is disputed. Also, the obligations of the dukes of Austria toward the empire were reduced. Their attendence at royal court days was only called for when court was held in Bavaria, and they were compelled to participate only in campaigns of the empire that were directed against Austria's neighbour; that is, Hungary. Henry II Jasomirgott and his wife, Theodora, a Byzantine princess, were granted succession through the female

line and the right, in the event of the premature death of their children, to appoint a candidate for the succession. The Babenbergs also were given the right of approving the exercise of jurisdiction by other powers within the new duchy, permitting Henry to exert pressure against such rival internal powers, secular as well as ecclesiastical. The rights of the duke were laid down by imperial charter (*Privilegium Minus*). For centuries, however, Austria continued to contain territorial dominions not subject to the duke. Henry moved his residence to Vienna, where he also founded the monastery of the "Scottish" (actually Irish) monks.

Acquisition of Styria by the Babenbergs

In 1192 the Babenbergs' territory was greatly extended when they won the duchy of Styria. In Styria the margraves of the family of the Otakars of Steyr had gradually asserted themselves—under conditions similar to those of the Babenbergs—over their rivals, the noble families of the Eppensteiner, Formbacher, and Aribonen. The most successful among the Styrian margraves had been Otakar III (reigned 1130–63). Then, in 1180, Emperor Frederick I, in the course of a renewed anti-Welf policy, raised Styria to the status of a duchy and granted it complete independence from Bavaria. A few years later, a treaty of inheritance (Georgenberg; 1186) was concluded between the dukes Leopold V of Austria (reigned 1177–94), a son of Henry Jasomirgott, and Otakar IV of Styria, the ailing last Otakar ruler. When Otakar died in 1192, Leopold succeeded him, and thus the Babenbergs came into the inheritance.

With the exception of a short intermission (1194–98), the reigning Babenberg henceforth ruled both duchies, Austria and Styria. Styria then included parts of the Traungau, which eventually was to become part of Upper Austria, and the province of Pitten, north of the Semmering, afterward assigned to Lower Austria. In logical continuation of the Babenberg policy, Leopold VI the Glorious and his successor, Frederick II the Warlike, the last representative of the dynasty, extended their domains farther south, gaining fiefs in Carniola.

Before he had inherited the duchy of Styria, Leopold V had taken part in the Third Crusade, during which, on the ramparts of Acre, he had become involved in a quarrel with the English king, Richard I the Lion-Heart. Later, on his return journey to England, Richard tried to make his way through Austria in disguise but was recognized near Vienna, taken prisoner, and later handed over to Emperor Henry VI. England had to pay a heavy ransom, a share of which Leopold obtained and invested in the foundation, extension, and fortification of towns as well as in the stamping of a new coin, the so-called Wiener Pfennig. The road connecting Vienna and Styria was improved, and the new town of Wiener Neustadt was established on its course to protect the newly opened route across the Semmering Pass.

On Leopold V's death the Babenberg domains were divided between his sons for four years, until the death of one of them, Frederick I, in 1198. His brother Leopold VI, the most outstanding member of the family, then took over as sole ruler (1198–1230). This was a time of great prosperity for the Babenberg countries. In imperial politics Leopold VI again took sides with the Hohenstaufen, backing Philip of Swabia. In church matters the Duke was a great supporter of the monasteries, founding a Cistercian monastery at Lilienfeld (*c.* 1206). He tried to concentrate patronage rights over ecclesiastical property in his own hands and took rigorous action against the heretics (Cathari and Waldenses). He participated in several crusades in Palestine, Egypt, southern France (against the Albigenses), and Spain (against the Saracens). Leopold VI's efforts to emancipate Austria ecclesiastically by creating a separate Austrian bishopric in Vienna came to naught because of the opposition of the church in Passau and especially Salzburg; nor did his son Frederick II succeed in the same matter. Leopold VI played some role in imperial politics, bringing about the peace Treaty of San Germano between Emperor Frederick II and Pope Gregory IX (1230). He met his death in San Germano, and his body was transported to Lilienfeld to be buried there.

The rule of Leopold VI

A change came about under the last representative of the dynasty, Frederick II the Warlike, Leopold's son. His harsh internal policy and military excursions against neighbouring lands, together with his opposition to the emperor Frederick II, led in 1237 to the temporary loss of both Austria and Styria. The crisis, however, was overcome, and fresh opportunities were about to open for the Duke when, on June 15, 1246, he was killed in battle against the Hungarians on the Leitha River. With him the male line of the family came to an end.

The political history of Austria from the end of the 10th to the middle of the 13th century is marked by the establishment and consolidation of territories. This process was most advanced in the Babenberg domains, but was not confined to them. Dukes Herman (1144–1161) and Bernhard (1202–1256) of Carinthia achieved a comparable status, and Count Albert of Tirol (died 1253) moved in the same direction. The archbishops of Salzburg strove to eliminate all secular powers and patrons of their see, but in the other territories, secular princes strengthened their rule.

Another milestone of this period was the completion of the colonization of the Austrian territories. New settlements were now established by clearing the woods and advancing to more remote mountain areas. Several old and new settlements grew into market centres and towns and were eventually granted charters. The colonization movement also affected the ratio of German to non-German population. Except for some places in the Alpine regions, the Slavs were gradually assimilated, and the same holds true of the remnants of the Roman population in Salzburg and northern Tirol.

The intellectual life of the period deserves mention, too. The Babenberg court was famous enough to attract some of the leading German poets. At the beginning of the 13th century the Nibelung saga was written down by an unknown Austrian. Historical writing flourished in the monasteries. The era also produced first-rate Romanesque and early Gothic architecture.

LATE MIDDLE AGES

The contest for the Babenberg heritage. Upon the death of Frederick II the Warlike, the Babenberg domains became the political objects of aspiring neighbours. The Emperor and the Pope also tried to intervene. Two female descendants of the Babenbergs, Frederick's niece Gertrude and his sister Margaret, were considered to embody the claims to the heritage. Gertrude married first the Bohemian prince Vladislav and afterward the Margrave Hermann of Baden, who died in 1250. After Hermann's death, Otakar II (Přemysl Otakar II), prince of Bohemia (from 1253 king), married the widowed Margaret. Thereupon Hungarian forces intervened. Under the Treaty of Ofen (1254) Otakar was to rule Austria, while King Béla IV of Hungary received Styria. Troubles in Salzburg, stemming from a conflict between Bohemia and Hungary, inspired a rising among the Styrian nobles. Otakar intervened and in the Treaty of Vienna (1260) took over Styria as well. The state of anarchy that prevailed in Germany during this period proved advantageous to Otakar, who was granted both Austria and Styria in fief from Richard, earl of Cornwall, the titular German king. The grant, however, was only by writ and was invalid according to German law. During the following years Otakar's energetic rule met with growing opposition among the Austrian nobility. He introduced foreigners into important official positions, broke fortresses that had been erected without his consent, and dissolved his childless marriage with Margaret. Otakar had two of the opposition leaders, Otto of Meissau and Seifried of Mahrenberg, executed. The inhabitants of the cities, on the other hand, and the gentry, as well, generally favoured Otakar, who supported the churches and the monasteries. To complete his success, Ulrich of Spanheim, duke of Carinthia, willed Carinthia and Carniola to Otakar in 1269.

Rise of Přemysl Otakar of Bohemia

Reverses came only when Count Rudolf IV of Habsburg was elected German king as Rudolf I on September 29, 1273. Cautiously but nevertheless energetically,

Rudolf set about to undermine the powerful position Otakar had created for himself. He challenged the legitimacy of Otakar's acquisitions and finally placed the Bohemian king under the ban of the empire. In 1276 Rudolf and his allies invaded Austria, forcing Otakar to do homage and to renounce his claims to Austria. Two years later, while trying to recover what he had lost, Otakar was defeated by the united forces of Rudolf and the Hungarians and was killed on the battlefield near Dürnkrut (August 26, 1278).

The accession of the Habsburgs. As the German princes had not cared to give Rudolf adequate support against Otakar, he did not feel bound to them and set about to acquire the former Babenberg lands for his own house. In 1281 he made his eldest son, Albert (later Albert I, king of Germany), governor of Austria and Styria; on Christmas, 1282, he invested his two sons, Albert and Rudolf II, with Austria, Styria, and Carniola, which they were to rule jointly and undivided. As the Austrians were not used to being governed by two sovereigns at the same time, the Treaty of Rheinfelden (June 1, 1283) provided that Duke Albert should be the sole ruler. In 1282 Carniola had already been pawned to Meinhard II of Tirol (of the counts of Gorizia), one of the most reliable allies of Rudolf who, in 1286, was also invested with Carinthia.

Rule of Albert of Habsburg

At first the Habsburg rulers were far from popular in Austria. Albert's energetic and relentless rule roused bad feeling, and the Swabian entourage that had arrived with the new dynasty to occupy key positions was despised by native nobles. There were conflicts with Bavaria, Salzburg, and Hungarian nobles who violated the Austrian frontier. After the death of King Rudolf (1291), all the neighbours and rivals of the Habsburgs and the counts of Gorizia united. Albert, however, succeeded in negotiating a peace with his most dangerous foes, the Hungarians and the Bohemians, and he broke the fortresses of the rebel nobility. Meanwhile, Meinhard II had stifled the uprising in Carinthia.

In 1292 Albert had been passed over in the German election, and Adolf of Nassau was called to the throne. When Adolf fell out with the electoral princes, however, they went over to Albert, who had just subdued another rebellion in Austria. After Adolf was defeated and killed near Göllheim (1298), Albert had himself elected a second time. In his Austrian lands Albert's main concern was to provide for an effective administration, in which he was assisted by his privy councillors, most of whom were foreign. Records were set up to codify the prerogatives and returns of the ducal property. Eventually Albert did not spare the church, either. When the Přemysl family died out in 1306, Albert aspired to the Bohemian throne. He had his eldest son, Rudolf III, elected Bohemian king, but Rudolf died the following year. Albert was preparing for a new campaign when he was murdered by his nephew, John, and some accomplices (1308).

On Albert's death the anti-Habsburg movement flared up again in Austria, but his sons, Frederick I the Fair and Leopold I, managed to maintain control. Frederick stood for election as German king (as Frederick III), and for the next years the Habsburg countries had to support the cost of the war with his rival, Louis IV of Bavaria, until 1322, when Frederick was defeated near Mühldorf. Earlier, another more decisive battle had been lost by the Habsburgs to the Swiss at Morgarten in 1315. From that time on, the Habsburg domains in the territory south of the Rhine and the Bodensee (Lake Constance) began to crumble away. Frederick the Fair spent his last years in Austria and was buried in the Carthusian monastery of Mauerbach (1330). He seems to have been the first of the Habsburgs for whom Austria meant home. From his time on, Habsburg rule and Habsburg territories were known as the Austrian domains (*dominium Austriae*), a term that was replaced, in the course of the 14th and 15th centuries, by the new concept of the House of Austria.

After Frederick's death the Habsburgs were for some time ruled out as possible candidates for the German throne; but, under the brothers Albert II and Otto, Habsburg Austria received its first important accession of territory. In 1335 Carinthia and Carniola were acquired after the death of Henry of Gorizia; while, with the help of Luxembourg troops, Henry's daughter Margaret (surnamed Maultasch) managed to retain the Tirol. Albert and his brother Otto had not gotten on too well, but when Albert came to rule on his own, he proved to be of sound judgment and keen on preserving the peace. It was a time of severe catastrophes: bad harvests, floods, and earthquakes, and in 1348–49 the plague, which brought a persecution of the Jews that was suppressed, however, by the Duke. Albert arranged several tours around his domains to establish contacts with the populace and improve jurisdiction. Two campaigns against the Swiss failed to yield any spectacular results, but they helped once more to consolidate the weakened Habsburg position. At his death in 1358, Albert left four sons. Though in 1355 a family ordinance had decreed that all the male members of the family were to rule jointly over the undivided domains, only the eldest among them, Rudolf, was then fit to rule. Throughout his short reign (1358–65), Rudolf IV showed himself extremely energetic and ambitious. He started to rebuild St. Stephen's Cathedral in the Gothic style, and he founded the University of Vienna (1365). With these two projects he imitated and rivalled his father-in-law, the emperor Charles IV, at Prague.

Acquisition of Carinthia, Carniola, and Tirol

In 1359 Rudolf's forged charter, the Privilegium Majus, by which he claimed immense privileges for Austria and its dynasty, as well as the title of archduke, caused a breach between him and the emperor Charles IV. Charles was not prepared to accept the Privilegium Majus to its full extent (although it later was sanctioned by the Habsburg emperor Frederick III in 1442 and again in 1453). Upon news of the death of Duke Meinhard in 1363, Rudolf prevailed upon the Duke's mother, Margaret, to make over the Tirol to him. On this occasion the Emperor backed the Habsburgs against the Wittelbachs, and the Tirol thus passed to the House of Austria.

The division of the Habsburg lands. Rudolf was succeeded in 1365 by his two brothers, Albert III and Leopold III. After some years of joint rule, however, they quarrelled and in 1379, by the Treaty of Neuberg, partitioned the family lands. Albert, as the elder brother, received the more prosperous countries on the Danube (Upper and Lower Austria). The rest of the widespread domains fell to Leopold (including Styria, Carinthia, Tirol, the old Habsburg countries in the west, and central Istria). The treaty also contained several points on mutual wardship, preemption rights, and common titles, by which some connection between the two lines was to be preserved.

Treaty of Neuberg of 1379

In 1381 the resourceful Duke Leopold took advantage of the weak position of Venice in its war with Genoa and seized Trieste, which had broken away from Venice. His efforts to expand his rule in the west, however, were less successful, though he seemed lucky enough at first. Envisaging a connection between the original Habsburg territories in the west and the new domains in the Tirol, the Habsburgs looked for a foothold in the region west of the Arlberg (modern Vorarlberg). Neuberg on the Rhine was won in 1363 and Feldkirch in 1375. Another important acquisition was the city of Freiburg in the Breisgau. But then Leopold came into conflict with the Swiss, which led to defeat and his death at Sempach in 1386. An army of his brother, Albert III, was likewise defeated near Näfels in 1388, and the Habsburgs suffered heavy territorial losses. Leopold's sons recognized the wardship of Albert, who acquired Bludenz and the Mantafon Valley west of the Arlberg in 1394. In his own domains Albert was forced to check the dynasty of the Schaunbergs (in Upper Austria), who tried to create an independent domain around Wilhering and Eferding. Albert III especially favoured the city of Vienna as his capital, and it was because of his reorganization that the university Rudolf IV had founded there was able to survive.

After Albert's death in 1395, new Habsburg family troubles arose, differences that the treaties of Hollenburg (1395) and Vienna (1396) tried to settle. Under the Vienna treaty, the line of Leopold III split into Styrian and Tirolian branches, resulting in three complexes of Aus-

trian territories—a state of affairs that was to reappear in the 16th century. The individual parts came to be known by the names of Niederösterreich (comprising modern Lower and Upper Austria), Innerösterreich (comprising Styria, Carinthia, Carniola, and the Adriatic possessions), and Oberösterreich (comprising the Tirol and the western domains, known as the Vorlande, or Vorderösterreich).

The early diets of the nobility, church, and towns

In 1396 the Austrian estates, or diets, were first assembled to consider the Turkish threat and henceforth were to play an important political role in Austria. In them the nobility usually took the lead, but they also included representatives of the monasteries, the towns, and the marketplaces. In the Tirol, in Vorarlberg, and, at times, in Salzburg, the peasants also sent their representatives to attend the diets. Because of the Habsburg partitions and frequent regencies, the estates were able to gain in importance. They did not obtain the right to pass laws, but they obstinately insisted on the privilege to grant taxes and duties.

After the short rule of Albert IV (1395–1404) and a troublesome tutelary regime (1404–11), Albert V came into his own, and with him the Danube countries again enjoyed a strong and energetic rule (1411–39). Albert, however, had married the daughter of the emperor Sigismund and was thus drawn into the Hussite religious wars, in the course of which the Austrian lands north of the Danube were ravaged. In the Austrian west, Duke Frederick IV of the Tirolian branch lost the Aargau to the Swiss but was able to assert himself in Tirol against a rebellion of his nobles.

When Sigismund died, Albert inherited his positions. In 1438 he was first elected Hungarian king, with the German (as Albert II) and the Bohemian crowns to follow later. Albert no doubt had many of the qualities of a born ruler, but he died prematurely in 1439 on an unsuccessful campaign against the Turks. Soon thereafter, his widow gave birth to a son and heir, Ladislas Posthumus, to whom Frederick V of Styria, as the senior member of the house, became guardian. Frederick also had Sigismund, the son of Frederick IV of Tirol, under his tutelage.

Thus began the long reign of Frederick V (as Roman emperor he was to become Frederick III). His reign was marked by almost ceaseless strife with the estates, with his neighbours, and with his jealous family. When he tried unsuccessfully to take advantage of a conflict among the Swiss Confederates, the Tirolians made Frederick release Duke Sigismund from tutelage (1446). A few years later, on his return from Rome, where he had been crowned emperor, his enemies at home and abroad in 1452 forced him also to give up Ladislas, who was then the recognized king of Hungary and Bohemia. The boy king's policies were made by Count Ulrich of Cilli. Ulrich was murdered at Belgrade in 1456, however, and a year later King Ladislas died. In Bohemia and Hungary national kings came to power. Frederick now won himself a foothold in the Austrian domains on the Danube and succeeded in acquiring the rich estates and fiefs of Ulrich.

The Burgundian and Spanish marriages. Maximilian I, the son of the emperor Frederick III, was married to the Burgundian heiress, Mary, at Ghent (1477). By that tie to Burgundy the Habsburgs became involved in long struggles with France. After Mary's death (1482), Maximilian, moreover, met with increasing difficulties in the Burgundian countries themselves. In the meantime, another crisis had arisen in the eastern Habsburg domains. Disagreement about the Bohemian succession and a political error of Frederick III, who tried to install the former archbishop of Gran (Esztergom) at Salzburg, led King Matthias I Corvinus of Hungary to march against Austria. Vienna was besieged and finally taken by the Hungarians (1485), as was Wiener Neustadt (1487). The harried Maximilian came into even greater distress in the Low Countries, where the rebellious citizens of Bruges put him under arrest (1488). Sigismund, the Habsburg ruler of the Tirol, who was heavily encumbered by debts, planned to sell his country to the Bavarians. A complete breakdown of the House of Habsburg threatened, but Maximilian was ultimately released. He prevailed upon Sigismund to abdicate in his favour. In 1490 the Habsburgs were able to take over Lower Austria. Maximilian even attacked Hungary but in the Treaty of Pressburg (1491) renounced claims to Hungary, though reserving the succession rights of his family.

Near collapse of House of Habsburg

After the death of his father, Emperor Frederick III, Maximilian came into a heritage that surpassed the endowments of all his predecessors. Furthermore, his son, Philip I the Handsome, who governed the Low Countries, was betrothed to the Spanish infanta, Juana (later Joan the Mad), and through the unexpected death of male members of the Spanish dynasty this marriage was to raise the Habsburgs to the throne of Spain. In the German Empire as well as in Austria, Maximilian introduced sweeping administrative reforms that were the first steps toward a centralized administration. In 1508 Maximilian assumed the title of elected emperor as he was unable to pass through hostile Venetian territory to go to Rome for his coronation, and henceforth Rome and the pope had no more say in the creation of new emperors.

During Maximilian's last years, Eastern politics again came to the fore. The great crusade he planned against the Turks, however, never materialized. In 1515 Maximilian arranged a double marriage between his family and the Jagiellon line that ruled Bohemia and Hungary, thus reviving earlier Habsburg claims to these countries. Maximilian's energetic reign added greatly to the prestige of the Habsburgs. Thus, his grandson Charles (V) was able to prevail against French intrigue to inherit the imperial crown. Charles's younger brother, Ferdinand I, took over the rule of the Austrian countries but encountered the opposition of the estates, which he cruelly suppressed. In the agreements of Worms (1521) and Brussels (1522) Charles V formally handed over the Austrian lands to his brother. The subsequent years of Ferdinand's reign were troubled by peasant risings in the Tirol and Salzburg and were followed by similar upheavals in Innerösterreich.

Summary of developments in the late Middle Ages

In the late medieval period the Alpine lands were assembled by the Habsburgs into a monarchical union comprising about the extent of the present Austrian state. The process of union was at times intercepted and hindered by the partitions among the dynasty. When the process was finished, however, the territories still preserved their individuality and their own legal codes. During this period the towns developed and prospered, but in the rural settlements a backward tendency had set in. Many settlements were abandoned, especially in Lower Austria. The leading classes lost their interest in rural colonization as they found other and more lucrative sources of income. Mining developed, but trade was impaired by political instability. To about 1450 the University of Vienna enjoyed some fame in the fields of theology and science. The literary culture of Austria was characterized by remarkable works, among them the rhyming chronicle of the Styrian Otakar aus der Geul, the work of the Carinthian abbot John of Viktring, the poetry of Oswald of Wolkenstein, and the works of the theologian and historian Thomas Ebendorfer. From the middle of the 15th century onward Austria came under the influence of Italian Humanism.

REFORMATION AND COUNTER-REFORMATION

The acquisition of Bohemia. The year 1526 saw the defeat and death of the Jagiellon king of Hungary and Bohemia, Louis II, who fell in the Battle of Mohács against the Turks. In view of the treaties of 1491 and 1515, Ferdinand I and the Vienna court envisaged Hungary and Bohemia plus the adjoining countries falling to the Habsburgs. Thus, the union of Austria, Bohemia, and Hungary became the leading concept of Habsburg politics. After clever diplomatic overtures, Ferdinand was elected king of Bohemia (October 23, 1526). In Hungary, however, there was a split election; János (John) Zápolya, *voivode* (governor) of Transylvania, was chosen by an opposition party, whereupon war broke out between the two candidates.

Ferdinand's troops in Hungary would have been in a

stronger position had Zápolya not been assisted by the Turks under Süleyman I. In 1529 the Turks advanced as far as Vienna, which they besieged in vain. Another Turkish offensive came to a halt at Güns in western Hungary in 1532. Ferdinand, on the other hand, failed in his attempt to take Ofen (Hungarian Buda), where the Turks had entrenched themselves. By around the middle of the century the frontiers had become fixed. Hungary happened to be divided into three parts: the west and the north remained with the Habsburgs, the central part came under Turkish rule, and Transylvania and adjoining territory were kept by Zápolya and his successors. This situation was anticipated in the truce of 1547 and became formalized in the Peace of Constantinople (1562).

Ferdinand I's administrative reforms

During a short truce in the fighting against Zápolya and the Turks, Ferdinand started to reorganize Austrian administration. In 1527 he created new central organs: the Privy Council (Geheimer Rat) for foreign affairs and dynastic matters; the Court Council (Hofrat) as the supreme legal authority; the Court Chancery (Hofkanzlei), which served as the central office and only later on was to deal with internal affairs; and the Court Treasury (Hofkammer) for finance and budgeting. As the Court Treasury proved inefficient in the financing of the Turkish war, the Court Council of War (Hofkriegsrat) was established in 1556 to take care of the pay, equipment, and supplies of the troops, acquiring some influence on military operations as well.

The advance of Protestantism. The Protestant movement gained ground in Austria very fast. The nobility especially turned toward the Lutheran creed. For generations eminent families provided the protagonists of Protestantism in the Lower and Inner Austrian territories. The sons of the nobility were often sent to the north German universities to expose them more fully to Protestant influence. From 1521 Protestant pamphlets were produced by Austrian printers. Bans on them, issued from 1523 onward, remained ineffective, however.

Among the peasant population the Anabaptists had a stronger appeal than the Lutherans. As they had no support from the estates and because of their radicalism, however, the Anabaptists were persecuted from the start. In 1528 Balthasar Hubmaier, their leader in the Danube countries and southern Moravia, was burned at the stake in Vienna, and in 1536 another Anabaptist, the Tirolian Jakob Hutter, was put to death in the same way in Innsbruck after he had led many of his followers into Moravia. Ferdinand, for his part, advocated religious reconciliation and looked for means to achieve it; but the dogmatic viewpoints proved irreconcilable. The Peace of Augsburg (1555) finally brought some respite in the religious struggles.

When Charles V abdicated, Ferdinand I became emperor (1558), and thus the leadership of the empire was taken over by the Austrian (German) line of the Habsburgs. Maximilian II, the eldest son, followed his father in Bohemia, Hungary, and the Austrian Danube territories (1564). The next son, Ferdinand, was endowed with Tirol and the Vorlande; Charles, the youngest of the brothers, received the Inner Austrian lands, and took up residence in Graz. Maximilian was known for his Protestant leanings but was bound by a promise he had given his father to remain true to the Catholic religion. The Protestants were therefore granted fewer concessions from him than they might have expected.

Catholic reaction and the Counter-Reformation

Meanwhile, Catholic counteractivity began, with the Jesuits particularly prominent in Vienna, Graz, and Innsbruck. A new generation of energetic bishops took part and proved a great asset to the cause. It was also of some importance that the monasteries, though they had been deserted by many of their members and were struggling for existence, had not been secularized. On the Protestant side, it proved impossible to reconcile the various reforming movements. Social differences between them, especially between the nobility and the peasants, also stood in the way of a united Protestant front. The Counter-Reformation scored its first successes in Gorizia and Carniola, where Protestantism had remained insignificant. And in other parts, official religious commissions started to replace the Protestant preachers with Catholic clergymen.

Rudolf II and Matthias. Maximilian's successor, Rudolf II (reigned 1576–1612), had been educated in Spain strictly in the Catholic faith. He had all Protestants dismissed from court service. The conversion of the cities and market centres of Lower Austria to Catholicism was conducted by Melchior Klesl, at that time administrator of the Vienna see but later to become bishop and cardinal. In Upper Austria, where the Protestants had their strongest hold, the situation remained undecided, with the Catholic governor, Hans Jakob Löbl of Greinburg, and the Calvinist Georg Erasmus of Tschernembl leading the opposing religious parties. When Charles's son, Ferdinand II, took over in Styria, he proved to be the most resolute advocate of the Counter-Reformation. It was he who eventually succeeded in uprooting Protestantism, first in Inner Austria and then in the other Habsburg countries, with the exception of Hungary and Silesia.

From local skirmishes along the frontier, a long, drawn-out war with the Turks developed (1592–1606). In 1598 Raab (Hungarian Győr), which served as a bastion of Vienna, was temporarily lost; Gran, Veszprém, and Stuhlweissenburg (Hungarian Székes-fehérvár) passed several times from one side to the other. The introduction of the Counter-Reformation in Hungary, moreover, resulted in a rising of Protestant elements under István Bocskay. But in 1606 at Vienna a peace was concluded between Austria and the Hungarian estates. At Zsitvatorok another peace was negotiated with the Turks, who for the first time recognized Austria and the emperor as an equal partner.

Conflict between Rudolf and Matthias

Political disagreements between Emperor Rudolf, who to an increasing degree showed signs of mental derangement, and the rest of the family led to the "Habsburg Brothers Conflict." Cardinal Klesl in 1607 brought about an agreement among the younger relatives of the Emperor to recognize the Emperor's brother Matthias as the head of the family. As the conflicts with Rudolf persisted, Matthias strove also to come to an understanding with the estates, which were mainly Protestant. The formation of opposing religious leagues in Germany, the Protestant Union and the Catholic League, also added to the general confusion.

Matthias advanced into Bohemia, and, in the Treaty of Lieben (1608), Rudolf conceded to him the rule of Hungary, the Austrian Danube countries, and Moravia, while Matthias had to give up the Tirol and the Vorlande to the Emperor. In 1609 the estates received a confirmation of the concessions Maximilian II had made to them. The cities were guaranteed only in general terms that their old privileges should not be interfered with. At the same time, Rudolf II was forced to grant to Bohemia the so-called Letter of Majesty, which contained far-reaching concessions to the Protestants. After a final defeat of Rudolf in Bohemia in 1611, Matthias was crowned king of Bohemia. Rudolf's death in 1612 finally ended the conflict.

After Matthias had been elected emperor, his principal councillor, Cardinal Klesl, tried in vain to arrange an agreement with the Protestants in Germany. The ensuing years were filled with wars in Transylvania, where Gábor Bethlen came to power. In the Peace of Tyrnau (1615) the Emperor had to recognize Bethlen as prince of Transylvania, and in the same year he extended the truce with the Turks for another 25 years. In the meantime, war had broken out with Venice (1615–17) because of the pirating activities of the Serb refugees (Uskoken) established on the Croatian coast. A settlement was reached in the Peace of Madrid. The situation in Bohemia then reached a critical point, the religious tensions in the country finding a vent in the "Defenestration of Prague" (May 23, 1618), in which two of the Emperor's regents were thrown from the windows of the Hradčany Palace.

The Bohemian rising and the victory of the Counter-Reformation. War became inevitable when Emperor Matthias died in 1619. Not that he had been master of the situation, but his death brought Ferdinand II, the

most uncompromising Counter-Reformer, to the head of the House of Habsburg. Ferdinand was hard pressed at first, as Bohemian and Moravian troops invaded Austria. A deputation of the estates of Lower Austria tried to make him renounce Bohemia in a peace treaty and demanded religious concessions for themselves, unsuccessfully, however. The Bohemians were forced to retreat, and imperial troops advanced into their country. The Bohemians deposed Ferdinand from the throne of Bohemia and elected Frederick V in his stead. But two days later Ferdinand II was elected German emperor at Frankfurt (August 28, 1619).

War over the Bohemian crown

War was the only means of resolving the issue. The conflict for the Bohemian crown developed into a European war when Spain, the Bavarian duke Maximilian I, and the Protestant elector of Saxony entered the struggle on the side of the Emperor. The Upper Austrian estates rashly joined Frederick, with the result that their country was occupied by the army of the Catholic League and afterward pledged to Bavaria. At the Battle of the White Mountain, Ferdinand became master of Bohemia, Moravia, and Silesia, while Lusatia was pledged to Saxony. King Frederick fled to the Netherlands. The leaders of the Bohemian rising were executed, and other nobles who had compromised themselves lost their property. Many Protestants left the country. In the new constitution of 1627 Bohemia and its associated lands became a hereditary kingdom. The diets were not dissolved entirely, as the government wanted to make use of their administration, but their influence was restricted to financial matters.

After the death of Matthias, Ferdinand had also inherited the Danubian territories. Tirol, however, retained a special status under a new Habsburg secundogeniture (inheritance by a second branch of the house). Upper Austria, pledged to Bavaria, was disturbed by a great peasant rising. The Protestant peasants were defeated after heavy fighting, and in 1628 the country passed into the hands of the Emperor again.

The Counter-Reformation was vigorously enforced in the Austrian domains. This led to the mass emigration of Protestants, including many members of the nobility. Most went to the Protestant states and to the imperial cities of southern Germany. After the Bohemian victory the war went favourably for the Emperor, and the Peace of Lübeck (1629) seemed to secure the hegemony in Germany for the Habsburgs. But in 1629 Ferdinand's attempt in the Edict of Restitution (Restitutionsedikt) to establish religious unity by force throughout the empire provoked the violent opposition of the Protestants.

The struggle with Sweden and France. July 1630 saw intervention in Germany's religious strife from a different quarter—Sweden. In that month the Protestant Swedish king, Gustavus II Adolphus, landed on the Baltic coast of Pomerania. His purpose was to defend the Protestants against further oppression, to restore the dukes of Mecklenburg, his relatives, who had been driven from their lands by Ferdinand's forces, and perhaps to strengthen Sweden's strategic position in the Baltic. In the ensuing conflict the city of Magdeburg was destroyed by fire after it had been taken by the troops of the Emperor under Gen. Johann Tserclaes, Graf von Tilly (1631). The north German Protestants, who had so far remained undecided, consequently went over to the Swedes. After victories near Breitenfeld and on the Lech, the Swedish troops entered Bavaria.

During the subsequent period of the Thirty Years' War (*q.v.*), Ferdinand adopted a rigorous and often unrelenting attitude, though he yielded a little when the Peace of Prague was being negotiated (1635). His successor, Ferdinand III (1637–57), was as loyal to Catholicism as the father had been but showed himself more of a realist. He was not able, however, to prevent the war from again dragging into Habsburg territory, so that in 1645 even Vienna was threatened. The extremist party that had rejected all concessions lost its influence at the Vienna court, and two able diplomats, Maximilian, Graf von Trauttmansdorf, and Isaac Volmar, were entrusted with the representation of a weakened Austria at Münster and Osnabrück, where extended negotiations were conducted until acceptable terms could be settled for Austria. In the Peace of Westphalia (1648) Austria lost its possessions in Alsace, and Lusatia had to be ceded for good to Saxony. The peace in many respects marked the beginning of a new epoch. The Holy Roman Empire from then on was reduced to a loose union of otherwise independent states, and Habsburg politics shifted its emphasis, falling back entirely on the political, military, and financial resources of the hereditary Habsburg lands, now including also Bohemia. The new central organs and the administrative bodies of the territories took on much greater importance than the remaining institutions of the Holy Roman Empire. The Emperor came to rely on a standing army rather than upon troops provided by the German princes.

Economic and cultural consequences of the religious struggle

The heavy drain the religious wars had made on the population of the Austrian territories was compensated for by immigrants from the Catholic parts of the empire and by Croatian refugees from the southeast. The economic position of the peasants on the whole deteriorated. Many members of the nobility, as well as the church, acquired new property. In mining, boom and depression followed quickly upon each other. The loss of many experienced miners during the Counter-Reformation resulted in difficulties, but the government took several steps toward improving and extending the salt mines. In 1625 it founded the Innerberg Union, under which the Styrian iron industry was reorganized. The Emperor also tried to interfere with the trade organizations of the towns, though without much success. Trade and finance in the Austrian territories was dominated by foreign capital.

The cultural life of the period was also dominated by the religious struggle. In the field of education the schools of the denominational parties rivalled each other. In 1585 a Jesuit university was founded at Graz, while at Salzburg a Benedictine university was established (1623). Austrian humanists produced some outstanding works of poetry and historical writings, and the sovereigns were great patrons of the arts, but on the whole this was an epoch dominated by Italian and Western influences.

AUSTRIA AS A GREAT POWER

After the Thirty Years' War, Austrian politicians were understandably reluctant to enter into another military conflict. In 1654 Ferdinand IV, the eldest son of the Emperor, died. His brother Leopold, who had been destined for a church career, then was considered as heir to the throne and was recognized as such by Austria, Bohemia, and Hungary. In Germany, however, difficulties arose when France declared itself against Leopold. Nevertheless, after the death of Emperor Ferdinand, Leopold was finally elected (1658), after having conceded constitutional limitations that restricted his liberty of action in foreign politics. West German princes under Johann Philipp von Schönborn, archbishop of Mainz, formed the French-oriented League of the Rhine. At the same time, Austria was engaged in the northeast, when it intervened in the war between Sweden and Poland (1658) in order to prevent the collapse of Poland. There were some military successes, but the Treaty of Oliva (1660) brought no territorial gains for Austria, though it stopped the advance of the Swedes in Germany.

During the Thirty Years' War the Turkish front had been quiet, but in the 1660s a new war broke out with the Turks (1663–64) because of a conflict over Transylvania, where a successor had to be appointed for György II Rákóczi, who had been killed fighting against the Turks. The Turks conquered the fortress of Neuhäusel in Slovakia, but the imperial troops succeeded in throwing them back. The Austrian military success was not, however, reflected in the terms of the Treaty of Vasvár: Transylvania was given to Mihály Apafi, a ruler of pro-Turkish sympathies. A minor territorial concession was also made to the Turks. The year after the Turkish peace, Tirol and the Vorlande reverted to Leopold I (1665), and the second period of the Habsburg partition (1564–1665) came to an end.

In Hungary dissatisfaction with the results of the Turkish war spread. Not only the Protestants, who were threatened by the Counter-Reformation, but also many Catholic nobles were alarmed by Habsburg absolutism. A group of Hungarian nobles and the Styrian count Hans Erasmus of Tattenbach entered into a conspiracy. The Austrian government, informed of their activities, had four of the ringleaders executed—an action that led to a rising by rebels known as Kuruzen (Crusaders).

In the meantime, the position of the Habsburgs in the west had again deteriorated. At first Leopold I's leading statesmen, Johann Weikhard, Fürst Auersperg (dismissed in 1669), and the president of the Court Council of War, Wenzel Eusebius, Fürst von Lobkowitz, remained rather passive in view of the expansionist policies of Louis XIV of France. They also stayed outside the Triple Alliance of Holland, England, and Sweden that was concluded in order to ward off the attacks of Louis against the Spanish Netherlands. When Louis actually invaded Holland, the Emperor finally entered the war, but in the ensuing Treaty of Nijmegen (1679) he had to cede Freiburg im Breisgau to France.

The threat to Vienna by the Turks

Another and still more menacing danger appeared in the southeast. After some deliberation, the leader of the Hungarian rebels, Imre Thököli, had asked the Turks for help, whereupon the grand vizier Kara Mustafa organized a large Turkish army and marched it toward Vienna. Habsburg diplomats succeeded in concluding an alliance between Austria and Poland. Meanwhile, imperial troops under Duke Charles of Lorraine tried to hold the enemy but had to retreat. From July 17 to September 12, 1683, Vienna was besieged by the Turks. Deciding against a direct assault, the Turks began to drill tunnels underneath the bastions of the city, when relief columns arrived from Bavaria, Saxony, Franconia, and Poland. King John III of Poland took over the command of the relieving army, which descended upon the Turks and dispersed them. The Emperor concluded a pact with Poland and the Venetian republic—the Holy League. In 1685 Neuhäusel was won back, and in September of 1686 Ofen was captured, despite fierce Turkish resistance.

In 1687 the Hungarian diet recognized the hereditary rights of the male line of the Habsburgs to the Hungarian throne. In 1688 Belgrade was conquered and Transylvania was secured by imperial troops. In the meantime, Louis XIV had begun an offensive against the German Palatinate. This meant that no further troops could be spared for the Turkish war, and in 1690 all recent conquests in the south, including Belgrade, were lost again. A victory of the imperial and the allied German troops under Margrave Louis of Baden near Slankamen (1691) prevented the Turks from advancing farther, but then the Margrave was ordered to the Rhine front. Eventually Prince Eugene of Savoy took over the command and gained a decisive victory over the Turks near Zenta (1697). After another offensive against Bosnia, the Turks finally decided to negotiate a peace. In the Treaty of Carlowitz (1699), Hungary, Transylvania, and large parts of Slavonia fell to the Habsburg emperor. In the meantime, the war in the west had come to an end (Treaty of Rijswijk, 1697), overshadowed already by the question of the Spanish succession.

The War of the Spanish Succession. From 1701 to 1714 Austria was involved in hostilities with France over the issue of the Spanish succession. The childless King Charles II of Spain, a Habsburg, had willed his entire possessions to a Bourbon prince—a grandson of Louis XIV of France. All those who disliked the idea of a French hegemony in Europe consequently united against the French. The Emperor declared war (1701) and was immediately supported by Brandenburg-Prussia and Hanover. In the spring of 1702 England and Holland entered the war in the Grand Alliance against France. Louis XIV, on the other hand, was able to win the electoral princes of Bavaria and Cologne as his allies. At this critical juncture another Hungarian rising, led by Ferenc II Rákóczi, occurred. The rebels were prepared to join forces with the enemies of Austria and for years engaged Austrian troops. The rebels even threatened Vienna, whose suburbs had to be fortified. In the war with France, imperial troops fought on four fronts: in Italy, on the Rhine, in the Spanish Netherlands, and in Spain. Much larger forces were mobilized than had been customary during the 17th century, with the result that the financial drain on the imperial treasury was so heavy that the Emperor had to resort to Dutch and English loans. When Bavaria entered the war on the side of the French, Austria was in further danger, until the Battle of Blenheim (1704), in which a joint English and Austrian army under the Duke of Marlborough and Prince Eugene defeated the French and Bavarian forces.

Scale of the Spanish war

After a reign of 48 years filled with almost endless troubles, Emperor Leopold died in 1705. He was succeeded by his son, Joseph I (1705–11). In the religious quarrels the new emperor, an ally of Protestant states, showed great restraint and allowed himself to be guided mainly by political motives.

In 1703 the Duke of Savoy, who had left the French to go over to the Habsburgs, found himself in a critical situation; his capital, Turin, had come under French siege. An imperial army under Prince Eugene and reinforced by a Prussian contingent was sent to his aid and succeeded in uniting with the Savoyan forces and relieving Turin after a victorious battle (1706). At the beginning of the next year an agreement was reached under which the French evacuated northern Italy. The same year a smaller imperial army under Wirich, Graf von Daun, conquered Spanish-ruled southern Italy; but an invasion of southern France, which the sea powers had instigated, failed. A quick success, however, fell to the Austrians in a campaign against the Vatican state over a conflict between the Emperor and the Curia over mutual feudal rights and because of Pope Clement XI's rather pro-French leanings.

The allies were victorious in the Netherlands, winning the Battle of Oudenaarde and conquering Lille (1708). Paris seemed within easy reach. The Battle of Malplaquet (1709) was another victory for the allies, but they had to pay dearly for it. In the meantime, peace negotiations had foundered. After reverses in Spain and a political change in England, the alliance itself was in danger of falling apart. The situation was further aggravated by the death (1711) of Emperor Joseph I, who left daughters only.

At this juncture, liquidation of the Hungarian rising became possible. Rákóczi, who in 1707 had declared the deposition of the Habsburgs, began to meet with growing opposition among his followers. Imperial troops forced Rákóczi to flee to Poland; and the rebels, who had been promised an amnesty and who were guaranteed religious liberty, made their peace in 1711. From then on the Vienna government tried to be more considerate of Hungary and its aristocracy.

The Peace of Utrecht

The election of Charles VI as emperor was effected without any difficulties. The English left the coalition, and after a military reverse most of the Habsburg's allies joined the Treaty of Utrecht (1713). In the peace negotiations between Austria and France that were begun at Rastatt, Prince Eugene showed himself an unyielding and successful agent of Habsburg interests. Of the Spanish heritage, Austria gained the Spanish Netherlands, a territory corresponding approximately to modern Belgium and Luxembourg. These gains were somewhat impaired, however, by the Dutch privilege of stationing garrisons in a number of fortresses. In Italy, Austria received Milan, Mantua, Mirandola, the continental part of the kingdom of Naples, and the isle of Sardinia. The Wittelsbachs of Bavaria regained their country, but the treaty contained an appendix that provided for the eventuality of Bavaria being exchanged for the Spanish Netherlands. Of its gains, the north Italian territories were of the greatest value to Austria; the possession of Naples and the Netherlands, on the other hand, posed considerable military and political risks.

The problem of the Austrian succession. The extinction of the Spanish line of the Habsburgs and the fact that the emperor Charles VI was the last male member

The Pragmatic Sanction of Charles VI

of that house posed serious problems for the Habsburg territories, which at the beginning of the 18th century were held together mainly by the person of the sovereign, notwithstanding the fact that there were some institutions of central administration. A settlement was made in the form of a family ordinance. On April 19, 1713, Charles VI issued a decree according to which the Habsburg lands should remain an integral, undivided whole. In the event of the Habsburgs becoming extinct in the male line, the daughters of Charles or their descendants and, in default of any descendants of Charles, the daughters of Joseph I and their descendants and, after them, all other female members of the house should be eligible for the succession. As the son that was born to the Emperor in 1716 died after a few months, and only daughters were born to him after that (Maria Theresa, 1717; Maria Anna, 1718; Maria Amalia, 1724), this Pragmatic Sanction (a term used to characterize a pronouncement by a sovereign on a matter of prime importance) became of great significance. Austrian diplomacy in the last decades of Charles's reign was directed toward securing acceptance of the Pragmatic Sanction from all the European powers. It was published in 1720 and by 1722 had been recognized by the estates of all the Habsburg countries. Even the unanimous consent of the Hungarian diet was eventually obtained.

New conflicts with Turkey and the Bourbons. During the War of the Spanish Succession, Turkey had remained neutral toward Austria. But the Turks had attacked the possessions of the Venetians on the Pelopónnisos and the Ionian Isles. Austria tried to intervene and finally declared war. Prince Eugene defeated the Turks near the fortress of Peterwardein and conquered the strong bastion of Temesvár (1716). In the summer campaign of 1717 Belgrade again came into the hands of the imperial troops after a battle had been won against a Turkish relief army. In the Treaty of Passarowitz (1718) a frontier line was agreed upon that corresponded to the de facto situation. The Turks had to cede to the Austrians the Banat, the Turkish part of Syrmia, Walachia Minor as far as the Aluta, northern Serbia, Belgrade, and a strip of land along the frontier in northern Bosnia. A favourable trade agreement was also concluded.

During the Turkish war another crisis emerged. The Spanish minister Giulio Alberoni tried to initiate a policy of expansion in Italy. When Spanish troops landed in Sardinia and Sicily, the Emperor formed an alliance with Great Britain and France, later joined by the Netherlands (the Quadruple Alliance). After the English defeated the Spanish fleet, Madrid recalled its troops from the disputed territories. Austria received the more prosperous Sicily in exchange for Sardinia, which fell to Savoy. Charles then agreed to recognize the Spanish Bourbons. The gains from the Quadruple Alliance plus those of the Treaty of Passarowitz gave the Habsburgs the largest territory they were ever to rule. Their domains were far from unified, however, with the individual provinces showing a wide national, economic, cultural, and constitutional diversity.

Trading interests soon interfered with the alliance with the maritime powers. At first the attempts of the Ostend Company, which was backed by Charles VI, to enter into trade with India were quite successful. Because of the antipathy of the maritime powers, however, it seemed advisable to find an alternative to trade with Dutch and English colonial markets in the vast transatlantic empire of Spain. In 1725 Charles entered into an alliance with Spain, whereupon France, Great Britain, and Prussia formed a rival alliance. But soon after Russia was won over to the Habsburg cause, Prussia changed sides. As the outbreak of a European war seemed imminent, attempts were made at the Congress of Soissons to relax political tensions. Spain abruptly changed its alliances and concluded a treaty (1729) with England and France, the Netherlands joining in later. When Russia also began to waver, Prince Eugene tried to fall back on the traditional alliance with the maritime powers. After prolonged and difficult negotiations, England in 1731 accepted the Pragmatic Sanction, the Emperor in return giving a promise not to marry off his daughter Maria Theresa, the Habsburg heiress, to a prince who was himself heir to important domains. Austria finally dissolved the Ostend Company, having already suspended its charter in 1727. Charles VI then invested a great deal of energy in his endeavours to secure the recognition and the guarantee of the Pragmatic Sanction in the German diet. In this he was opposed by Bavaria and the elector of Saxony, but Austria finally obtained the guarantee of the Pragmatic Sanction at the Regensburg Diet (1732).

The question of the Polish succession

The question of the Polish succession led to a revival of the Austrian conflict with the Bourbon countries. Austria, with Prussia and Russia, favoured Augustus III of Saxony, the son of the deceased king, whereas France backed Stanisław I (Stanisław Leszczyński). On the military intervention of Russia in Poland, the Bourbons attacked Austria. The issue came to be mixed up with the problem of Lorraine, France dreading that on the impending marriage of Maria Theresa to Francis Stephen, duke of Lorraine, the latter's domains would be united with Austria's, so that French plans for the acquisition of Lorraine would be thwarted. France, Sardinia, and Spain simultaneously opened the war against Austria (1733). Prince Eugene, who was now aged, was able only to prevent a major success of the enemy on the Rhine. On the Italian front the Habsburgs fared even worse. The Battle of Parma ended undecided, but the Austrians were finally beaten near Guastalla. The small Austrian force that was stationed in southern Italy was unable to resist the Spanish attack, and Sicily and Naples were occupied by the Spaniards. In 1735 a Russian relieving corps reinforced the Habsburg front on the Rhine, and in northern Italy also there were a few successful operations of some local importance.

Direct contacts between Austria and France eventually led to the preliminary Peace of Vienna (October 3, 1735). Austria lost Naples and Sicily, which fell to a secondary branch of the Bourbons, and had to cede a tract of Lombard territory to Sardinia. As some compensation, Austria received Parma and Piacenza. Francis Stephen of Lorraine was promised Tuscany but had to renounce his hereditary duchy. On these conditions, France agreed to recognize the Pragmatic Sanction. The final peace was then concluded at Vienna in 1738.

Prince Eugene had died during the War of the Polish Succession. It soon proved disastrous that a successor of similar capacity of the prince was not found. During the second Turkish war of Charles VI (1737–39), Austria had joined in the Turkish–Russian conflict but without coordination of military operations. The Austrians, furthermore, underrated the Turkish forces and were themselves reduced by epidemics. The fortress of Niš was taken but was lost again soon thereafter. Peace negotiations conducted at Nemirov were broken off, and the war went on. The Austrians lost another battle at Grocka. Again peace negotiations were launched, in the course of which the larger part of the gains of the Peace of Passarowitz were lost. More disquieting even than the territorial losses was the loss in prestige. The epoch that had been the rise of Austria to a great power thus ended with reverses.

Social, economic, and cultural trends in the Baroque age. The Thirty Years' War and the Turkish wars had resulted in the devastation of large parts of the country and in great losses among the population, which suffered further reduction during the plague years of 1679 and 1713. The territories that had been wrested from the Turks had to be resettled systematically by German and other immigrants. The initiative for resettlement projects came from the official bureaucracy, the settlements being concentrated mainly in the south of Hungary. During the period of religious conflicts many Protestants had been exiled, but in the 18th century transportation to the various underpopulated parts of the empire was often resorted to.

Mercantilist policies after the 1660s

In the industrial and commercial field, mercantilist ideas, encouraged by the government, were prominent from the 1660s on. The situation of the peasantry was a thoroughly unfavourable one. Tentative measures in the

reigns of Leopold I and Charles VI to protect the peasants had little effect. Certain "model industries" (mostly textile factories) were established but were only partly successful. The economic policy of the absolutist state also resulted in strong interference with trade organizations. The guilds were suppressed or at least debarred from the new manufactures.

Trade was encouraged but yielded only small gains for the state. Industrial and commercial undertakings were managed in part directly by the state but largely through privileged corporations or private persons. Of some importance were the first (1667) and the second (1719) Oriental trading companies and the Ostend Company (1722). Trade in the Mediterranean was also intensified. Promising colonial ventures in India were discontinued for political reasons, however, in the middle of the 18th century. Under Charles VI new roads came to be planned and built on a large scale.

The state was in permanent want of money. This was a period of perpetual war as well as great economic investments, both entailing an excessive strain on state finances. At first the government resorted to the rich bankers such as Samuel Oppenheimer and his successor Samson Wertheimer for funds. Soon, however, it attempted to establish banking firms that were state controlled. The Banco del Giro, founded in Vienna in 1703, quickly failed, but the Vienna Stadtbanco of 1705 managed to survive; the Universalbancalität of 1715 was liquidated after a short period of operation.

After the victory of the Counter-Reformation, education was almost exclusively in the hands of the Catholic Church. The grammar schools of the religious orders, especially of the Jesuits and the Benedictines, set a very high standard for the most part. In 1677 another university was established at Innsbruck, the theological school of which was to acquire some fame. Historical writing flourished, the most outstanding works being those of two Benedictine brothers, Bernard and Hieronymus Pez; Gottfried Bessel, abbot of Göttweig; and Leopold I's official historiographer, the Jesuit Franz Wagner. The Austrian Jesuits were famous for their scientific and geographic researches, taking part in exploring China.

Baroque poetry, theatre, and music

Among the achievements of Baroque poetry, mention should be made of Wolf Helmhart of Hohberg, whose works offer interesting insights into the life of the nobility, and of Katharina of Greiffenberg. The theatre of the Baroque was of great appeal, being remarkable for the splendour of its decorations and the ingenuity of stage machinery. The plays produced ranked from the elaborate Italian opera to the blunt humour of the popular play. Music attained an especially high standard, encouraged by three emperors who were composers themselves (Ferdinand III, Leopold I, and Joseph I). Charles VI was also a skillful musician, and he engaged the services of Johann Joseph Fux, who came from eastern Styria and developed into an important composer and teacher.

Austrian Baroque culture is, however, most clearly revealed by the splendours of its architecture. At first the field was dominated by the Italians, but soon native architects stepped forth. Preeminent was Johann Bernhard Fischer von Erlach (first plan of Schönbrunn Palace, Karlskirche in Vienna, Kollegienkirche in Salzburg) and his son Josef Emanuel (Hofbibliothek). They were rivalled by Jakob Prandtauer (Herzogenburg, Melk, and part of Sankt Florian monasteries) and especially by Johann Lucas von Hildebrandt (Schwarzenberg Palace, Belvedere, Peterskirche in Vienna, monastery of Göttweig). Among native sculptors Georg Raphael Donner was the first in rank and quality of work. Fresco painting was represented by Johann Michael Rottmayr from Salzburg, Daniel Gran from Vienna, and Paul Troger from the Tirolian Pustertal. (E.Zo.)

II. The period 1740–1866

FROM THE ACCESSION OF MARIA THERESA TO THE CONGRESS OF VIENNA

The war period, 1740–63. In October 1740 the Holy Roman emperor Charles VI, the last male Habsburg ruler, died and was succeeded by his daughter Maria Theresa, the wife of the grand duke of Tuscany, Francis Stephen of Lorraine. Until the election of her husband as emperor in 1745, Maria Theresa was referred to only as queen of Hungary and Bohemia. Her descendants represented the House of Habsburg-Lorraine.

Charles VI had established a reasonably unified order of succession for all of the lands under the Habsburg sceptre (the so-called Pragmatic Sanction). By making broad concessions to foreign powers, he had secured their recognition of this act, and he died expecting a smooth succession for Maria Theresa. It was her misfortune that the poor state of Austria's military defenses during the last years of Charles's reign had vitiated his careful diplomatic manoeuvres, and it was of particular importance that four months prior to the Emperor's death another young ruler, Frederick II the Great, had succeeded to the throne of Prussia, which he wanted to raise to great power status. Thus it was that a major German state, which previously had been consistently loyal to the Austrian and imperial cause, became throughout Maria Theresa's entire reign the most determined foe of the Habsburg Empire. The specific issue in the conflict over supremacy in the German orbit was Frederick's intent to wrest the rich province of Silesia from the Habsburgs. The kings of Spain and of Piedmont-Sardinia and the electors of Bavaria and Saxony in full or in part challenged Maria Theresa's claims to succession, despite the fact that they had previously all acknowledged the Austrian heiress' right to rule. Frederick, on the other hand, cared less about the succession than about Silesia.

Conflict with Prussia over Silesia

By October of 1741 Frederick's army, then the best trained in Europe, had occupied Lower Silesia, and Maria Theresa in that month concluded a treaty in which she ceded this major and richest part of the province to Prussia under threats from other quarters. France, the hereditary foe of Austria and main inspiration of the anti-Habsburg coalition, had, in an alliance of Nymphenburg (May 1741), pledged to support Spanish and Bavarian claims on Maria Theresa's inheritance. Charles Albert, the elector of Bavaria, was promised part of the Alpine Hereditary Lands and Bohemia, and, with French support, was elected emperor as Charles VII. On the day of his coronation, Austrian forces occupied his own capital, Munich.

Maria Theresa was helped by the fact that two major European powers, Great Britain and Russia, did not wish to see the Habsburg Empire dismembered. The British and the Dutch soon gave Austria active support, and Saxony and Sardinia dropped out of the anti-Habsburg coalition to support Maria Theresa, whose chances of consolidating herself on the throne thus became more promising.

But the situation remained serious. Habsburg forces still had to face Spanish troops in northern Italy, and in the spring of 1744 France declared war on Austria. One year later, in May, French troops defeated British and Dutch forces at Fontenoy in the Austrian Netherlands (the major part of present-day Belgium).

Frederick II of Prussia, who had watched developments closely after the definitive peace with Austria (Treaty of Berlin, July 1742), was afraid that the Austro–British–Dutch coalition might eventually be victorious and that Maria Theresa might be able to turn her main forces against him and regain Silesia. Stealing a march on the Habsburgs, he invaded Bohemia in August 1744. This time Austrian resistance was more determined, and though Frederick failed in his designs on Bohemian territory, he succeeded in confirming the victory of his first campaign by a new peace treaty, the Treaty of Dresden, December 1745. Having thus secured his Silesian spoils, he could afford to support the election of Francis Stephen of Lorraine as Holy Roman emperor in succession to Charles VII (died January 1745).

Francis Stephen becomes Holy Roman emperor

The war over the Austrian succession continued. As far as the Habsburg power was concerned, the results of the fighting in Italy, Belgium, and Holland were on the whole indecisive. A defensive alliance with Russia (1746) protected Austria from the danger of a new Prussian attack.

During these latter stages of the conflict, the question of the preservation of the Habsburg Empire was no longer an issue. The problem was merely the price that had to be paid for survival. In the long, drawn-out peace negotiations of Aix-la-Chapelle, Wenzel Anton, Graf von Kaunitz (subsequently Prince von Kaunitz-Rietberg, from 1753–92 Austria's state chancellor) showed for the first time his mettle as a diplomat of consummate skill. Austria had to return some minor principalities in Italy to Spanish Bourbons until such time as the Bourbon lines there became extinct. A small frontier rectification in favour of Piedmont-Sardinia also had to be made, but neither of these adjustments was of great consequence. Far more important and painful was, of course, the loss of the major part of Silesia to Prussia, which opened the way to the rise of a new European great power as the most serious rival of the Habsburgs in Germany. It also meant a considerable drop in the number of Germans living within Habsburg lands, with important consequences in the rise of the national problem in the following century. On the other hand, Maria Theresa, a determined woman of courage, moderation, and charm, had managed to establish good relations with the frequently rebellious Magyars and had secured at least their nominal support at the height of the succession crisis.

The Habsburg Empire was not dismembered. Nor did it become a satellite state under the tutelage of other great powers. The outcome of the War of the Austrian Succession, except for the loss of Silesia, was a genuine defensive victory and proved to the world that Austria represented more than an agglomeration of lands under the same rule, acquired by wars and marriage contracts. In the two centuries since Ferdinand I, regent in the so-called Hereditary Lands, had become king in Bohemia, Hungary, and Croatia, a certain cohesion between the major historic units of the empire had clearly, after all, been established.

As far as the worldwide colonial aspects of the struggle were concerned, the results of the war were indecisive; certainly they were not to the advantage of Austria's traditional ally, Great Britain. A more permanent settlement of the conflict was reached only at the Treaty of Paris between Britain, France, Spain, and Portugal in February 1763—by and large to the disadvantage of the French, by then Austria's nominal ally.

During the brief era of peace—the eight years (1748–56) between the end of the War of the Austrian Succession and the Seven Years' War—Austrian military organization was revamped, and a state-wide system of conscription was introduced. The artillery in particular was greatly improved, and at the end of the period the Empress was ready once more to take up the struggle with Prussia. The new state chancellor, Kaunitz, had set an entirely new diplomatic stage for this difficult enterprise. The diplomatic actions taken at that time are generally referred to as the "reversal of alliances," actually a treaty system intended to isolate Prussia. It was in the interests of both Austria and France, hereditary enemies for two centuries, to become allies, since both were concerned about the rapid rise of an aggressive and unpredictable Prussia in the north. In the event of an Austrian reconquest of Silesia, the Austrian Netherlands were to be ceded to France. Russia, worried about Prussia's future designs in the east in regard to Poland, joined the coalition in January 1757, as did Sweden.

Cooperation with France against Prussia

Prussia forestalled the Austrian diplomatic and military preparations by means of a surprise preventive attack in 1756. But it was able to counterbalance the weight of the formidable coalition rallied against it only by a subsidy treaty agreement with Britain, the former traditional ally of the Habsburg Empire. The tremendous superiority of the great coalition in territory and population figures was, however, in part offset by divisions among its members, in whose success Austria had by far the major stake. In a purely military sense, Frederick II had the full advantage of an inner line of defense. His brilliance as a military leader compared to that of the careful but too cautious Austrian commander in chief, Leopold Joseph, Graf von Daun, also counted heavily in his favour. The French army was in full decline, and the forces from the Holy Roman Empire outside of Austria, convoked by the emperor Francis of Lorraine against the Prussian aggressor, failed lamentably.

Though Austrian and Russian victories were sparse compared to a string of Prussian successes, the superior power of the coalition would have inevitably resulted in victory if the new tsar, Peter III, a great admirer of the Prussian King, had not withdrawn in 1762 from the war. Further designs to reverse Russia's position and to side openly with Frederick were foiled by Peter's overthrow a few weeks later. The initiator of the plot, his former consort and now ruling empress, Catherine II, did, however, sign a separate peace with Prussia on the basis of the status quo. Only a few weeks later the French withdrew from the war as well, since the obvious inability of the Austrian troops to regain Silesia killed France's hope for the acquisition of Belgium. In consequence of this, Maria Theresa, in the hardest decision of her reign, was forced to give in. Thus it was that the Treaty of Hubertusberg, concluded with Prussia in February of 1763, merely confirmed the outcome of the two previous Silesian wars. The only minor concession made by Frederick was a pledge to cast the electoral vote of Brandenburg (Prussia) in the next imperial election in favour of Maria Theresa's oldest son, Joseph.

The most decisive result of the Austro-Prussian Seven Years' War was the rise of Prussia to great power status. Austria, though weakened, remained a great power, and compensatory acquisitions for the loss of Silesia were impending. But, taking a long-range view, the Prussian victory represented a decision in the first round of the struggle for supremacy in Germany between the Habsburg Empire and Prussia, a conflict that the Habsburg Empire was to lose decisively within a century.

Foreign policy, 1763–92. In 1772 Austria participated in the first partition of Poland and acquired Galicia. The initiative for the partition came from Frederick II. Catherine II the Great of Russia would have preferred to have an undivided satellite Poland as her neighbour, but Maria Theresa, anxious to prevent a further shift of the balance of power to Austria's disadvantage, participated in the partition, though she regretted the breach of all traditions of international order. In 1774 the Habsburg Empire was given the Bukovina to the southeast of Galicia from the Ottoman Empire as reward for Austrian mediation in a Russo-Turkish conflict. These enterprises were due in part to the influence of the young Joseph II, who, after the death of his father, Francis, in 1765, succeeded as shadow emperor of the Holy Roman Empire and became co-regent with his mother in the Habsburg lands.

Austria and the first partition of Poland

Joseph's ambitions went much further. In 1778, after the older line of the House of Wittelsbach had expired, he came to an understanding with the head of the younger Palatine line, according to which the Habsburgs would become heirs to the rule of Bavaria. Frederick II quite naturally opposed this agreement, which would much more than offset the loss of Silesia and would have given the Habsburgs renewed predominance in Germany. Joseph refused to yield, even though he was only lamely supported by his mother, and a new confrontation became inevitable. Military operations initiated in Bohemia in the summer of 1778 remained indecisive, however, and in May 1779 the Treaty of Teschen was negotiated between Maria Theresa and Frederick, both then in their old age and reluctant to fight another major war. Thus, Joseph had to put off his ambitious scheme, and the cession of the Innviertel by Bavaria, now incorporated into Upper Austria, had to serve as meagre consolation. Though Joseph's original plan would, in fact, never have been tolerated by the European great powers, in 1785, five years after the death of the great Empress, her intemperate son and successor reverted to his old scheme, this time proposing the exchange of Bavaria against the Austrian Netherlands. Again Frederick thwarted the plan, and Joseph's designs to undo the balance established by the Silesian wars were put to rest.

Joseph's foreign policy lacked success in other respects,

too. His eagerness to barter the noncontiguous Austrian Netherlands for Bavaria resulted in part from the fact that he was frustrated in his understandable designs to undo the restrictions imposed on the Austrian rule there by the peace treaties of Utrecht and Rastatt of 1713 and 1714. These restrictions included the closing of the Scheldt River by the Dutch and Dutch rights to garrison the Austrian border fortresses against France at Austrian expense. But Dutch and French objections blocked Joseph's plan.

Failing thus in the west, Joseph hoped, as Russia's ally, to make at least some gains in a new Austrian war against the Turks. This struggle, started in 1788, ended the year after Joseph's death (1791), again in a meagre compromise. Thus the Emperor, whose great gifts included skill neither in diplomacy nor in military strategy, was in his conduct of foreign affairs far more unfortunate than most of his predecessors, who were not in other respects his intellectual peers.

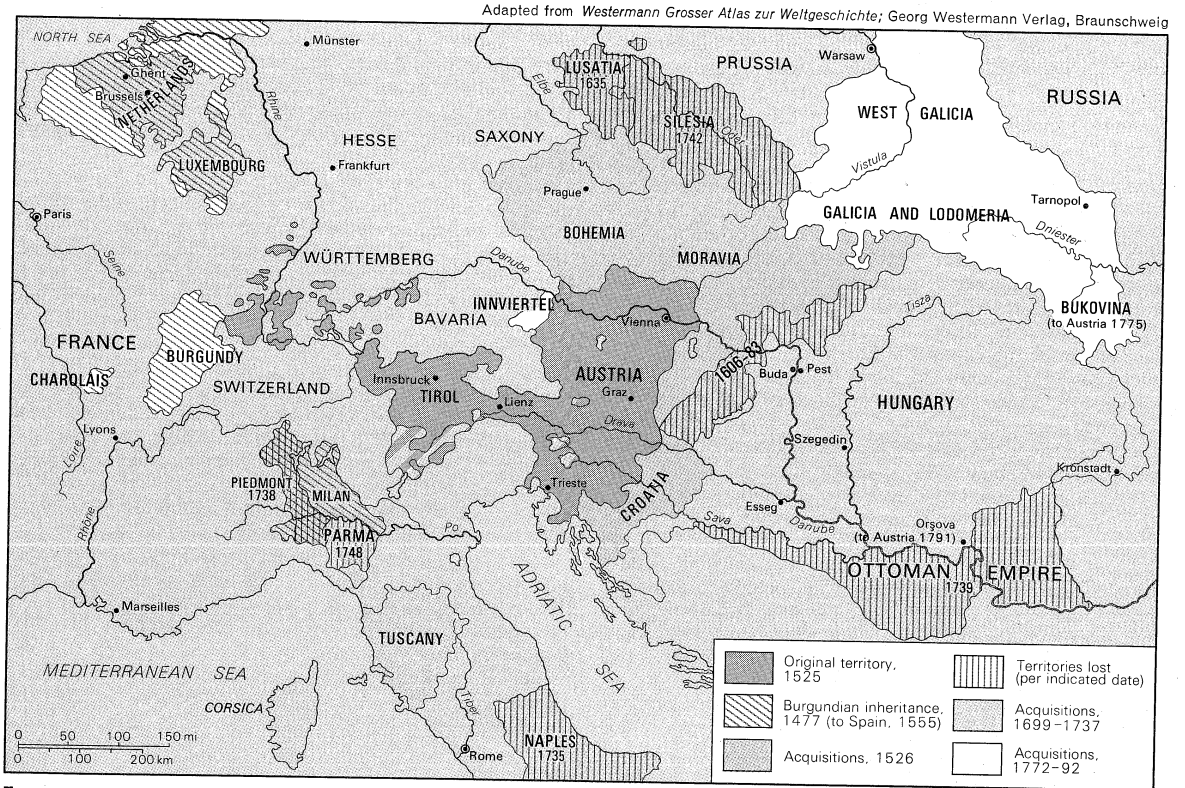
Accession of Leopold II, 1790

His brother and immediate successor, Leopold II, heretofore Grand Duke of Tuscany, certainly was Joseph's intellectual peer, particularly in his ability as diplomat. He inherited from his brother a domestic situation in which Hungary and Belgium were in full revolt. In international relations he had to carry the mortgage of the sterile Turkish War begun by his brother, and he was faced as well by the possibility of a new partition of Poland, which might exclude Austria as beneficiary. But above all he had to deal with the spectre of an ever more serious revolutionary situation in France, which might well involve Austria in a long war of unpredictable outcome.

In domestic matters, Leopold, an enlightened ruler in Tuscany, had to retreat cautiously and regretfully from some of the reform legislation in the Habsburg Empire. In foreign affairs he had to end the Turkish War by the best possible compromise. As to the Polish question, encouraged by the spirit of reform in truncated Poland, he wanted to avoid a new partition altogether, but at the same time he wished to keep relations with Russia and Prussia on an even keel. This included, of course, an understanding in regard to the foremost international problem of his reign—the relationship to revolutionary France. Though naturally opposed to the revolutionary spirit, Leopold's course was remarkably free from considerations arising from close family ties, specifically the fate of his brother-in-law, King Louis XVI, and sister, Queen Marie-Antoinette. Leopold approved of the concept of a French constitutional monarchy, but to be prepared for all contingencies, in July 1790, he secured an understanding with Frederick William II, the new king of Prussia, concerning common interests of both countries in the east. At this time relations with France were aggravated further by the French seizure of the estates of great nobles in the Alsace, since they could be considered subjects of the Holy Roman Empire as well. The establishment of the headquarters of anti-revolutionary French political refugees in the German Rhineland created additional tension. In a second meeting with Frederick William II at Pillnitz in Saxony in August 1791, the two rulers publicly expressed their concern with the French situation. But joint intervention would depend on British and Russian support. In February 1792 the Emperor considered it opportune to go one step further and concluded a defensive alliance with Prussia. Whether he would eventually have taken the final step and gone to war with France is conjectural; on March 1, 1792, he died unexpectedly. Less than two months later Austria did find itself at war, and while the declaration to that effect was made by France, the new emperor, Francis II, Leopold's son, a young man of modest abilities, had done little to prevent it.

The struggle with France, 1792–1815. In the ensuing era of almost continuous warfare that lasted for nearly a quarter of a century, the Habsburg Empire was involved more heavily than any other continental European power. This was partly because of its geographic location in the centre of Europe, partly because of Francis' position as Holy Roman emperor until 1806. The gradual transition from enlightened regimes of Joseph II and Leopold II to an increasingly conservative spirit under his rule was of additional significance. At the beginning of the war period Austria fought the spirit of the French Revolution. Supported by Spain, Portugal, and, above all, Prussia, it was engaged in the War of the First Coalition from beginning to end (1792–97). Yet Prussia dropped out of the war in 1795 to prepare for the third partition of Poland, after Prussia and Russia had jointly excluded

Adapted from *Westermann Grosser Atlas zur Weltgeschichte;* Georg Westermann Verlag, Braunschweig



Expansion of the Austrian Habsburg domains until 1795.

Austria from the second partition. The Austrians fought against the French revolutionary armies in Belgium, Holland, and the Rhinelands with indifferent success. While the emperor Francis' far abler younger brother Archduke Charles checked the French in southern Germany, the young Bonaparte routed the Austrians in Italy, crossed the Alps, and invaded Carinthia and Styria. In April 1797 Austria had to agree to the armistice and Preliminary Peace of Leoben and in October of the same year to the permanent one of Campo Formio. The Habsburg Empire was let off relatively lightly. While the Austrian Netherlands (an indefensible province anyway) and Lombardy had to be ceded, territories of the Republic of Venice to the east of the Adige were gained. This forced participation in the destruction of the ancient Venetian republic put Austria in no better light than Prussia and Russia in the second partition of Poland. On behalf of the Holy Roman Empire, Francis also had to cede the left bank of the Rhine to France and accept a scheme by which the German ecclesiastical princes were deprived of their secular powers (Congress of Rastatt, 1797–99). This action further weakened Francis as Holy Roman emperor, although his position as ruler of the Austrian lands was strengthened somewhat by the third and final partition of Poland in 1795, in which Austria gained West Galicia as far as the Bug River.

Gains and losses at the Peace of Campo Formio, 1797

In the War of the Second Coalition (1798–1802), three major European powers—Prussia, England, and Austria—supported by Portugal, Naples, and the Ottoman Empire, represented the anti-French alliance. The theatres of war were again southern Germany and Italy, and, this time, Switzerland. When, in 1800, the Russian armies withdrew, the Austrian position became precarious. French advances in northern Italy and southern Germany were followed by the Armistice of Steyr in December 1800 and by the Treaty of Lunéville in February 1801. According to the terms of the latter, the provisions of the Treaty of Campo Formio of 1797 and the cession of the left bank of the Rhine to France were confirmed. The system of French satellite republics in Holland and Italy had to be recognized. Thus the main Habsburg territories were left intact, but Austria's position in Germany was further weakened.

Austria's chances in the War of the Third Coalition (1805–07) at first seemed promising. There was genuine military cooperation with Russia, and the coalition was able to count on the naval and financial support of Britain. Prussia, however, delayed a decision concerning its participation. The Austrians were defeated in Germany and the critical situation thus created forced their retreat from Italy as well. In November 1805 French troops occupied Vienna for the first time. On December 2 the Austrians and Russians were badly defeated by Napoleon at Austerlitz in Moravia, and the harsh Treaty of Pressburg was imposed on Austria hardly three weeks later. According to its principal provisions the Venetian territories gained in 1797 had to be ceded together with Tirol, Vorarlberg, Brixen, and the Trentino. The confirmation of Austria's incorporation of Salzburg and Berchtesgaden (1803) could not serve as adequate compensation for these severe losses. The Habsburg Empire had ceased to be a great power, and the Holy Roman Empire, in the face of French advances toward the east, was doomed.

Battle of Austerlitz, 1805

It had not been difficult to foresee these developments when Bonaparte proclaimed himself emperor of the French in 1804. To maintain Austria's position as best he could, and anxious to preserve the imperial title for his house, Emperor Francis proclaimed himself emperor of Austria on August 14, 1804. This purely declaratory manifesto, never submitted for the consent of the Estates of his lands as the Pragmatic Sanction had been a century before, pertained to all the Habsburg realms and might well therefore have been challenged, above all by Hungary. Yet this was not a major issue in the turmoil of the times. A month after Napoleon established the Confederation of the Rhine, Francis accordingly abdicated the empty title of Holy Roman emperor (August 1806). Invested only with an imperial title devoid of a great historical tradition, he ruled henceforward as Emperor Francis I of Austria.

The period from 1805 to 1809, when Austria was compelled to join the French continental blockade system, was one that saw the rise of feelings that were truly national, even though they were largely confined to the Austro-German orbit. A new, energetic, and idealistic foreign minister—Johann Philipp, Graf von Stadion—attempted to prepare the diplomatic ground for a renewed struggle against French expansion, and the archduke Charles organized a national militia. Encouraged by continued British resistance and by the guerrilla warfare of the Spanish people against the French occupation, Austria, on April 9, 1809, declared war on France. In contrast to a prostrate Prussia and to the German princes who catered to the whims of the foreign conqueror, the Habsburg Empire—bravely and vainly—tried to redress the balance in Europe. By mid-May 1809 Vienna was in French hands, though this time only after an artillery bombardment. The Austrians fought on, and a week later, at Aspern, on the left bank of the Danube opposite Vienna, Archduke Charles administered the first defeat to a French army commanded personally by Napoleon. He failed, however, to take advantage of this victory, and six weeks later the Austrians were decisively defeated at Wagram. Peace was dictated by the victors at the imperial castle of Schönbrunn in October 1809. This time Salzburg, a part of Upper Austria, and northern Tirol became parts of Bavaria; southern Tirol became part of the satellite kingdom of Italy, and West Galicia part of a new puppet Grand Duchy of Warsaw. The western parts of the Austrian southern Slav territories were ceded to France, and a heavy indemnity had to be paid.

A by-product of the defeat was the crushing of the heroic rising of the Tirolean peasants and the execution of their brave leader, Andreas Hofer. A far-reaching economic consequence of the war period was the state bankruptcy of 1811, which reduced the Austrian currency to one-fifth of its previous value.

Now a remarkable change in Austrian foreign policy took place. Shortly before the conclusion of the Treaty of Schönbrunn, Count (later Prince) Clemens Metternich, formerly ambassador to Napoleon's court, was appointed minister of foreign affairs in place of Stadion, whose anti-French policy had failed. Metternich conducted foreign policy for almost 40 years, until the outbreak of the Revolution of 1848. His first task was to shift from an anti-French policy, based partly on ideological grounds, to one of expedient cooperation until the day when new opportunities to restore Austria to its former rank as a great power would present themselves. He pursued this policy with great skill. One of his first moves was to persuade his imperial master to agree to Napoleon's request for the hand of Francis' eldest daughter, Marie-Louise (1810). Humiliating as acceptance of this demand by an upstart ruler was to an ancient dynasty, Metternich considered it conducive to Austrian interests.

Metternich becomes minister of foreign affairs

In the War of 1812, Austria—France's new ally by marriage—was forced to put up an auxiliary corps under Karl Philipp, Prince zu Schwarzenberg, but Metternich's instructions were that any significant encounter with the Russian forces should be avoided. After the annihilation of the great French army in Russia and the continuation of the war in Germany by Russia and Prussia with the support of England and Sweden, the Habsburg Empire managed to stay neutral, thus raising the price for its subsequent intervention. It was clear that such intervention could only be on the side of the anti-French alliance, but Austria first offered Napoleon its "armed mediation," provided Napoleon would give up his overlordship of the German territories on the right bank of the Rhine and dissolve the Grand Duchy of Warsaw. The demand for the return of all Austrian territories ceded to France in the Treaty of Schönbrunn was raised. Napoleon, as expected, rejected the offer, and in August 1813 Austria became a partner in the War of Liberation, although the national issues previously raised by Stadion were played down by the conservative Metternich.

By its belated entry into the grand coalition, Austria had secured particularly favourable terms. Even though the Habsburg forces were smaller in number than those of Russia and Prussia, Prince zu Schwarzenberg was appointed commander in chief of the allied armies. Although a strategist of no better than average ability, he proved to be a skillful moderator among the ambitious generals of the coalition. His much abler chief of staff, Joseph, Graf Radetzky, must be credited with drawing up the plans for the great Battle of Leipzig (October 16–19), which broke the back of the French operations in Germany. Napoleon rejected an offer on the part of the allies that would have conceded the Rhine and Alp frontiers including Belgium and Holland to France. It is probable that Metternich at this time would have liked to keep a greatly weakened Napoleon in power as the best bulwark against the danger of a new French revolution.

On New Year's Eve, 1813, the war was carried into France. Had it not been for the tremendous superiority of the allies and the exhaustion of the French people, Schwarzenberg's indifferent strategy would hardly have secured victory. At it was, the allies were able to take Paris by the end of March 1814, force Napoleon's abdication, and secure the restoration of the Bourbons. The Treaty of Paris of May 30, 1814, gave France the relatively favourable frontiers of 1792. The reorganization of Europe was to be arranged at a congress that was to meet in Vienna in September 1814 under Metternich's chairmanship.

Austria and the Vienna Settlement, 1815

In the negotiations, which lasted until June 1815, Austria sided in general with Bourbon France and with England. Austria regained most of the territories that it had lost in the Treaty of Schönbrunn. Its former rule in the Austrian Netherlands and in southwest Germany was not restored, since Metternich saw little advantage in the control of noncontiguous areas. The Habsburg Empire was compensated in Italy: not only were Lombardy and the Venetia recovered; Tuscany, Modena, and —for the lifetime of Marie-Louise—Parma and Piacenza, also, were established as Habsburg semi-independent appendages in Italy. Napoleon's consort, now separated from him, was to rule in these two last-mentioned principalities. It is questionable whether the restoration of the Holy Roman Empire might not have served Austria's interests better than the establishment of the German Confederation (June 1815) under Austria's presidency.

In line with Emperor Francis' and Metternich's wishes, the confederation gave Austria, though only jointly with Prussia, far-reaching control over German affairs. This control was aimed in particular at preventing the establishment of representative constitutional governments and liberal institutions in any of the 50 German states. Reestablishment of the Holy Roman Empire would have struck a strong chord in the heart of the Austro-German peoples but this was precisely what Metternich, fearful of any expression of the popular will, did not want. Following the same line of thought, he believed that the dismemberment of Italy would prevent the future unification of the country under liberal national banners.

Metternich's skill restored Austria to great power status and even resulted in his diplomatic leadership as the so-called coachman of Europe. But the two major issues that were to severely weaken the position of the Habsburg Empire within the next half-century—the German and the Italian problems—were accentuated rather than solved by the settlement of 1815.

Reforms and their reversal, 1740–1815. The enlightened era in the Habsburg Empire comprises the reigns of Maria Theresa (1740–80) and those of her two brilliant sons, Joseph II (as co-regent from 1765 to 1780 and in his own right from 1780 to 1790) and Leopold II (1790–92). Maria Theresa, much better adjusted to reality than her immediate successor, drew chiefly on pragmatic experience. She wanted to secure the well-being of her subjects and to affirm the control of the state as far as compatible with traditions and customs. Joseph II intended to move faster and more in line with abstract, enlightened principles. He had little patience with references to tradition and was in some respects less prejudiced than his mother. Leopold II was, in his basic views, closer to his brother than to his mother, whom he resembled in his judicial temperament. While he would presumably have liked to expand Joseph's reforms, it was the main task of his all-too-short reign to quell the revolts in Hungary and the Austrian Netherlands and to settle the unrest in the Hereditary Lands and Bohemia. Thus he was forced to retreat in essence from the reforms of Joseph II to those of Maria Theresa. Under his son, Francis (1792–1835), the trend gradually became one of extreme conservatism and outright reaction, in part because of the preferences of this mediocre ruler but largely as a result of the counter-revolutionary spirit of the times.

Administrative reforms

The basic idea of the administrative reforms was to transform the estates system into a partially bureaucratic administration based on civil service rules, although Maria Theresa considered it expedient to preserve at least the external, but by then rather hollow, shell of the estates structure. Thus, on the provincial level, the speakers of the crownland Estates presided over most of the agenda of the provincial administration. On the top level of administration for the whole empire, the state chancellery of old was divided into a court chancellery entrusted with domestic agenda and a state chancellery proper for foreign affairs. An advisory state council was composed in part of great nobles and in part of high bureaucratic officials. A directory on public affairs for the Bohemian and Hereditary Lands was subsequently converted into the United Austrian and Bohemian Court Chancellery. A new commerce directory for the whole empire was established.

Maria Theresa introduced restrictions on the largely arbitrary patrimonial jurisdiction of the lords on their estates and transferred venue in matters of capital offenses to fewer and better courts. She steadfastly opposed the abolition of torture, though not on account of any basic cruelty of her character. She simply could not imagine how sufficient evidence could be gathered without forced confession. It was not until 1776 that she was finally persuaded to abolish torture. In this respect, Austria had lagged behind France, Prussia, and even Russia.

Joseph, who had much more understanding of judicial questions than did his mother, did much to improve civil and criminal procedure. He provided free legal counsel for peasants in litigation with lords. Although criminal justice administered on the basis of the strictly utilitarian philosophy of the Emperor was still extremely harsh, Joseph did much to improve civil law. The superb code of civil law of 1811, initiated in substance under Joseph, came to fruition only under Emperor Francis.

Economic policy

As to public finance, both the nobles and the church lost various privileges in regard to exemption from taxation. Craft guild restrictions, particularly those concerning the admission of apprentices, were loosened. Much was done to help the peasants in the newly gained trans-Carpathian territories (Galicia and Bukovina) by direct and indirect government assistance, the abolition of customs duties for exports to other crownlands, and by the granting of various privileges for new settlers. The basically mercantilist economic policy of Charles VI's reign was somewhat changed in line with the influence of physiocratic and so-called populationist theories. Henceforth, skilled human labour, and not precious metal, was gradually to become the yardstick of national wealth. This led, on the one hand, to restrictions on emigration, and, on the other, to improvements in vocational training. Severe import restrictions on so-called luxury goods, which frequently had killed demand rather than encouraged the rise of competitive domestic industries, were somewhat modified. Joseph II went further. In regard to industrial and commercial enterprise, he was, much in contrast to his overall political philosophy, a strong supporter of private initiative.

Maria Theresa went a long way toward alleviating the

lot of the unfree peasants. For the first time in Austrian history, the service obligations of the peasants were strictly defined. Above all, the ambiguous double role of the lord as landowner and official of local government was put under government supervision.

Joseph expanded the reforms of Maria Theresa. He further reduced the governmental functions of the lord and by legislation of 1781 abolished serfdom altogether outside of Hungary. Similar legislation was introduced in Hungary between 1784 and 1786. The mistaken belief of the Hungarian peasants that abolition of serfdom made them free owners of the land was, however, a major cause of the revolutionary situation at the time of the Emperor's death. The full conversion of the peasant's personal services outside of Hungary into obligations to be paid off in cash was somewhat precipitate, for money, in an agricultural economy still largely based on barter, was lacking. Leopold II reversed Joseph's peasant legislation so that things stood much as they had been at the time of the death of Maria Theresa. There was little further change until 1848.

Educational reforms of Maria Theresa

The greatest single achievement of Maria Theresa's reign was her educational reform at the elementary and intermediate levels. Three types of schools were introduced: (1) elementary schools in all but the smallest villages, in which reading, writing, and arithmetic were taught, with attendance being compulsory; (2) district schools in every administrative district, in which history, more advanced study of the vernacular language, geometry, drawing, and some vocational training were offered; and (3) so-called normal schools, established in the crownland capitals. Meant to be terminal schools for the urban middle class, they served also as teacher-training institutions. As for higher education, reforms under Maria Theresa were handicapped by her mistrust of—as she perceived them—radical enlightened ideas. She was, however, persuaded to transfer censorship from Jesuit control to a new and somewhat more liberal state agency. Some new chairs in the natural sciences and law were established at the University of Vienna, but progress in this respect was limited, and non-Catholics were altogether barred from graduating. Maria Theresa also founded some outstanding special schools. One of them was meant to train young nobles for public service, another, officers for the armed forces.

Joseph's reign proved to be disappointing as far as education was concerned. Censorship, it is true, was eliminated—at least as far as criticism of the government was concerned. But the Emperor's utilitarian objectives precluded practically any purpose in higher education other than the training of civil servants. The publication of literature that could not serve as a textbook for disciplines of immediate practical use was not encouraged. Under Emperor Francis, censorship was reintroduced in full force.

Radical church reforms are usually associated with Joseph's reign, but here also Joseph followed in the footsteps of his mother. Both rulers were devoutly religious, but both believed in firm state control of ecclesiastical matters outside of the strictly religious sphere. Following populationist doctrines, the Empress ordered restrictions of religious holidays and the prohibition of ecclesiastic vows prior to the 24th birthday. She insisted that clerics were subject to the jurisdiction of the state in nonecclesiastical matters. The acquisition of land by the church was to be controlled by the government. She took action against the Society of Jesus (Jesuits), but only in 1774, after the Pope had ordered its suppression.

Joseph's most radical measures were the issue of the Edict of Tolerance of 1781 and his monastic reforms. The edict and the legislation attached to it gave Protestants near equality and gave Jews the right to enter various trades, as well as permission to study at universities. In this respect the difference between the Emperor and his mother was fundamental. While Maria Theresa viewed Protestants as heretics and Jews as the embodiment of the Antichrist, Joseph fully respected other Christian denominations and entertained secret plans to establish an Austrian state church independent of Rome.

As to the monasteries, Joseph held that institutions not engaged in useful work for the community, above all agriculture, care of the sick, and education, should be dissolved. Consequently, about a third of the Austrian monasteries ceased to exist, their former members being ordered to learn skills adapted to secular life. The property of the dissolved institutions was used to pay for the upkeep of parishes and to finance the establishment of new parishes.

Control of church discipline and church property were further tightened by Joseph; seminaries for the training of the clergy were secularized. He even tried, without success, to simplify radically the Catholic liturgy. Many of his religious policies were discontinued in the reaction that followed, but the Edict of Tolerance and the monastic reform were maintained.

Joseph was unsuccessful in his efforts to achieve empire-wide administrative centralization. In particular, his attempt to enforce the use of German as the language of administration in Hungary was a failure and certainly accelerated the growth of Magyar nationalism and anti-German feeling there. On his accession, Leopold II was forced once more to recognize Hungary as a separate unit of the Habsburg lands.

THE AGE OF METTERNICH, 1815–48

Foreign and domestic policy. Austria's leading diplomatic position after the conclusion of the second peace treaty of Paris in November 1815 was never anchored primarily in the power potential of the Habsburg Empire but in the skill of Metternich and his adviser, Friedrich von Gentz, in establishing a common conservative platform acceptable to east and west. Its ideological foundation was meant to be the Holy Alliance of the European powers of September 1815, as proposed by Tsar Alexander I and as opposed by the British government. It was meant to organize Europe on the basis of Christian authoritarian principles, implying the possibility of intervention by foreign powers in revolutionary or even merely liberal movements abroad.

The Concert of Europe

In the ensuing period of the Concert of Europe (an attempt to establish a directorate of the great powers over European affairs), which saw limited cooperation among the great powers at the conferences of Aix-la-Chapelle (1818), Troppau (1820), Laibach (1821), and Verona (1822), France was commissioned to put down a revolutionary rising in Spain and Austria was to suppress a revolt against the brutal oppression by the Bourbon regime in the Kingdom of the Two Sicilies. These actions caused Britain to withdraw from the concert of the five great powers. The shattering blow to Metternich's concept of the anti-revolutionary unity of the European powers, however, was Russia's support of the Greek independence movement between 1821 and 1830. Therewith his system in international relations had for all practical purposes broken down and could not be restored but merely patched up.

Two immediate problems for Metternich were the maintenance of Austria's hold in Italy and the struggle with Prussia for supremacy in Germany. As to Italy, the secret Carbonari Movement and the liberal Young Italy promoted by Giuseppe Mazzini were in full swing throughout the whole Restoration period. The French July Revolution of 1830, another defeat of the Metternich system, encouraged open revolt in the Austrian appendages of Parma and Piacenza and in the Papal states. Austrian troops had to restore the old order by force of arms.

The handling of the Italian question, while it did not really undermine Austria's great power position, gave Austria the image of a tyrannical oppressor of a freedom-loving, highly cultured people. That the Austrian administration in Lombardy-Venetia was neither corrupt nor by contemporary standards particularly inefficient counted little in Austria's favour.

Even more serious was the German problem, which involved issues of real power for the whole Habsburg Empire. Such issues had great emotional impact on the Austro-Germans in general and in particular on the liberal

intelligentsia. In their rejection of liberal, potentially revolutionary trends as dangerous to their rule, the German princes saw eye to eye, but the Emperor of Austria and the King of Prussia were deeply divided as regards the issue of supremacy in Germany.

The Carlsbad Decrees, 1819

The Wartburg Festival (1817) of German academic youth celebrating the tercentenary of the beginning of the Reformation had worried the princes deeply. The assassination of the playwright August von Kotzebue as an alleged Russian spy by a radical German student gave the conservative powers the opportunity for which they had been waiting. Under the leadership of Metternich and with the active cooperation of the increasingly reactionary Tsar Alexander, the Carlsbad Decrees were passed in August 1819. These put the German and Austrian universities under strict government control. Student associations were forbidden, censorship was strengthened. An investigatory commission was set up in Mainz, and teachers, writers, and students suspected of liberal views were blacklisted throughout Germany and Austria. In 1824 the German Confederal Assembly in Frankfurt renewed these provisions at the instigation of Metternich for an indefinite period. New oppressive measures on an even larger scale were again introduced at Metternich's behest by the German Confederation as answer to republican demonstrations at Hambach (Palatinate) in 1832 and at Frankfurt in 1833. The restrictions were in essence still in force at the time of the outbreak of the Revolution of 1848.

In the conflict between the two leading powers in the confederation, Prussia took the lead first in the economic field. In 1819 Prussia and its largely noncontiguous domains were merged into one customs association. Treaties of adherence by small neighbour states followed. By 1829 most German states had joined the association. There were some major exceptions—most notably Austria. The whole plan was indeed directed primarily against Austria. Counter movements such as a south German customs union and a more limited central German one collapsed. When, on January 1, 1834, the German customs union (the Zollverein) was completed, Frankfurt, Baden, and the Hansa cities stayed outside with Austria. Prussia had however won an important political as well as economic victory. It made the German north and west (the Rhineland) the centres of gravity of further industrial development.

Metternich's decline

At this time Metternich's power was not only externally but also internally in decline. In 1824 Franz Anton, Graf von Kolowrat, a great noble of Czech origin, was appointed minister of state. Somewhat more enlightened than Metternich, particularly in regard to the Austrian Slavs, he carried great weight in matters of domestic administration. The situation changed further when in 1835 the emperor Francis died and was succeeded by his eldest son, the feeble-minded crown prince Ferdinand. Metternich believed that even a feeble-minded prince could not be bypassed, but the actual power of the government was transferred to a state conference consisting of two archdukes (both of limited intelligence) and Kolowrat and Metternich as permanent members.

Metternich's last success occurred in 1846 when a revolt, initiated by Polish nobles in Galicia, was suppressed after the peasants, indirectly supported by the government, had turned against their oppressive masters. Because the revolutionary ferment had been fed from the city republic of Cracow, Russia and Prussia agreed that Austria should incorporate the ancient Polish coronation city into Galicia.

REVOLUTION AND COUNTER-REVOLUTION, 1848–59

The revolutions of 1848–49. Metternich's victory was short-lived. Early in March 1848 a revolution began in Austria as soon as news about the success of the revolution in France had spread. In Vienna, crowds of people led by students and young academicians asked for liberalization of the regime. A clash with the military led to bloodshed. The riots, in which workers participated, thereupon spread further. The demand for the aged Metternich's resignation was met by the crown at once, for it was expected that after this symbol of reaction was overthrown, order would quickly be restored. But encouraged by this first and unexpected quick success, the revolution spread across the empire.

Three main aspects of the Austrian Revolution may be distinguished: social, democratic liberal, and national or multi-national.

As to the social revolution, the Habsburg Empire was less industrialized than any major European state west of Russia. Thus, while workers participated courageously in the risings in Vienna, the impact of their intervention was small and had little effect on governmental policies. There existed a far stronger movement for the full emancipation of the peasants, which meant giving them clear title to the lands without any further obligation of service to the lords. The newly elected constituent assembly in Vienna, in September 1849, passed legislation to this effect. It was sanctioned by the Emperor, but the actual enforcement of this sweeping reform, in particular in regard to the financial indemnity to be paid by the peasants, was left to the postrevolutionary regime. Still, the influence of peasant emancipation on the revolution was a powerful one. The largest and basically most conservative social class appeared satisfied and by the autumn of 1848 was no longer a major factor in the revolution.

Concerning the liberal democratic revolution, a moderately conservative new cabinet had by April 1848 drawn up a new constitution usually referred to as the Pillersdorf Constitution, after the name of the minister of the interior. This document was, on the whole, fairly democratic. At about the same time, a liberal government under Count Lajos Batthyány was formed in Hungary, in which some of the great leaders of the nation and the dominant figure of the more radical reform policy, Lajos Kossuth, participated. These men drew up a new constitution for Hungary. It provided far-reaching home rule for the country and, except for two provisions, was agreed to by the Emperor. The unresolved provisions were a demand for a separate budget and a more important demand that Hungarian troops not be called to action without the approval of the Hungarian government.

Developments in Hungary

In the meantime František Palacký, the great Czech historian and recognized intellectual leader of his people, had in a public letter rejected the participation of Czech representatives in the first freely elected German national assembly, which was to convene in May 1848 in Frankfurt. While the lands of the Bohemian crown were legally a portion of the territory of the Habsburg Empire that belonged to the German Confederation, Palacký held that the Czechs, who strongly affirmed the existence of a multi-national Austria as a bulwark against Russian advances to the west, could never be part of Germany.

German Austrian liberals and enlightened conservatives participated gladly in the National Assembly in Frankfurt. One of the few enlightened, though not liberal, members of the imperial house, Archduke John, was elected temporary regent of the new Germany, while a moderate Austrian liberal, Anton Ritter von Schmerling, was appointed prime minister in September 1848. Inasmuch as the Frankfurt assembly had no executive power, these positions meant little in practical terms. The assembly immediately began to work on the draft of a constitution, and here the Austrian question represented one of the foremost divisive issues. The Grossdeutsch (Large German) faction wanted to include the Austrian Germans but was opposed to any overall association with the non-German majority of the Habsburg Empire. The Kleindeutsch (Small German) faction thought the only way to avoid this difficulty was to exclude Austria altogether. The solution adopted in the German federal constitution in May 1849 affirmed the incorporation of the German Austrian lands as long as they had a constitution and administration separated from that of the non-Germanic parts of the Habsburg Empire. But the practical effect of these provisions was nil, because by the time the constitution was adopted the counter-revolution was in full swing. Hardly a month later the Austrian deputies at Frankfurt were recalled. The Prussian government fol-

lowed suit and about a month later the remainder of the parliament, which had been transferred to Stuttgart, was dissolved by simple police action. In essence this meant the end of the revolution in Germany, but not the end of the Austro-Prussian conflict as it had developed during the revolution.

The Prussian government wanted to exclude Austria in a roundabout way from the confederation. A union scheme was therefore proposed by Prussia and put before a new, no longer revolutionary, assembly at Erfurt. Austria did not participate in these proceedings. According to the Erfurt plan, supported for a time by Saxony and Hanover, the German Confederation should be divided into two sections, an inner German section under the leadership of Prussia and a second, purely nominal, section consisting only of Austria in alliance with the inner section. The Austrian government, under its energetic prime minister, Felix, prince zu Schwarzenberg, rejected this plan and forced the states that supported Prussia to withdraw from the scheme. Finally, the old confederation was again called into session at Frankfurt. Schwarzenberg would in fact have liked to create a central European union including the whole of the Habsburg Empire. But this goal, which would have made Austria supreme in the whole of central Europe, was unobtainable.

Provisions of the Erfurt plan

In deciding to take his stand on a relatively minor issue, the question of the right of Prussia or Austria to intervene in a revolutionary situation in Electoral Hesse, Schwarzenberg forced Prussia to withdraw and hence to recognize the restoration of the Confederation of 1815 under Austria's presidency (the Punctation of Olmütz, November 1850). This was the last, and a merely diplomatic, victory of Austria over Prussia. It was achieved only because Tsar Nicholas I backed the Habsburg Empire, which seemed to him the more reliable conservative power. The Austrian triumph proved to be short-lived. It merely strengthened the Prussian resolve to undo the humiliation as soon as an opportunity offered itself.

Such an opportunity was not long in coming, since the lack of cohesion within the Habsburg Empire had become painfully obvious during the course of the revolution. While riots in Prague, Lwów, and Cracow could be quelled, events in Hungary took a collision course. The new government there, though eagerly introducing educational and social reforms, was at the same time set on a policy of rapid Magyarization. The result was risings among non-Magyar Hungarian national groups—Croats, Serbs, Slovaks, and Romanians. The conflict with the government in Vienna stiffened when the diplomatic Batthyány was replaced by the far more radical Kossuth, in September 1848.

A further aspect to be considered was Austria's involvement in a war with Piedmont-Sardinia, whose king, Charles Albert, intended to drive the Austrians out of northern Italy. His army invaded Lombardy in March 1848. By August, however, the Sardinians were decisively defeated in the Battle of Custoza. An armistice was concluded in late summer of 1848, but the war, because of violation of its provisions, was resumed by Sardinia in March 1849. After the new attack was repelled, peace was restored on the basis of the status quo in August 1849. Although the Sardinians were no match for the Austrian army, the Italian campaigns made it impossible for Austria to deal with the Hungarian revolution effectively.

The Slav problem

In the meantime, the Slavic question had become a major issue in the western part of the empire as well. In June 1848 a Slav congress was convoked in Prague, in which Czechs, Croats, Poles, Ruthenians, Serbs, Slovaks, and Slovenes were represented. Before the congress could terminate its deliberations, street riots offered the government a pretext to terminate the proceedings. Even though a final program of action could not be agreed upon, the results of the Prague Congress were of great importance in the history of Pan-Slavism. Beyond this, various plans for federalization and local autonomy within the empire were discussed.

By September 1848 the Hungarian situation had deteriorated further. Troops led by the ban (viceroy) of Croatia, Josip von Jelačić, entered the country under orders to restore royal authority. The action created general indignation in Magyar Hungary, and the imperial commander in Budapest was lynched during a riot. Thereupon the Hungarian Reichstag was dissolved by royal decree.

The situation in Hungary encouraged a new revolutionary rising in Vienna, in early October 1848. It was the last major attempt in the German-speaking orbit to regain the revolutionary initiative. The imperial minister of war, Theodor, Graf Latour, was lynched by a mob, and an imperial army under the command of the arch-reactionary Prince Alfred Windischgrätz occupied Vienna, meting out harsh retribution to the revolutionary leaders.

A momentous change took place on December 2, 1848. The feeble-minded Emperor Ferdinand, dubbed by official historiography "the Benign," abdicated in favour of his 18-year-old nephew, Francis Joseph, from whom a more energetic stand against the revolution could be expected. The first prime minister of the young Emperor was Schwarzenberg, a gifted conservative and an utterly ruthless and shrewd statesman, whose advice Francis Joseph accepted more indiscriminately than he did that of any other of his ministers.

Schwarzenberg was even less inclined to compromise in the internal affairs of the Habsburg Empire than in those concerning the German Confederation. Thus, an open break with Hungary occurred in mid-December 1848, and Windischgrätz's army marched into the country on a full wartime footing. Under this mediocre commander, the Austrians did poorly against the effective tactics of the insurgents. The Hungarian Reichstag, encouraged by this turn of events, in April 1849 declared that the Habsburgs had forfeited their right to rule. Hungary was accordingly proclaimed a republic under the presidency of Kossuth.

One factor in the Hungarian decision was the changed situation in Austria. By early March of 1849 the constitutional committee of the Reichstag, after prolonged and thorough discussion on a high intellectual level, had adopted the draft of an Austrian constitution (the Kremsier draft, after the Moravian town in which the Reichstag was reconvened). The draft provided for a constitutional monarchy, giving the sovereign only a suspensive veto against parliamentary legislation. Administrative organization along national lines was provided on the district level, yet crownland administration, the old historic boundaries, and the essence of the centralistic structure of the empire remained intact. A weakness of the draft was that it did not deal with the Hungarian problem. But, all things considered, the bill represented a reasonable compromise between the federalist and centralistic concepts, the former promoted mostly by Slavs, the latter primarily by German liberals.

The Kremsier constitution

The Kremsier constitution was the first reform plan drawn up by genuine representatives of the Austrian peoples, and it was precisely for this reason that the cunning Schwarzenberg decided to foil it. He had the Reichstag dissolved by police and warrants issued for the arrest of some of the reformers. At the same time he had the minister of the interior, Stadion, issue by decree a new, strictly unitary and centralistic constitution for the empire as a whole, including Hungary. The Stadion constitution was more conservative than the Kremsier draft, but it still subscribed to representative government. Enactment was, however, held in abeyance. Thus not only the Austrian peoples felt deceived but Stadion and other members of the Cabinet as well, since the constitution was never put into force and was even formally rescinded in December 1851. In the long history of Austrian government, the dissolution of the Reichstag of Kremsier was one of the most fatal governmental mistakes.

In March 1849 Piedmont-Sardinia had resumed military operations against Austria. This fact and the tense political situation in Germany caused Hungary's military situation to deteriorate rapidly. In April 1848 Windischgrätz was replaced as commander, and after an interim the cruel and ruthless Julius, Freiherr von Haynau, was

appointed as his successor. By mid-May 1849 the Hungarian revolutionary army had reconquered Budapest, presenting the young Francis Joseph with the painful alternative of risking his throne or asking the Tsar for help. Nicholas I, afraid of the possibility of the spread of the revolution to Russian Poland, complied with the Austrian Emperor's request. Russian armies entered Hungary and cooperated with Austrian forces. Budapest was reconquered by mid-July, and on August 13, 1849, the Hungarian troops capitulated. Kossuth and some of his followers managed to flee to Turkey. Even Tsar Nicholas recommended mercy for the gallant Hungarian military commanders. The answer of the Schwarzenberg-Haynau regime was the execution of 14 Hungarian generals—those who had surrendered to the Russians rather than to the Austrians—by hanging. About a hundred other executions, including that of Kossuth's moderate predecessor, Count Lajos Batthyány, followed. These actions as well as many long prison sentences and property confiscations imposed on minor rebels had only a limited deterrent effect, but they revolted public opinion across Europe. Hungary was dismembered and Transylvania, Croatia, and other areas were organized as separate crownlands under strictly authoritarian rule. The prostrate country was divided into five military districts and put under the administration of the stern Archduke Albrecht, whose intervention in the March 1848 Revolution in Vienna had made him hated by all liberals.

Defeat of the revolt in Hungary

All things considered, the revolution across the empire had not accomplished very much. Absolutism was seemingly more firmly entrenched than before, and the political clock had been put back beyond the regime of Maria Theresa. And yet a regime so badly shaken as Austria's could not hope to rule unchallenged in the future. The unresolved social, constitutional, and national issues became more intense, and new changes were soon in the offing.

The neoabsolutist era, 1849–60. The neoabsolutist era in Austria from the breakdown of the revolution in 1849 to 1860 must be judged from the point of view of the domestic policies of the regime—which were controversial but not entirely negative—and of a foreign policy that proved to be of disastrous consequence.

Positive domestic achievements were the establishment of a unified customs territory for the whole empire, distinct progress in industrialization, the promulgation of a code for trades and crafts, and some rudimentary beginnings of social legislation. Enactment of the full emancipation of the peasants, initiated by the revolutionary legislation, worked well. One-third of the cost was to be carried by the former owners of the land, one-third by the peasants themselves, and one-third by the government. This was, on the whole, a solution more equitable than that adopted in Prussia, and in Russia in the 1860s. Some improvements in standards were made in the universities and in the curricula of the Gymnasia. Overall credit for the reform policies belonged largely to Alexander Bach, the successor of Schwarzenberg, who had died in April 1852, lamented by few but the young Emperor.

Emancipation of the peasants

The regime's policies on other matters were more typically reactionary. Freedom of the press and jury and public trials were abandoned; corporal punishment by police orders was reintroduced. Informers flourished: the observation of the liberal reformer Adolf Fischhof that the regime rested on the support of a standing army of soldiers, a kneeling army of worshippers, and a crawling army of informers was exaggerated but not entirely unfounded. One of the most unwholesome developments was the conclusion in 1855 of a concordat with the Holy See that gave the church more power than it had possessed since the middle of the 18th century. Jurisdiction in marriage questions was handed over to the church, as well as control of censorship and elementary education. Church control extended indirectly to secondary education, also, because the priests, who were entrusted with compulsory religious education, had the right to see to it that instruction in any other field, be it physics or history, did not conflict with their teachings. In the second half of the 19th century, a regime of this type could possibly have stayed in power if it had been successful in the conduct of foreign affairs. But this was not to be the case.

Exclusion from Germany and Italy. In the protracted Crimean War crisis (1853–56) the Habsburg Empire took a stand against the Russian occupation of the Danube principalities (Moldavia and Walachia). By threatening military intervention on the side of the Western allies, Austria forced their evacuation by Russian troops and from 1854 occupied the territories itself for the duration of the war. In December 1854, Austria joined the Western alliance but refrained from a declaration of war on Russia and from active military participation. The continuing threat of such action forced Russia to accept the humiliating Treaty of Paris in March 1856.

The consequences for the Habsburg monarchy were grave. Russia, Austria's ally for well over a century, moved into the camp of its enemies. The tsars, Nicholas I and his son Alexander II, never forgave the alleged ingratitude for the Russian intervention in Hungary in 1849. The Western powers, on the other hand, were highly dissatisfied that the Habsburg Empire had refused to take the last decisive step and become a combatant in the war. In retaliation, and in payment of political debts, they supported the cause of Italian unification, skillfully engineered by the Sardinian prime minister, Count Cavour.

The end of Austria's alliance with Russia

Though Karl Ferdinand, Graf von Buol-Schauenstein, Metternich's successor as minister of foreign affairs, was blamed for allowing Austria to fall between two chairs, the fact is that the Habsburg Empire simply had been caught in an insoluble dilemma. By the nature of its own absolutist regime, it was drawn to tsarist despotism; but in justified fear of encirclement and of the effect of Russian-inspired Pan-Slavism on Austria's Slav population, the empire had to look for the support of the West with its alien political values. This was the basic contradiction in the policy of the Habsburg power.

Napoleon III, largely to distract attention from his own semidictatorial regime and to assuage the Liberals, deemed it advisable to support the cause of Italian independence. In the secret agreement of Plombières of July 1858, he pledged French military support for the liberation of Lombardy and Venetia. This was technically a defensive pact, but in face of a political situation in which three major European powers—France, Great Britain, and Russia—sided with the Italian cause, the Austrians would have been well advised to ignore any Piedmontese provocations. But Buol blundered Austria into an ultimatum by which in April 1859 it demanded Piedmont-Sardinia's demobilization within three days. This was all the justification Cavour and Napoleon III needed to start the Austro-French Piedmontese War. French and Italian military leadership and preparations were far from satisfactory, and conditions on the Austrian side were even worse. Under an inferior commander, Franz, Graf von Gyulai, the Austrians hesitated to take the offensive. In the Battle of Magenta on June 4, they were defeated and Milan had to be evacuated. Gyulai was discharged and the inexperienced emperor Francis Joseph took over the supreme command. On June 24, 1859, the Austrians were decisively beaten in the Battle of Solferino, and two weeks later Napoleon III and Francis Joseph, who was deeply shocked by the carnage of the encounter, concluded the preliminary Peace of Villafranca, followed in November 1859 by the permanent Peace of Zürich. According to its terms Austria had to cede Lombardy—except for Mantua and Peschiera—to Napoleon III, who in turn would give it to Piedmont-Sardinia. The rulers in the Austrian appendages, Modena and Tuscany, who had been driven out by their subjects in the course of the war, were to be restored. These terms represented a kind of compromise. Napoleon III was not anxious to continue a war that might have brought about Prussian intervention on the side of Austria.

The underlying conflict between Austria and Prussia over supremacy in Germany erupted into a crisis as a result of the involvement of the two powers in Danish affairs. When in 1863 the King of Denmark decreed a union of Schleswig and Holstein under one constitution

The Schleswig-Holstein crisis



Austrian Empire, 1815–59.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte;* Georg Westermann Verlag, Braunschweig

and finally incorporated Schleswig into Denmark, he thereby violated international agreements and the charter of the German Confederation, to which Holstein belonged. Protests by Austria and Prussia on behalf of the Confederation were of no avail, nor were they meant to be. Consequently it came in 1864 to war between the two major German powers and Denmark.

The Habsburg Empire's interest in this conflict was not as peripheral as it seemed, for the government was afraid that if action was left to Prussia alone the war would end with unilateral Prussian aggrandizement. Secondly, and perhaps equally important, the issue of Schleswig-Holstein was of major concern to German nationalists. After a brief war, Denmark was forced, by the Peace of Vienna of October 1864, to cede Schleswig, Holstein, and the small duchy of Lauenburg to the Austrian and Prussian sovereigns. But it became increasingly clear that Bismarck did not want to establish the principalities as states within the German Confederation under a sovereign prince. He worked instead in an underhand way for annexation by Prussia. A temporary arrangement was, however, reached at the Convention of Gastein (August 1865), according to which Holstein was put under Austrian military administration, Schleswig under Prussian. Lauenburg was sold outright to Prussia.

The real reason behind this temporary arrangement was the desire of both sides to gain time to make military preparations for the showdown that Bismarck considered inevitable and the Austrian government likely. Further efforts to settle the status of the duchies on a permanent basis by compromise failed, chiefly because this issue served Bismarck with a useful pretext for war. In further preparation Bismarck in April 1866 concluded an alliance treaty with Italy, which promised Italy Prussia's support for the conquest of the Austrian-held Venetian province. The treaty was to lapse if Italy did not open hostilities against Austria within three months. Thus a time limit was set for the beginning of the war.

The pretext for the openings of hostilities was a minor one. When Austria convoked the Estates of Holstein in early June 1866, Prussia declared the Convention of Gastein violated and occupied the duchy by military force. Thereupon Austria asked for mobilization of the confederal armies against the aggressor, Prussia. All member states agreed. Prussia in turn declared the charter of the confederation void and invaded Saxony, Hanover, and Electoral Hesse. The great war for supremacy in Germany, sometimes also referred to as the war between brethren or the Seven Weeks' War, began on June 16, 1866. Four days later Italy joined in the hostilities.

Francis Joseph's initial and grave error was his selection of the commander in the northern theatre of war. Gen. Ludwig August, Ritter von Benedek, a subcommander with many years of experience in the Italian theatre, was against his will put in charge of the main forces against Prussia. Archduke Albrecht, cousin of the Emperor, was given the far less risky command in the south against Italy, so that a member of the dynasty would not be compromised in case of defeat.

The Prussian forces, in accordance with the superior strategy of the chief of staff, Helmuth von Moltke, started a three-pronged attack on Bohemia. Austrian army units were defeated in most of the preliminary engagements in Bohemia. Benedek wanted to withdraw to Moravia, but when the Emperor appealed by implication to his sense of honour, he accepted battle on July 3 at Sadowa (Königgrätz) against three converging Prussian armies. The Austrians, their brave resistance notwithstanding, were defeated and withdrew in disarray. Before the Prussian offensive could be carried to the gates of Vienna, a temporary truce was arranged at Nikolsburg, followed by the preliminary peace concluded there on July 26. Bismarck, concerned with the possibility of immediate French intervention and with long-range plans for future friendly relations with Austria, wanted to avoid unnecessary humiliation of the Habsburg power. The terms confirmed by the permanent Treaty of Prague of August 23 were therefore moderate. Austria would have to recognize the dissolution of the German Confederation and the reorganization of Germany without its participation. Austria's rights in Schleswig-Holstein were transferred to Prussia and it had to pay a relatively minor indemnity. Yet no territorial cessions were demanded, and as a point of honour it was allowed to secure the preservation of the territorial integrity of its most faithful ally, Saxony.

Defeat by Prussia at Sadowa

Before evaluating the peace, the outcome of the war against Italy must be considered. By 1866 the Emperor and most Austrian statesmen had belatedly realized that Austrian rule in Venetia in the face of the opposition of the overwhelming majority of the population there made little sense. Nonetheless, a secret semi-official Italian offer to buy the province was rejected as dishonourable. A conditional offer to cede Venetia outright to Italy came too late in view of the Prusso-Italian alliance. Instead Emperor Francis Joseph concluded on June 12, 1866, a strange agreement with Napoleon III. According to its main provisions in regard to Italy, Austria would cede Venetia, win or lose, to Napoleon, who in turn would give the province to the new Italian kingdom. This meant in effect that the Habsburg power would fight a regular war for the sake of prestige and medieval chivalry, although the outcome was settled before the fighting started. The Italians in turn refused to accept a gift from Napoleon and preferred to fight. The Austrians defeated the Italians at Custoza in Lombardy as well as in the naval Battle of Lissa (Vis) off the Dalmatian coast. But this was of no consequence because of Italy's partnership with Prussia and the latter's victory at Sadowa. By the Peace of Vienna with Italy on October 3, 1866, the cession of Venice, this time to Italy directly, was confirmed.

Significance of the defeat

Of all the wars in the long history of the Habsburg monarchy prior to World War I, the brief Seven Weeks' War of 1866 had probably the most far-reaching effect. Its outcome banished Austria from the rank of the genuine first-rate powers. Beleaguered German nationalism within the empire increasingly assumed a belligerent tone, quite different from the moderate German-directed centralism that had been dominant for many generations. This spirit was challenged not only by a proud Magyarism but also by the combined force of the empire's Slavs —about half the population. This in turn made the Habsburg power increasingly vulnerable to Russian pressure from the east and within 13 years forced it into a German alliance, not as an equal but as a junior partner. Equally important were the internal changes brought by military defeat in two wars within seven years.

THE TRANSITION TO CONSTITUTIONAL GOVERNMENT, 1860–66

In March 1860 Francis Joseph ordered that the Reichsrat, a kind of empire-wide, purely advisory council of

state, should be enlarged by the addition of 38 members proposed by the diets and selected by the crown itself. Only in matters of public finance was this body given a share in legislation, yet it was entrusted with the formidable task of advising the emperor concerning the promulgation of a new imperial constitution. No agreement, however, could be reached. The moderately liberal centralists, represented largely by the Germans, demanded a strong empire-wide legislature and restriction of the agenda of the old Estates diets. They were faced by conservative federalists, largely Magyar, Czech, and Polish nobles, who wanted to strengthen the diets' position. Magyar influence was again on the rise. The Emperor himself sided naturally enough with the conservative forces, which were federalist mainly on the strength of historic and not ethnic claims. The result was the October Diploma of 1860, a constitution proclaimed by decree. It enlarged the diet to 100 representatives and broadened the legislative functions of the Reichsrat in matters of finance, commerce, and industry. Foreign and military agenda remained the exclusive domain of the emperor. Hungary's constitutional status within the empire was restored as it had existed prior to the revolution of 1848, but the concessions agreed to in March 1848 were not recognized. The federalists, particularly the Magyars, objected because their demands had been met only halfway. The centralists rejected the constitution just as strongly, as fully at variance with their claims. In effect, the whole unfortunate legislation had to be abandoned, and on the advice of the new German centralist cabinet, a new one, technically a revision of the October Diploma, was decreed (the so-called February Patent of February 26, 1861).

Constitutional changes

The October Diploma was really no constitution at all in the representative sense; the February Patent was at least an inadequate one. It provided for a bicameral system, an empire-wide house of representatives composed of dietal delegates, and a house of lords, consisting partly of hereditary members and partly of men of special distinction appointed for life. Furthermore, a parliamentary body for the Habsburg lands exclusive of Hungary was established. Opposition of the national groups rendered the constitution unworkable, and in 1865 it had to be suspended. Absolutism was, for all practical purposes, restored under a new prime minister, Richard, Graf Belcredi. The crown could hardly expect that this renewed elimination of constitutional government would be permanent, and the outcome of the war of 1866 made this assumption a certainty. In the meantime, negotiations with the Magyars, who were under the leadership of the highly respected moderate liberal Ferenc Deák and Count Gyula Andrássy, continued. (R.A.K.)

III. Austria-Hungary and the republics of Austria, since 1867

AUSTRIA-HUNGARY, 1867–1918

The liberal ascendancy. *The Ausgleich of 1867.* The economic consequences resulting from the defeat in the war of 1866 (Seven Weeks' War) made it imperative that the constitutional reorganization of the Habsburg monarchy, under discussion since 1859, be brought to an early and successful conclusion. Personnel changes facilitated the solution of the Hungarian crisis. Friedrich Ferdinand, Freiherr (later Graf) von Beust, who had been prime minister of Saxony, took charge of Habsburg affairs, first as foreign minister (from October 1866) and then as chancellor (from February 1867). By abandoning the claim that Hungary be simply an Austrian province, he induced Francis Joseph I to recognize the negotiations with the Hungarian politicians (Déak and Andrássy) as a purely dynastic affair, excluding non-Hungarians from the discussion. On February 17, 1867, Francis Joseph I restored the Hungarian constitution; a ministry responsible to the Hungarian parliament was formed under Count Gyula Andrássy; and in May 1867 Law XII was approved by parliament, legalizing what became known as the *Ausgleich* ("compromise"). This agreement was a compromise between the Hungarian nation and the dynasty, not between Hungary and the rest of the empire; and it is symptomatic of the Hungarian attitude that Hungarians referred to Francis Joseph and his successor as their king and never called him emperor.

Beust becomes chancellor

In addition to regulating the constitutional relations between the king and the nation, Law XII accepted the unity of the Habsburg lands for purposes of conducting certain economic and foreign affairs in common. The compromise was thus the logical result of an attempt to blend traditional constitutional rights with the demands of modern administration. In December 1867, the *engerer* Reichsrat, the section of parliament representing the non-Hungarian lands of the Habsburg monarchy, approved the compromise. Though, after 1867, the Habsburg monarchy was popularly referred to as the Dual Monarchy, the constitutional framework actually was tripartite, comprising the common agencies for economics and foreign affairs, the agencies of the kingdom of Hungary, and the agencies of the rest of the Habsburg lands—commonly but incorrectly called "Austria." (The official title for these provinces remained "the kingdoms and lands represented in the Reichsrat" until 1915, when the term "Austria" was officially adopted for them.)

Under the Compromise of 1867, both parts of the Habsburg monarchy were constitutionally autonomous, each having its government and its parliament composed of an appointed upper and an elected lower house. The "common monarchy" consisted of the emperor and his court, the minister for foreign affairs, and the minister of war. There was no common prime minister and no common cabinet. The common affairs were to be considered at the "delegations," annual meetings composed of representatives from the two parliaments. For economic and financial cooperation, there was to be a customs union and a sharing of accounts, which was to be revised every ten years. (This decennial discussion of financial quotas became one of the the main sources of conflict between the Hungarian and Austrian governments.) There would be no common citizenship, but such matters as weights, measures, coinage, and postal service were to be uniform in both areas.

Although there was no common prime minister or cabinet, there soon developed the so-called *gemeinsamer* Ministerrat, a kind of crown council in which the common ministers of foreign affairs and war and the prime ministers of both governments met under the presidency of the monarch. The common ministers were responsible to the crown only, but they reported annually to the delegations representing both parliaments.

The December constitution

The Compromise of 1867 for all practical purposes set up a personal union between the lands of the Hungarian crown and the western lands of the Habsburgs. The Hungarian success inspired similar movements for the restoration of states' rights in Bohemia and Galicia. But the monarch who only reluctantly had given in to Hungarian demands was unwilling to discontinue the centralist policy in the rest of his empire. Public opinion and parliament in Austria were dominated by German bourgeois liberals who opposed federalization of Austria. As a prize for their cooperation in compromising with the Hungarians, the German liberals were allowed to amend the February constitution of 1861; the Fundamental Laws, which were subsequently adopted in December 1867 and became known as the December consitution, lasted until 1918. They granted equality before the law and freedom of press, speech, and assembly and protected the interests of the various nationalities, stating that

> all nationalities in the state enjoy equal rights, and each one has an Inalienable right to the preservation and cultivation of its nationality and language. The equal rights of all languages in local use are guaranteed by the state in schools, administration, and public life.

The authority of parliament was also recognized. Such provisions, however, were more a promise than a reality. Although parliament, for instance, did theoretically have the power to deal with all varieties of matters, it was, in any case, not a fully representative parliament (suffrage was restricted, and it was tied to property provisions until 1907); and the king was authorized to govern without

parliament in the event that the assembly should prove unable to work. Austrian affairs from 1867 to 1918 were, in fact, determined more by bureaucratic measures than by political initiative; Josephinistic traditions (see above *Foreign Policy, 1763–92*) rather than capitalist interests characterized the Austrian liberals.

Domestic affairs. After the December constitution had been sanctioned, Emperor Francis Joseph appointed a new Cabinet, named the "bourgeois ministry" by the press, because most of its members came from the German middle class (though the prime minister belonged to the Austrian high aristocracy). In 1868 and 1869, this ministry was able to enact several liberal reforms, undoing parts of the Concordat of 1855. Civil marriage was restored; compulsory secular education was established; and interconfessional relations were regulated, in spite of a strong protest from the Catholic Church. In 1870, the Austrian government used the promulgation of the dogma of papal infallibility as pretext for the total abrogation of the concordat.

The progressive legislation of the bourgeois Cabinet stood in sharp contrast to its inability to cope with the demands of the non-German nationalities. In 1868 the Czechs and the Poles issued declarations demanding a constitutional status analogous to that of the Hungarians. The government in Vienna did give the Poles in Galicia a considerable amount of self-government, which was later used to Polonize the Ruthenian minority. In 1871 a ministry for Galician affairs was set up, and the Poles remained the staunchest supporters of the Austrian government well into World War I.

Czech demands for autonomy

The bourgeois ministry was split into a liberal-centralist and a conservative-federalist faction; its members could not reach an agreement among themselves on policies to be adopted. The liberal members of the Cabinet opposed Czech demands; the conservatives showed themselves willing to consider them. Francis Joseph, indignant because of the anti-clerical policy of the liberals, dismissed the prime minister, Fürst Carlos Auersperg, in 1868, replacing him with the conservative Eduard, Graf von Taaffe, his boyhood friend. A period of indecision nevertheless persisted. The Emperor wavered between the liberals, whose anti-clericalism and parliamentarianism he disliked but with whom he sympathized in their centralist, German-oriented policy, and the conservatives, who had his favour in political legislation but aroused his fears by their demands for federalization. Neither Taaffe nor his successors, Leopold Hasner (from December 1868) and Potocki (from April 1870), could solve the Czech problem. The Franco-German War of 1870–71 temporarily diverted public attention from the Czech demands, though public opinion was divided strictly along lines of nationality: Austro-Germans celebrated the victories of the Prussian army, whereas the Slavs were decidedly pro-French. The Austrian government remained neutral because conflicting international interests had blocked the Austro-French negotiations that had culminated in a meeting of Francis Joseph and Napoleon III at Salzburg in 1867. The victory of Chancellor Otto von Bismarck and the establishment of the Second German Empire under the leadership of the Prussian king gave finality to the military decision of 1866. Austria was definitely excluded from the German scene, and a reorientation of dynastic interests seemed a logical consequence. Francis Joseph decided to explore the possibilities of satisfying the Czechs with some measure of federalism. On February 5, 1871, he appointed as prime minister Karl Siegmund, Graf von Hohenwart, a staunch clericalist. The driving mind in Hohenwart's Cabinet, however, was the minister of commerce, Albert Schäffle, an economist whose socialism may not have appealed to the Emperor but whose federalism did.

Failure to satisfy Czech demands

As a first step toward conciliation with the Czechs, the Hohenwart Cabinet dissolved parliament and the provincial diets. When the Bohemian elections improved the federalist position, Hohenwart proceeded to deal directly with the Czechs, copying in certain measure the method used to conclude the compromise with Hungary. Secret talks with Czech leaders František Ladislav Rieger and František Palacký led to the issuance of an imperial rescript by Francis Joseph on September 12, 1871, promising the Czechs recognition of their ancient rights and showing his willingness to take the coronation oath. The Czechs answered this rescript on October 10, 1871, by submitting a constitutional program of 18 articles, called the Fundamental Articles. According to this program, Bohemian affairs should be regulated along the principles of the Hungarian compromise, raising Bohemia to a status equal to Hungary. With this, Hohenwart, who had been up against violent German opposition from the very first day of his appointment, aroused Hungarian resistance, too. Andrássy, fearing that the Czech program could incite minority groups in Hungary, succeeded in convincing Francis Joseph that the stability of the Habsburg monarchy was endangered by the Czech program. On October 27, 1871, Hohenwart was dismissed and Francis Joseph returned the government to the hands of the German liberals.

The new prime minister, Prince Adolf Auersperg, entrusted the key ministries of his Cabinet to university professors and lawyers. The "ministry of doctors," as it was nicknamed by the people, concentrated on legal and administrative reform and endeavoured to strengthen the German control in parliament. After the dismissal of Hohenwart, the Czechs turned to passive resistance, withdrawing from the Bohemian diet and again abstaining from attendance at the parliament in Vienna. This attitude gave the government the chance to weaken the federalist position by introducing a bill for electoral reform. Instead of the existing modus, whereby the diets selected the deputies that were sent to parliament, the new bill set up electoral districts, each of which was to elect one deputy directly to the Reichsrat. The new system, however, preserved the old division of the electorate into *curiae* (socio-economic classes), making parliament in this way a representation of German bourgeois interests.

The economic crisis of 1873

By a strange coincidence, the political victory of German capitalism took place at the very moment of a severe economic crisis. In April 1873, the opening of the Vienna International Exhibition had been thought of as a manifestation of the material progress and economic achievements of the Habsburg monarchy. The so-called *Gründerjahre*, or years of expansive commercial enterprise during the late 1860s and early 1870s, however, were characterized not only by railroad and industrial expansion and the growth of the capital cities of Vienna and Budapest but also by reckless speculation. Warning signs of an imminent crisis were disregarded, and in May, soon after the opening of the exhibition, the stock market collapsed.

The ensuing depression forced the government to abandon liberal bourgeois principles. The state took over the railroads and instituted public-works projects in an attempt to alleviate popular distress. A far-reaching consequence of the stock-market crash of 1873 was the permeation of anti-Semitism into Austrian politics. Jews were accused of being responsible for the speculative stock-market activities, even though official investigations proved that many elements of the population, including some ministers and high aristocracy, had participated in the *Gründungsfieber*, or "speculative fever," and the attendant scandals. The government survived the crisis, however, and German liberal political rule continued for five more years. German liberalism passed into eclipse not because of economic or domestic crisis but as a consequence of its opposition to foreign expansion.

International relations: the Balkan orientation. After his appointment as foreign minister, on November 14, 1871, Count Gyula Andrássy conducted the foreign affairs of Austria-Hungary with the intention of preserving the status quo. Discarding the anti-Bismarck bias of his predecessor, Beust, he sought the friendship of the German *Reich* with the intention of strengthening his position in the unavoidable confrontation with Russia over Balkan problems. The Three Emperors' League (*Dreikaiserbund*) of 1873, by which Francis Joseph and the German and Russian emperors agreed to work together

The Three Emperors' League of 1873

for peace, gave expression to this policy. It also represented Andrássy's intention to strengthen Austria's position in a possible confrontation with Russia over Balkan problems, because the league made a change of the status quo in the Balkans dependent on German consent.

The continuing decline of Ottoman power encouraged the Balkan nations in their opposition to Turkish rule, and in 1875 there were revolts and upheavals. Andrássy failed to induce the Turkish government to adopt a reform program, and by 1876 Russian intervention seemed to be imminent. Russia offered to join with Austria-Hungary in partitioning the Balkans between them, but Andrássy was convinced that Austria-Hungary was a "saturated state" unable to cope with more nationalities and lands, and thus he temporarily resisted the offer. He was aware, however, that Russia could not be restrained altogether; and thus, through Bismarck's mediation, there were concluded two secret agreements, at Reichstadt (Zákupy) in July 1876 and at Budapest in January 1877, whereby Russia gave up its plans for a "great partition" and settled for the territory of Bessarabia and, in return, acquiesced in Austria-Hungary's acquiring Bosnia and Hercegovina. Austria-Hungary and Russia further agreed, however, to refrain from intervention for the time being, and it was only when great-power mediation proved unable to settle the conflict between Serbia and Turkey that Russia declared war on Turkey, in April of 1877, after having once more secured Austro-Hungarian neutrality. In February 1878, with the war won, the Russians did not content themselves with Bessarabia, but, in the Treaty of San Stefano, violated Austria-Hungary's Balkan interests by creating a Great Bulgaria. Having Great Britain as an ally in his opposition to the Russian advance in southeastern Europe and Bismarck as an "honest broker," Andrássy managed at the Congress of Berlin in July 1878 to force Russia into retreating from its excessive demands. Bulgaria was broken up again, Serbian independence was guaranteed, Russia retained Bessarabia, and Austria-Hungary was allowed to occupy Bosnia and Hercegovina. Military occupation of the two provinces turned out to be more than the expected mere formality. It took 150,000 Habsburg troops and several weeks of fighting before the lands were under Habsburg authority. Since no agreement could be reached on whether the newly acquired lands should aggrandize the Hungarian or the Austrian part of the monarchy, an ingenious solution placed them under the jurisdiction of the common Habsburg ministry of finance.

Decisions of the Congress of Berlin, 1878

National conflict and reform. *Taaffe's ministry.* The German liberals had opposed the Balkan policy of Andrássy; and, out of fear that the Slav element in the monarchy would be strengthened by the addition of new Slav population, they voted against the occupation of Bosnia and Hercegovina, in this way withdrawing support from the government. When Prime Minister Auersperg resigned, the era of German liberal predominance came to an end. In 1879, the same year in which the Dual Alliance with the German *Reich* bound the Habsburg monarchy to Germany's foreign policy, the appointment of Taaffe as prime minister signified a reorientation in domestic affairs. From 1879 onward, the German element in the Habsburg monarchy was on the defensive, fighting stubborn and senseless rearguard actions against the Slav drive for political and national equality.

For the elections of 1879, a coalition had formed, consisting of clericals, German aristocrats, and Slavs, which gave itself the name of the Iron Ring. Taaffe had first tried to form a cabinet above parties. It was to include even the liberal Karl, Edler von Stremayr, who had presided over a caretaker government after Auersperg's resignation. The situation in parliament had decisively changed when the Czechs were persuaded by Taaffe, in 1879, to give up their boycott. In April 1880, as a first step, language ordinances were issued that made Czech and German equal languages in the "outer [public] services" in Bohemia and Moravia. In 1882 the University of Prague was divided, giving to the Czechs a national university. In the same year, an electoral reform reduced the tax requirement for the right to vote from ten to five florins, thus enfranchising the more prosperous Czech peasants and weakening the hold of the German middle class. The Taaffe government is also remembered for social-reform legislation; the laws of 1884 fixed the maximum working day (at 11 hours), outlawed the employment of children under 12, required a Sunday rest day for workers, and set up compulsory insurance against accidents and sickness.

Reforms of the 1880s

Despite the conservative character of the government, political life in the Habsburg monarchy underwent a decisive change during the Taaffe period. In the 1880s, the traditional party lineup decomposed, and new alignments and parties formed that were essentially radical and aggressive. The Slav orientation of the Taaffe Cabinet did not satisfy the Czechs but rather encouraged a mood of belligerence; because the moderate Old Czechs failed to live up to radical demands, the nationalistic Young Czechs were able to gain support from the electorate. Similarly, in German Austria and, especially, in Vienna, the moderate liberals were increasingly challenged by extremist groups—notably, German nationalists. In 1882 their "Linz program" proposed the restoration of German dominance in Austrian affairs by detaching Galicia, Bukovina, and Dalmatia from the monarchy, reducing relations with Hungary to a purely personal union under the monarch, and establishing a customs union and other close ties with the German *Reich*. This Pan-Germanic program found its chief protagonist in Georg, Ritter von Schönerer, a deputy to the Reichsrat, who also introduced, for the first time, a note of anti-Semitism into German nationalism. Although his version of extreme chauvinism and racialism never attracted more than a small number of followers, in a modified and moderate way Pan-Germanism and anti-Semitism became the ideological support of the bureaucracy and officer corps; though these elements did not favour union with Germany, they did feel that the Habsburg monarchy had the task of bringing German culture to the "inferior" non-German nationalities. The period also witnessed the founding of parties for the masses. While Schönerer and Pan-Germanism appealed to the educated classes, Karl Lueger transformed the Christian Socialism of Karl, Freiherr von Voegelsang, into a political organization which appealed to small shopkeepers, artisans, tradesmen, and lower bourgeois circles of Vienna and the surrounding countryside. The 1880s finally saw the transition of the workers' movement from the welfare and adult education societies into a political party. Although workers' movements had been weakened in Austria by personal rivalries and government persecution, Victor Adler in 1889, at a conference in Hainfeld, managed to unite the competing Marxist groups into the Social Democratic Party. Political life in Austria from 1890 well into the 1920s was dominated by these three movements originating in the 1880s: Pan-Germanism, Christian Socialism, and Democratic Socialism.

Reaction of the German nationalists

Taaffe continued to probe for compromises between nationalities that were becoming increasingly radical in their demands. In 1890 he tried to negotiate an agreement between the Old Czechs and the German liberals whereby Bohemia would have been divided for administrative and judicial purposes along lines of nationality, but he was balked by the more chauvinistic Young Czechs and German nationalists, and his efforts led to riots in Prague in 1893. When Emil Steinbach joined Taaffe's Cabinet as minister of finance in 1891, he encouraged Taaffe and the Emperor to try electoral reform as an instrument of breaking nationalist opposition. It was hoped that, by extending the franchise, nationalistic antagonism could be allayed and the growing unrest among urban workers could be placated. On October 10, 1893, a suffrage bill was introduced, giving the vote to virtually every literate adult male (though preserving the traditional system of voting in *curiae*). The conservative groups of all nationalities joined forces against this bill, and, under pressure from the Hungarian government, Taaffe had to resign on November 11, 1893. Though failing in political matters, the Cabinet had been successful

Taaffe's later attempts to lessen national tensions

in introducing some economic and social reforms: between 1888 and 1892 a system of cooperative banks for farmers was organized; the taxation system was revised; Austrian currency was stabilized by a return to the gold standard; and the florin was replaced by the crown, which remained the Austrian currency until 1924.

Badeni's ministry. The franchise question continued to dominate Austrian domestic affairs and became closely welded to the nationality conflicts. Alfred, Fürst zu Windischgrätz, sought to win the support of parliament by forming a cabinet in which the clerical conservatives, the Poles, and the German liberals were represented. They were united, however, only in opposition to universal suffrage. Each minister defended his national cause, and the ministry was torn by ceaseless conflict. The end came in June 1895, when the government fulfilled an old promise and introduced Slovene classes into the grammar school at Cilli (Celje), in Styria. Because the school had been exclusively German, this was regarded as a grave blow to the German cause, and the German liberals resigned, forcing Windischgrätz himself to resign.

Deeply embittered by the conduct of the German liberals, Francis Joseph on October 2 entrusted the task of solving the country's problems to a Polish aristocrat, Kazimierz Felix, Graf Badeni, known as a "strong man" for the high-handed way in which he had acted as governor of Galicia. Little noticed at the time, the appointment of Badeni symbolized the breakdown of German control over the Habsburg monarchy. For the first time in Habsburg history, the Germans controlled none of the key positions of government. Not only the Prime Minister (Badeni) but also the Finance Minister (Leo, Ritter von Biliński) and the Foreign Minister (Gołuchowski, who had succeeded Count Gusztáv Kálnoky von Köröspatak the year before) came from the Polish part of the empire. Badeni managed to induce parliament to accept a compromise franchise bill that introduced qualified universal male suffrage but preserved the system of class voting (a fifth *curia* was even added).

Significance of Badeni's appointment

The shortcomings of the new system enraged the parties representing the masses of the population. In the 1870s and 1880s, decisive economic changes with far-reaching social consequences had occurred in the Habsburg lands. Though remaining primarily agrarian, the Habsburg lands had undergone an industrialization that had resulted in an unprecedented growth of urban centres. Vienna, which had about 430,000 inhabitants in 1851, found itself a metropolis of 1,800,000 at the turn of the century; and this phenomenon was paralleled in other areas, especially in Bohemia, which had become the industrial centre of the western part of the Habsburg lands. The socio-economic development naturally began to affect politics. From 1890 on, the advance of the Social Democrats and Christian Socialists caused considerable tension in Vienna. In October 1894, the Social Democrats were able to organize their first impressive, orderly mass demonstration in the capital, and the communal elections of 1895 had made the Christian Socialists the strongest party in Vienna, ending the long liberal rule. When the Emperor refused to confirm Karl Lueger, the popular leader of the Christian Socialists, as mayor of Vienna, there were demonstrations and protests. Not until Lueger was elected mayor the fifth time did Francis Joseph agree to confirm him, in April 1897. Furthermore, a few weeks earlier, the elections held on the basis of the new suffrage had strengthened the radical elements in the Reichsrat; the Young Czechs, for instance, had completely overwhelmed the conservative Old Czechs.

Growing strength of socialist parties in the 1890s

Counting on support from the Slav and conservative parties in parliament, Badeni dared to take up the Bohemian language question again. In April 1897, he issued a famous language ordinance that introduced Czech as a language equal to German even in the "inner service"—that is, for communications within government departments. This decision would have meant that civil servants in Bohemia and Moravia would have to be able to speak and write Czech as well as German. Since the Germans refused to learn Czech, this would have put them in a definite disadvantage in Bohemia's administration. The publication of the ordinance thus provoked violent German reactions: university professors signed resolutions of protest; mass meetings incited the public; and German deputies in the Reichsrat began to obstruct all legislative activities. The protest reached its climax in November 1897, when parliamentary sessions turned into bedlam, and popular protests against Badeni led to street demonstrations. The mass protest was not restricted to Vienna; it was even worse in some German towns in Bohemia; in Graz, clashes between soldiers and the masses ended in the death of one demonstrator. For a moment it seemed as if 1848 was about to return. To pacify the public, Francis Joseph gave in; on November 28, 1897, he dismissed Badeni, and asked Paul, Freiherr Gautsch von Frankenthurn, a former minister of education, to form a government out of the German parties of parliament. Gautsch's attempts to appease the Germans ran into obstruction from the Czechs. The stage of violence was shifted from Vienna to Prague and from the Reichsrat to the Bohemian diet. In March 1898 Gautsch was replaced by the former governor of Bohemia, Franz Anton, Fürst von Thun und Hohenstein, who in turn failed within a year. Neither Manfred, Graf Clary und Aldringen, who formally revoked the Badeni language ordinance, nor his successor, Wittek, who headed a short-lived cabinet of a few weeks, managed to solve the nationality problem.

Koerber's ministry. Finally, on January 18, 1900, Francis Joseph asked Ernst von Koerber, a former minister of the interior, to form a new Cabinet. Koerber was the first and only commoner to be appointed prime minister by Francis Joseph; as a leading bureaucrat, he formed his ministry from the ranks of other bureaucrats, concentrating, in the subsequent years, on the administration of public affairs and economic programs rather than trying to deal with political problems. First by imperial decree and then, after some political bargaining, by consent of parliament, Koerber carried through a program of economic expansion, social legislation, and administrative reform, liberating the press from government and police control. By devious politicking, Koerber managed to keep government activities free from national strife, but he could not prevent national emotions from becoming more and more extremist. The national conflict came to be fought over educational matters, and in the final years of Koerber's government the desire for national universities aroused the sentiment of the Italians, Slovenes, and Ruthenians, turning the traditional Czech–German conflict into a multi-national one. In December 1904, Koerber's various manoeuvrings faltered, and he was driven from office by a combination of parties.

Nationalism and electoral reform. The political climate in Austria was further complicated by the worsening of relations between the Emperor and the Hungarian government. Magyar separatists had agitated for the separation of the Habsburg army, and when Francis Joseph used an address to the troops at Chlopy in 1903 for an unequivocal reaffirmation of the common and unified character of his army, a controversy developed that had repercussions in the Austrian half of the Dual Monarchy. The plan to use universal suffrage to break the opposition in Hungary furthered the cause of political democracy in Austria. The demand for universal and equal suffrage had increased since the Russian revolution in the winter of 1905; after Koerber's Cabinet had run aground over a minor financial matter toward the end of 1904, Gautsch was chosen by the Emperor to introduce universal franchise in Austria. Though the first bill, introduced to parliament by Gautsch in February 1906, ran into the opposition of the middle-class and conservative parties that still controlled parliament, the realization of this program could no longer be blocked. Imperial interest and popular pressure—the Social Democrats had organized mass rallies to support the bill—combined to overcome parliamentarian opposition. After Gautsch had resigned in March 1906 and Prince Conrad von Hohenlohe-Schillingsfürst had failed to master the situation, Max Wladimir, Freiherr von Beck (prime minister from June 1906), managed to carry the bill through

Universal suffrage

parliament. In January 1907 Francis Joseph sanctioned the law giving the vote to every male of 24 or over, and abolished the *curiae*. Membership of the Reichsrat was increased from 425 to 516; the returns of the election of 1907 made the Germans now inescapably a minority, with 233 members, though certainly the strongest national group. The Czechs could count on 107 seats, the Poles 82, the Ruthenians 33, the Slovenes 24, the Italians 19, the Serbo-Croats 13, and the Romanians 5.

Universal suffrage brought the expected decline of the chauvinistic parties. The Young Czechs as well as the Pan-Germans were reduced to small factions without parliamentarian influence, whereas the Christian Socialists and the Social Democrats returned as the two strongest parties out of more than 30 represented in parliament; the Socialist delegation in the Austrian parliament was, in fact, larger than in any other country. The Austrian constitution, however, did not force the emperor to form his government according to the composition of the parliament. Neither the Social Democrats nor the Christian Socialists were able to acquire any significant influence on the shaping of Austrian government affairs.

Beck remained in office and satisfied the Christian Socialists with some concessions but for the most part based his policy on the support of the conservative parties. In 1905 the diet of Moravia had succeeded in finding a compromise between German and Czech national demands, and it was hoped a similar compromise could be achieved for Bohemia. But, within a short time, national conflicts got the upper hand again, and parliamentary debate and public opinion were once more excited by national strife. In 1908, however, international complications diverted attention from domestic affairs.

Foreign policy, 1878–1908. *The alliances.* The occupation of Bosnia and Hercegovina in 1878 had reasserted Habsburg interests in Balkan affairs. Facing the possibility of conflict with Russia in this area, Austria-Hungary looked for an ally, with the result that in 1879 Austria-Hungary and the German Reich had joined in the Dual Alliance, by which the two sovereigns promised each other mutual support in the case of Russian aggression. The signing of the Dual Alliance was Andrássy's last act as foreign minister, for he resigned shortly afterward, but the alliance survived as the main element in the international position of the Habsburg monarchy until the very last day of the empire. Under Andrássy's successors—Heinrich, Freiherr von Haymerle, and Kálnoky—Habsburg foreign policy continued its conservative course.

The Three Emperors' Alliance of 1881

In 1881 an alliance with Serbia, which after the Congress of Berlin turned to Austria-Hungary for protection, made this Balkan state a satellite of the Habsburg monarchy. The Three Emperors' Alliance (Russia, Germany, Austria-Hungary) of the same year brought Russian recognition of Habsburg predominance in the western part of the Balkan peninsula. The signatories of this alliance promised to consult one another on any changes in the status quo in the Ottoman Empire, and, while Russia was given assurances that its position regarding the Straits and Bulgaria would be recognized, Austria-Hungary received from Russia the promise that there would be no objection to a possible annexation of Bosnia-Hercegovina in the future.

The Three Emperors' Alliance was an important element in the structure of alliances that Bismarck set up to stabilize the European continent. Having decided to rely on Austria-Hungary as the fundamental partner in international affairs, Bismarck had to endeavour to neutralize all the areas in which the Habsburg monarchy might possibly be drawn into a conflict. It was essential to avoid being involved in a controversy at an inopportune moment and in a region of little interest to Germany. Bismarck therefore attempted to lessen the possibility of a conflict between Austria-Hungary and Russia by making them partners in the Three Emperors' Alliance. And when, in 1882, Italy approached Germany to find a partner in its anti-French policy, Bismarck used the opportunity to neutralize another European trouble spot. He informed the Italian foreign minister that the road to Berlin led through Vienna, with the result that the Triple Alliance (Italy, Germany, Austria-Hungary) was signed in May 1882. It was primarily a defensive treaty against a French attack on Italy or Germany. It further stated that, in the event of any signatory coming to war with another power, the partners of the alliance would remain neutral. The treaty did not settle the problems still existing between the Habsburg monarchy and the Italian kingdom, but for Bismarck it sufficed that they were neutralized.

The Triple Alliance

In 1883 Bismarck acted once more to reduce the danger of war in "Europe's backyard" by arranging a defensive agreement between Austria-Hungary and Romania. The Triple Alliance and the Romanian Alliance not only strengthened the international status quo but also gave security to the internal order of the Habsburg monarchy by weakening the irredentist movements in Transylvania and the Italian parts of Austria-Hungary.

The deterioration of German–French relations in the following years convinced Bismarck of the indispensability of the Triple Alliance, and he made every effort to force Vienna to renew the alliance in 1887. By threatening to withdraw protection against Russian aggression, Bismarck forced Kálnoky to consent to his demands, but there can be no doubt that Austria-Hungary was clearly impeded in its national interests by having to adapt its foreign policy to the German and Italian demand for the isolation of France. Although Kálnoky succeeded during the negotiations in avoiding any new obligation in western Europe, he was less successful in defending more immediate Austrian interests. He managed to evade the Italian request for the support of an active Italian colonial policy, but he was unable to keep Italy out of involvement in Balkan affairs. It might be that in view of his own conservative and defensive policy he saw an advantage in having Italy as a third partner in the maintenance of the status quo against possible Russian expansion. At any rate, it was on Kálnoky's initiative that the original Italian demand for a declaration in favour of the status quo along the Ottoman coasts and the Adriatic and Aegean seas was extended to the interior of the Balkan peninsula.

Kálnoky's Balkan policy

On top of this Kálnoky granted the Italians the right to ask for compensation in case of any change in the territorial status quo without defining this term. In a certain way, all the differences and clashes between Austrian and Italian Balkan policy in the first decade of the 20th century can be traced to the introduction of this clause (later formulated in article VII of the treaty) at the renewal of 1887. In the same year, Bismarck built around the Triple Alliance a system of alliances and agreements which amounted to complete isolation of France and obliged the major European powers to guarantee the status quo along the borders of the Ottoman Empire. The First and Second Mediterranean Agreements of 1887 joined Great Britain to the powers (Austria and Italy) interested in blocking Russia from the Straits, and enabled Kálnoky to abandon direct agreements with Russia. The Three Emperors' Alliance of 1881 was allowed to expire, and Austria-Hungary was thus left without any formal understanding with Russia. Count Agenor Gołuchowski, who followed Kálnoky as foreign minister in 1895, decided, however, that direct relations with St. Petersburg should be renewed. In April 1897, Francis Joseph and Gołuchowski visited St. Petersburg. The agreements signed as a result of this initiative aimed at excluding Italy from Balkan affairs and sought to entrust preservation of the Balkan order to the bilateral cooperation of the two eastern monarchies rather than to a multilateral alliance system. The final years of the 19th century were marked by a change from static continental policy to a more dynamic world policy, and the ensuing mobility in international relations reduced the value of the Triple Alliance.

The Bosnian crisis. The Austro-Russian agreements of 1897 came to bear when, in 1903, a major revolt occurred in Macedonia. Following a meeting of Tsar Nicholas II and Francis Joseph in October 1903, their foreign ministers drafted a reform program for the Otto-

man Empire. A mutual neutrality agreement was added the following year, leaving Austria-Hungary a free hand in the event of a conflict with Italy and enabling Russia to turn and face Japan. Explicitly excluded from the agreement with Russia were Balkan conflicts. When King Alexander of Serbia was assassinated in a military revolt in 1903 and the Obrenović dynasty was replaced by the Karageorgević, Serbian relations with the Habsburg monarchy deteriorated. The Serbs adopted an expansionist policy of unifying all southern Slavs in the Serbian kingdom, and, in order to block a Serbian advance, the Habsburg monarchy applied economic pressure. In 1906 all livestock imports from Serbia into the Habsburg monarchy were prohibited. This conflict, the so-called Pig War, did not crush Serbia but rather pushed it into the Russian camp.

When, in 1906, Count Gołuchowski was replaced as foreign minister by the former ambassador to St. Petersburg, Alois, Freiherr Lexa von Aehrenthal, a turning point in Austrian foreign policy was signalled. Aehrenthal made a belated effort to free Austria-Hungary from its submission to German interests and to engage in a dynamic Balkan policy. A first step in this direction was his proposal for the construction of a railroad through the Sandjak of Novi Pazar. The combined Russian and Serbian opposition forced Aehrenthal to abandon the project temporarily and made it clear that any advance in the Balkans would probably result in war with Serbia and perhaps with Russia as well. The danger of such a conflict arose within a short time. In July 1908, following a revolution in Turkey, the Young Turk movement announced the reform of the Turkish constitution. Afraid that this constitutional change could undermine the Habsburg position in Bosnia and Hercegovina, two provinces that nominally were still under Ottoman suzerainty, Aehrenthal decided to use the opportunity to fortify the Austro-Hungarian position in the Balkan peninsula. In September 1908, he met with the Russian foreign minister, Count A.P. Izvolsky, and secured, so he thought, Russian approval of the proposed annexation in return for Austria's support in having the Straits opened to Russian warships. On October 6, 1908, the annexation was announced, immediately bringing a violent reaction from Serbia. When Izvolsky found that his plans for the Straits were opposed by Great Britain and France, he retracted his tentative support of Austria and supported the Serbian position. The situation became serious, and for a while war seemed imminent. Franz, Freiherr Conrad von Hötzendorf, the chief of the general staff of the Habsburg monarchy, who had long advocated preventive war, pushed for an aggressive move, but Aehrenthal had apparently never planned more than going to the brink of war. In March 1909, a German ultimatum forced the Russians to withdraw their support from Serbia, and, since the Turkish government had agreed to the annexation of the two provinces in return for a monetary compensation, Serbia also had to come to terms with the Habsburg monarchy. The Bosnian crisis was settled, but the Serbians felt their national pride deeply wounded and continued to stir unrest in the southern Slav provinces of the Habsburg monarchy.

Annexation of Bosnia and Hercegovina

The last years of peace. *Conflicts of nationality.* The annexation crisis had repercussions among the other Slav nationalities in the monarchy. For several years, Czechs had been attracted by the Pan-Slav movement, and in July 1908 a Pan-Slav congress was held in Prague. During the diplomatic crisis of the following winter, the Czechs unabashedly took the side of the Serbs, and, on the day of the 60th anniversary of Francis Joseph's accession to the throne, martial law had to be declared in Prague. National strife broke out once more all over the monarchy, and parliamentary activities were all but blocked by filibustering and the riotous activities of the deputies. Prime Minister Beck had resigned in November 1908; and his successor, Richard, Freiherr von Bienerth, after having accomplished little with a cabinet of civil servants, tried to appease the nationalities by including *Landsmannminister* (national representatives) into his Cabinet (February 1909).

Obstruction in parliament continued. The Germans, in control of the government and the central administration, continued to assign to the monarchy the role of an outpost of German culture; the Slavs, however, increasingly wanted to make Austria the home of Slav national aspirations. The Czech agrarian leader František Udržal stated in parliament: "We wish to save the Austrian parliament but we wish to save it for the Slavs of Austria who form two-thirds of the population." A population census taken in 1910 confirmed the Slav claim: out of the 28,324,940 inhabitants of the western half of the Austro-Hungarian Empire, only 35.58 percent regarded themselves as German; 17.77 percent Poles; 12.58 percent Ruthenians; 23.02 percent Czechs and Slovaks; 4.48 percent Slovenes; 2.8 percent Serbs and Croats; and 2.75 percent Italians. The Slav predominance was weakened by the attitude of the Poles, who remained loyal to the central government, thus allowing the national conflict to assume the character of a primarily Czech–German quarrel. Even the Social Democratic Party could not overcome nationalist antagonism. In 1899, at the party congress at Brünn, the Social Democrats had presented a national reform program based on democratic federalism, granting the right of national decisions to territorial units formed on a basis of nationality. Karl Renner and Otto Bauer, who later became leaders of German–Austrian Socialism, drafted various programs for the solution of the nationality problem in books published between 1900 and 1910. But these efforts could not prevent the Socialists from splitting along national lines, too, and in 1910 the Czech Socialists declared themselves independent of the Austro-German Socialist Party.

National composition of the empire in 1910

Party rivalries. Such national differences weakened the Socialist position in the elections of 1911. Over 50 parties had competed in the campaign, and, since the German nationalist parties had allied in the Deutscher Nationalverband, they managed to return to parliament as the strongest single party, gaining 104 seats out of 516. The Christian Socialists, weakened by personal rivalry, suffered heavy losses, winning only 76 seats. The German Social Democrats received 44 seats and the Czech Socialists 24. The Czech parties were badly divided, those representing the Czech middle class gaining 64 seats. Bienerth found himself unable to form a workable ministry, and he was replaced by Gautsch, who tried to reconcile the Germans and the Czechs. For a while negotiations seemed quite successful, but extremist incidents deadlocked the talks, and the Gautsch Cabinet was replaced by a new ministry headed by Karl, Graf von Stürgkh (November 1911). Unable to deal with the nationality problem in a parliamentarian fashion, Stürgkh repeatedly suspended the Reichsrat. It was characteristic of the general political climate in Europe that Stürgkh had to concentrate his legislative program on the improvement of Austrian armament, for international crises overshadowed the nationality conflict.

Stürgkh's suspensions of the Reichsrat

Conflict with Serbia. Ever since the Bosnian crisis of 1909, Austrian diplomats had been convinced that war with Serbia was bound to come. Aehrenthal had fallen sick soon after the annexation of Bosnia and Hercegovina, and after a long illness he died (February 1912), at a moment when an Italian–Turkish conflict over Tripoli had provoked anti-Turkish sentiment in the Balkan states. Leopold, Graf Berchtold, who directed Austro-Hungarian foreign policy from 1912 on, did not possess the qualities required in such a critical period. Aehrenthal had been able to silence the warmongering activities of Conrad, who continued to advocate preventive war against Italy and Serbia, but Berchtold yielded to the aggressive policies of the military and the younger members of his ministry. During the Balkan Wars (1912–13), fought by the Balkan states over the remnants of the Ottoman Empire, Austria-Hungary twice tried to force Serbia to retract from positions gained by threatening it with an ultimatum. In February and October 1913, military action against Serbia was contemplated, but in both instances neither Italy nor Germany was willing to guarantee support. Austria-Hungary had to acquiesce in the territorial changes in the Balkan peninsula, changes that

eliminated the Turks from Europe. By supporting Bulgaria against Serbia, Austria-Hungary alienated Romania, which country had shown resentment against the Habsburg monarchy because of the treatment of non-Magyar nationalities in Hungary. Romania thus joined Italy and Serbia in support of irredentist movements inside the Habsburg monarchy. By 1914, leading government circles in Vienna were convinced that offensive action against the foreign protagonists of irredentist claims was essential to the integrity of the empire.

The assassination of Francis Ferdinand

In June 1914, Archduke Francis Ferdinand, the heir of Francis Joseph, participated in army manoeuvres in the provinces of Bosnia-Hercegovina, disregarding warnings that his visit would arouse considerable hostility. When Francis Ferdinand and his wife were assassinated by a Bosnian nationalist at Sarajevo on June 28, 1914, the Austro-Hungarian foreign office decided to use the opportunity for a final reckoning with the Serbian danger. The support of Germany was sought and received, and the Austro-Hungarian foreign office proceeded to draft an ultimatum putting the responsibility for the assassination on the Serbian government and demanding full satisfaction. The attitude of the foreign office was shared by Conrad and by Stürgkh but was opposed by the Hungarian prime minister, Count István Tisza, who wanted an assurance that a military move against Serbia would not result in territorial acquisitions and thus increase the Serb element in the monarchy. His demand satisfied, Tisza joined the advocates of war. In ministerial meetings on July 15 and 19, a deliberately provocative ultimatum was drafted in words that supposedly excluded the possibility of acceptance by Serbia. The ultimatum was handed to the Serbian government on July 23. The Serbian answer, handed in on time on July 25, was declared as being insufficient, though Serbia had agreed to practically all Austro-Hungarian demands with the exception of two that, in effect, entailed constitutional changes in the Serbian government. These were that certain unnamed Serbian officials be dismissed at the whim of Austria-Hungary and that Austro-Hungarian officials participate, on Serbian soil, in the suppression of organizations hostile to Austria-Hungary and in the judicial proceedings against their members. In its reply, the Serbian government pointed out that such demands were unprecedented in relations between sovereign states but nevertheless agreed to submit the matter to the Permanent Court of Arbitration or to the arbitration of the great powers. On receiving this reply, the Austro-Hungarian ambassador immediately left Belgrade, severing diplomatic relations between the two countries. Berchtold and his government were clearly determined to make war on Serbia, regardless of the fact that such action might result in war between the great powers. While the European governments frantically tried to offer compromise solutions, Austria decided on a *fait accompli*. On July 28, 1914, Berchtold asked Emperor Francis Joseph to sign the declaration of war, informing him that

> it cannot be excluded that the [Triple] Entente powers [Russia, France, Great Britain] might make another move to bring about a peaceful settlement of the conflict unless a declaration of war establishes a fait accompli [*eine klare Situation geschaffen*].

In the meantime the German government had taken control of the situation and, placing German strategic and national plans over Austro-Hungarian interests, had changed the Balkan conflict into a Continental war.

World War I. *Subordination to Germany.* The German declaration of war against France and Russia subordinated the Austro-Serbian conflict to the German aim of settling its own rivalries with France and Russia. According to the terms of the military agreement between Germany and Austria-Hungary, the Austro-Hungarian army had to abandon plans to conquer Serbia and instead protect the German invasion of France against Russian intervention. The setbacks that the Austrian army suffered in 1914 and 1915 can be attributed, to a large extent, to the fact that Austria-Hungary became a military satellite of Germany from the very first day of the war, though it cannot be denied that the Austrian high command proved to be quite incompetent. Conrad had clamoured for preventive war since 1906, but, when he received his chance, in July 1914, it turned out that the

Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York



Peoples of Austria–Hungary in 1914.

Austrian army had no plans for an expeditious offensive. Similarly, after Italy entered the war on the side of the Allies in May 1915, Conrad was unprepared. The fact that only after the Germans had taken command could the Russian front be stabilized and that Serbia and Romania were not defeated until 1915 and 1916, respectively, did little to enhance the prestige of the Austrian government.

Internal disintegration. In July 1914, parliament had been out of session, and Stürgkh had refused to convene it. This and the military censorship that was established immediately after the outbreak of the war concealed the discontent of the non-German population. While German public opinion in Austria had welcomed the war enthusiastically, and while some Polish leaders supported the war out of their anti-Russian resentment, the Czech population openly showed its animosity. The Czech leader Tomáš Garrigue Masaryk, who had acted as one of the most prominent spokesmen for the Czech cause, emigrated to western Europe. Karel Kramář, who had supported the Pan-Slav idea, was put on trial for high treason and declared guilty on the basis of shaky evidence. German nationalism was riding high, but in reality the German Austrians had little influence left. In military matters, they were practically reduced to executing German orders; in economic affairs, the Hungarians, who controlled the food supply, had the decisive influence. Count Tisza, who had opposed the war in July 1914, became the strong man of the empire. On his advice, Berchtold was dismissed in January 1915, and the foreign office was once again entrusted to a Magyar. But Count István Burián, who was in charge of foreign affairs until December 1916, failed to keep Italy and Romania out of the war. German attempts to pacify the two states by concessions were unsuccessful, because Francis Joseph was unwilling to cede any territory in response to the irredentist demands of the two nations. How little the outward calm in the Habsburg lands corresponded to the sentiment of the population became apparent when Stürgkh was assassinated by Friedrich Adler, the pacifist son of the leader of Austrian Socialism, in October 1916. Francis Joseph made Ernst Koerber prime minister, but Koerber had no chance to develop a program of his own. On November 21, 1916, Francis Joseph died, at the age of 86, leaving the throne and the shaky empire to his 29-year-old grandnephew, Charles, who had had little preparation for his task until the death of Francis Ferdinand had made him the heir apparent. Full of the best intentions, Charles set out to save the monarchy by searching for peace in foreign affairs and by recognizing the rights of the non-German and non-Magyar nationalities of his empire. Charles relied heavily on the advice of politicians who had had the confidence of Francis Ferdinand. He dismissed Koerber in December 1916 and made Count Heinrich Clam-Martinic, a Czech aristocrat, prime minister. At the foreign office, he replaced Burián with Ottokar, Graf Czernin.

Assassination of Stürgkh

When parliament was reconvened in May 1917, it became manifest how far internal disintegration of the Habsburg monarchy had progressed. Parliament once again became the stage of unrelenting national conflicts. Finding so little support from the Czech side, Charles turned back to the German element, and in June 1917 made Ernst von Seidler, once his tutor in administrative and international law, prime minister. But, even though he tried to appease the Czechs, the stubborn insistence of the Germans not to yield any of their prerogatives made reform of the empire impossible.

Early peace feelers

At the same time, various moves to get Austria-Hungary out of the war ended in failure. After a U.S. offer of general mediation had miscarried in December 1916, Charles tried through secret channels to deal directly with the Entente powers. In the spring of 1917, an exchange of peace feelers took place through the mediation of his brother-in-law, Sixtus of Bourbon-Parma, but Italy's unwillingness to abandon some of the concessions granted to it in the Treaty of London (1915) made these talks abortive. Similarly, negotiations with Allied representatives carried on in Switzerland brought no results.

Since the Austro-Hungarian government was unable to extricate itself from the ties of the Dual Alliance, France and England ceased to have regard for the integrity of the Habsburg monarchy. Furthermore, the revolutionary events in Russia in 1917 and the entry of the United States into the war introduced a new, ideological element into Allied policy toward the Central Powers. The German-directed governments represented an authoritarian system of government, and national agitation in the Habsburg monarchy henceforth assumed the character of a democratic liberation movement, winning the sympathies of west European and American public opinion. From early 1918 on, the Allied governments began officially to promote the activities of the emigrés from Austria, foremost among them Tomáš Masaryk, the Czech leader, and in April 1918 a Congress of Oppressed Nationalities was organized in Rome. But the collapse of the Habsburg monarchy cannot be ascribed to the new Allied policy of supporting the independence claims of the Habsburg nationalities, which was only a belated adjustment to the changed conditions within Austria-Hungary. From the summer of 1917, the activities of the nationalist movements within the empire made the situation increasingly untenable, and, two days before U.S. Pres. Woodrow Wilson proclaimed his Fourteen Points, one of which demanded the reorganization of the Habsburg monarchy in accordance with the principles of national autonomy, the Czechs demanded outright independence (January 6, 1918). Within a month, Polish and South Slav deputies, together with the Czechs, presented to the Reichsrat a program demanding the establishment of independent constituent assemblies for nationally homogeneous areas.

The end of the Habsburg Empire. During the same period in which the national-independence movement reached its final stage, another dangerous development manifested itself. From 1915 on, the supply situation had worsened increasingly, and by January 1918 there were dangerous shortages, especially of food. Prompted by the difficult food situation and inspired by the Bolshevik victory in Russia, a strike movement developed. Demands for more bread and a demand for peace were combined with nationalist claims into open opposition to the government. The strikes among the civilian population were followed by mutinies in the army and navy. In January and February 1918, however, the army and the government succeeded in suppressing social unrest and antiwar demonstrations. But, from the same date, the national opposition movement gathered momentum.

Strikes, mutinies, and demonstrations

The hopes that the government had placed in peace settlements with the eastern states were not fulfilled, either. The peace treaty with the Ukraine (signed in February 1918), the Treaty of Brest-Litovsk with Soviet Russia (March 3, 1918), and the Treaty of Bucharest, which settled the peace with Romania (May 7, 1918), did not alleviate the supply situation and irritated the Poles because of certain provisions of the Ukrainian settlement.

In April 1918, Czernin was replaced as foreign minister by Burián. This change was the result of a conflict between Czernin and Charles over the desirability and possibility of Austria concluding a separate peace with the Allies. Unknown to Czernin, Charles had in 1917 made certain secret overtures to the Allies, which were revealed by French premier Georges Clemenceau. The Germans were outraged, and Czernin was dismissed on their orders. Burián returned to the foreign office on April 16 and immediately reported to the German high command at Spa, where Emperor Charles and Burián had to assure the German Emperor of their unchanging loyalty. While this act of submission satisfied the German Austrians, it further incensed the Slav opposition. In May 1918 a Slav national celebration in Prague demonstrated the strength of the independence movements. But the Emperor and the German elements in the central government were still not aware of the extent of the disintegration. In July 1918, Prime Minister Seidler resigned, and his successor, Max Hussarek von Heinlein, began a belated effort to reorganize the Habsburg monarchy. Hussarek's efforts to federalize the empire in the moment of imminent military defeat unintentionally turned out to provide

Belated attempts at reform

the basis for the formal liquidation of the Habsburg monarchy. On October 16, 1918, Emperor Charles issued a manifesto announcing the transformation of Austria into a federal union of four components (German, Czech, South Slav, and Ukrainian). The Poles were to be free to join a Polish state, and Trieste was to be given a special status. The lands of the Hungarian crown were to be excepted from this program. Within a few days, national councils were established in all the provinces of the empire and for all practical purposes acted as national governments. The Poles proclaimed the union of all Poles in a unified state and declared their independence at Warsaw on October 7, 1918. The South Slavs advocated union with Serbia, and on October 28, 1918, the Czechs proclaimed the establishment of an independent republic. The dissolution of the Habsburg monarchy was thus consummated by the end of October 1918—that is, before the war actually ended.

It was impossible for the country to survive another winter of hostilities, and on September 14, 1918, Burián published an appeal to all belligerents to discuss the possibilities of ending the war. When this move was opposed by the Germans as well as by the Allied powers, Burián tried for a separate peace settlement for Austria-Hungary. On October 14, 1918, he sent a note to President Wilson asking for an armistice on the basis of the Fourteen Points. On October 18, 1918, U.S. Secretary of State Robert Lansing replied that, in view of the political development of the preceding months and, especially, in view of the fact that Czechoslovakia had been recognized as being at war with the Central Powers, the U.S. government was unable to deal on the basis of the Fourteen Points anymore. On October 27, 1918, Count Gyula Andrássy, who had replaced Burián three days before as foreign minister, sent a new note to President Wilson; in asking for an armistice, he declared full adherence to the statements set forth in the U.S. note of October 18, thus explicitly recognizing the existence of an independent Czechoslovak state. From this moment on, it remained only to liquidate the war. On October 22, 1918, Heinrich Lammasch, a renowned authority in the field of international law and a respected pacifist, formed a new Cabinet. He hoped to save the Habsburg monarchy by drawing up a federative structure. But instead of saving the state, he found himself charged with the task of supervising the dissolution of the empire and bringing about an orderly transfer of power. The government could not influence events outside of Vienna anymore and from October 30, 1918, was even challenged in the central agencies by the German–Austrian state council. The hostilities were ended by an armistice signed on November 3, 1918. The Austro-Hungarian high command, which had blundered into the war unprepared in 1914, did little better at its conclusion. Due to inaccuracies in the wording of the documents, more than 300,000 soldiers were taken prisoner by the Italian army. For some days the government hoped that, in spite of the secession of the Slav areas, the Habsburg dynasty could survive in the remaining lands. But even the German Austrians had lost faith in the Habsburgs, and, with revolutionary agitation on the rise and republican passion widespread, Charles adhered to the advice of Lammasch and decided to waive his rights to exercise political authority. On November 11, 1918, he issued a proclamation acknowledging "in advance the decision to be taken by German Austria" and stating that he relinquished all part in the administration of the state. The declaration of November 11, 1918, marks the formal act of dissolution of the Habsburg monarchy.

THE FIRST REPUBLIC AND THE ANSCHLUSS

The war's aftermath. On October 21, 1918, the 210 German members of the imperial parliament (Reichsrat) of Austria formed themselves into a national assembly for Deutschösterreich, or German-Austria; and on October 30 they proclaimed this an independent state under the direction of a State Council (Staatsrat) composed of the leaders of the three main parties and other elected members. Revolutionary disturbances in Vienna and, more important, the news of the German revolution forced the State Council on the republican path. On November 12, the day after the emperor Charles's abdication, the National Assembly resolved unanimously that "German-Austria is a democratic republic" and "German-Austria is a component part of the German republic." Karl Renner, a leading Socialist, became head of a coalition government, with Otto Bauer, the acknowledged spokesman of the left wing of the Social Democrats, as foreign secretary. On November 22, the territory of the republic was further defined: the National Assembly claimed for the new state all the Habsburg lands in which a majority of the population was German. It also claimed the German areas of Bohemia and Moravia.

Dangers of a Communist revolution

From the first day, the republic was faced with the disastrous heritage of the war. Four years of war effort and the breakup of the Habsburg Empire had brought economic exhaustion and chaos. The resulting social distress and poverty inspired revolutionary activities, thus making Bolshevism appear the greatest danger to the new republic, especially after a Soviet republic was established in Hungary at the end of March 1919. The Austrian Social Democrats were determined to resist Bolshevism with their own forces without making an alliance (as the German Social Democrats did) with the old order. A Volkswehr (People's Guard) was organized and was twice effective (April 17 and June 15) against Communist attempts at a *Putsch*. Otto Bauer and Friedrich Adler staked their popularity on defeating the Communist agitation in the workers' and soldiers' councils, which had been set up on the Soviet model. By mid-1919, political and social order was restored on parliamentary lines, the Communist Party relapsing into insignificance. More dangerous was the tendency of the *Länder* (provinces, or states) to break away from Vienna or to claim almost complete independence. Though the principal motive of this was reluctance to send food supplies to Vienna, it also represented a genuine social, political, and ideological conflict: the administration of the industrialized capital was Socialist controlled, while the *Länder*, being predominantly agrarian, remained conservative and faithful to the Catholic tradition. This difference was aggravated by the fact that the monarchy had been the only bond between the German Austrian lands; with the abdication of the Emperor, no symbol of loyalty common to all *Länder* remained. Vorarlberg voted for union with Switzerland in May 1919, and Tirol also attempted to secede.

The constitutional settlement. In February 1919, elections for a constitutional assembly were held. The Social Democrats were returned as the largest single party, with 69 seats. Sixty-three were won by the Christian Socialists and 26 by the German Nationalists. When this assembly met (March 4), it had to make wide concessions to federalism in order to appease the provinces. In exchange, Vienna was also elevated to the rank of a state and the mayor made the equivalent of a state governor. This proviso subsequently enabled "Red Vienna" to pursue an autonomous policy, despite the fact that the *Bundesregierung* ("federal government") was controlled by the conservative parties from 1920 to 1934.

Autonomy of Vienna

The constituent assembly also settled the constitution of the federal republic (October 1, 1920). The State Council was abolished, and a bicameral legislative assembly, the Bundesversammlung, was established. The Bundesrat (upper house) was to exercise only a suspensive veto and was to be elected roughly in proportion to the population in each state. This represented a defeat for the federal elements in the states, which had wanted the Bundesrat to exercise an absolute veto and to be composed of equal numbers of members from each state. The lower house, or Nationalrat, was to be elected by universal suffrage on a basis of proportional representation. The Bundesversammlung in full session elected the president of the republic for a four-year term, but the federal government, with the chancellor at its head, was elected in the Nationalrat on a motion submitted by its principal committee; this committee was itself representative of the proportions of the parties in the house.

The foreign policy of Otto Bauer and representatives of

the major political parties had insisted firmly on *Anschluss* ("union") with Germany, and as late as 1921 unauthorized plebiscites held in the western provinces returned overwhelming majorities in favour of union with Germany. But article 88 of the peace treaty of Saint-Germain, signed on September 10, 1919, forbade *Anschluss* without the consent of the League of Nations and also stipulated that the republic should cease to call itself Deutschösterreich (German-Austria); it became the Republik Österreich. The Austrian claim for the German-speaking areas of Bohemia and Moravia was denied by the peace conference, and Austria had to recognize the frontiers of Czechoslovakia along slightly rectified historical administrative lines. The southern frontier with Yugoslavia was threatened by Yugoslav armed invasion, and it was finally decided that the question should be settled by a plebiscite, which, on October 10, 1920, returned a majority of 59 percent in favour of Austria. The German-speaking districts of western Hungary were to be ceded to Austria outright; but Austria, in the face of Hungarian resistance, was obliged to hold a plebiscite. The area of Sopron was finally restored to Hungary.

After the elections of February 1919, Renner formed another coalition government, but after a government crisis in the summer of 1920 a caretaker cabinet under the Christian Socialist Michael Mayr was formed. This government prepared the draft of the constitution and introduced it into parliament. After its approval, new elections were held, on October 17, 1920. The Christian Socialists were returned as the strongest party, gaining 82 seats, while the Social Democrats were reduced to 66 and the German Nationalists to 20. Mayr formed a new cabinet composed of Christian Socialists; the Social Democrats went into opposition and never returned to the government throughout the First Republic. This political division hardened, and no decisive change took place during the following years. The system of proportional representation combined with the ideological background of Austrian parties made oscillations of political allegiance unlikely. Of the two mass parties, the Social Democrats had an unshakable majority in Vienna (in which about a third of all the inhabitants of the republic lived), while the Christian Socialists had an equally secure majority among the Catholic peasants and the conservative classes, the latter consisting largely of army officers, landowners, and big business. The urban middle classes, hostile to both workers and peasants, became German nationalists. But German nationalism was not limited to the middle classes. The workers and even the peasants felt themselves to be Germans and also responded to the national appeal.

Christian Socialist electoral victory (1920)

Economic reconstruction. The main task of the non-Socialist governments in power in Austria from the autumn of 1920 on was to restore financial and economic stability. Between 1919 and 1921, the urban population of Austria lived largely on relief from the United States and Great Britain, and, although production improved, distress was heightened by inflation that threatened financial collapse in 1922. In October 1922, the chancellor, Ignaz Seipel, secured a large loan through the instrumentality of the League of Nations, enabling Austrian finances to be stabilized. In return, Austria had to undertake to remain independent for at least 20 years. The controller general appointed by the League of Nations reported in December 1925 that the Austrian budget had been balanced satisfactorily, and in March 1926 international financial supervision was withdrawn.

Seipel's success in October 1922 gave Austria some years of stability and made economic reconstruction and relative prosperity possible. In Socialist-controlled Vienna, an ambitious program of working class housing, health schemes, and adult education was carried out under the leadership of Karl Seitz, Hugo Breitner, and Julius Tandler. "Red Vienna" thus acquired a unique reputation in Europe.

Political strife. In 1920 all three major parties spoke in democratic terms. Despite democratic phrases, however, preparations for possible civil war had never been abandoned. The Christian Socialists, led by Seipel, a believer in strong government, were convinced that they had to protect the existing social order against a Marxist revolution. In the provinces, reactionary forces (the *Heimwehr*, or "home defense forces"), originally formed for defense against the Yugoslavs or merely against international disorder, gradually acquired Fascist tendencies. The Social Democrats felt that their social-reform program was endangered by reaction. They possessed their own armed force, the *Schutzbund* (Defense League), descended from the People's Guard of 1918, and they and the reactionary forces regularly demonstrated against each other. In 1927, in the course of a clash between members of the Schutzbund and certain reactionary forces at Schattendorf, an old man and a child were accidentally shot by the reactionaries. When the latter were acquitted by a Vienna jury on July 14, the Social Democrats called for a mass demonstration, which got out of hand and ended in the burning down of the ministry of justice. In street fighting between the police and the demonstrators, almost 100 persons were killed. The Social Democrats then launched a general strike, but this had to be called off after four days. Seipel had used the opportunity for a violent assertion of government authority. The balance between Socialist and non-Socialist forces in Austria was never secure after this decisive date.

Emergence of Fascist groups

The Christian Socialists, pressed increasingly by the *Heimwehr*, now began to take the offensive against the Social Democrats. Wilhelm Miklas, a leading Christian Socialist, was elected president as successor to the non-party Michael Hainisch, who had been in office since December 1920. There were repeated attempts to revise the constitution, principally with the object of strengthening the power of the executive. After protracted negotiations, a compromise was reached late in 1929. On December 7, 1929, a series of constitutional amendments gave increased powers to the president. Of particular importance were the rights to appoint ministers and issue emergency decrees. But Vienna preserved its autonomy, and the democratic principle was preserved against the far-reaching authoritarian demands of the *Heimwehr*. In the elections of November 1930, the Social Democrats were returned as the largest single party, with 72 seats. The Christian Socialists held 66, the German Nationalists 19, and the *Heimwehr*, now posing as a Fascist party on the Italian model, 8.

These political events were overshadowed by the great world economic crisis. Though the Social Democratic leaders believed that the crisis should be met by the orthodox means of deflation and spending cuts, they were resolved not to be compromised by supporting these measures and refused to enter a coalition government. On the other hand, in October 1931 they acquiesced in suspending the election of the president by direct popular vote, as had been provided by the constitution of 1929, and agreed to the re-election of Miklas by parliament for a further four years. The government, meanwhile, led by Otto Ender and Johann Schober, was driven to desperate devices in order to stave off collapse. Schober, leader of the middle class German Nationalists, launched the project for a customs union with Germany in March 1931; this provoked violent opposition from France and the alliance of the Little Entente (Czechoslovakia, Yugoslavia, and Romania) and was subsequently condemned by a majority of the International Court at The Hague. The bankruptcy in May 1931 of the Creditanstalt, the most influential banking house in Austria, brought Austria close to financial and economic disaster. This, together with the rise of the National Socialists in Germany, resulted in considerable support being given to the Nazis in Austria; and the provincial elections in 1932 showed that they were draining off votes from the conservative parties. The Nationalists began to demand a general election, and this demand was taken up by the Social Democrats, who saw a chance of winning a majority in parliament.

Proposed customs union with Germany (1931)

Authoritarianism: Dollfuss and Schuschnigg. After the election, when Engelbert Dollfuss came to form a Christian Socialist government on May 20, 1932, he could count on a majority of only one vote. Dollfuss be-

longed to a new generation that had been educated in the conservative conviction that the Western form of parliamentary government had been forced upon the central Europeans as a result of military defeat and Socialist revolution and that the political and social order could be restored only by the establishment of some kind of strong authority. The leaders of the Christian Socialist Party found themselves under attack from two ideological enemies, the Marxists and the Nazis, who apparently threatened the very basis of the conservative order. In reaction, Dollfuss thus determined to replace parliamentary government by an authoritarian system. The opportunity to do this came in March 1933, when, during a debate on a minor bill, an argument arose over alleged irregularities in the voting procedure. The president of the Nationalrat resigned; the two vice-presidents followed his example; and Dollfuss declared that parliament had proved unworkable. It never met again in full, and Dollfuss governed thenceforth by emergency decree.

The end of parliamentary government

By this time (spring of 1933), Adolf Hitler was in power in Germany, and Nazi propaganda for the incorporation of Austria was greatly increased. In this situation, Dollfuss turned to Italy for help, convinced that British and French aid would be ineffective. This shift in foreign policy can also be attributed to the fact that Dollfuss had to rely on the help of the *Heimwehr* to stay in power: the anti-Marxist *Heimwehr* had been in close contact with the Italian Fascist Party for a number of years, and when Dollfuss visited Benito Mussolini at Riccione in August 1933, he secured Mussolini's promise to protect Austria's independence only on the condition that he would give the *Heimwehr* a free hand in the destruction of the Social Democratic Party in Austria.

The Social Democrats were subjected to increasing provocation and on February 12, 1934, took to arms. Civil war followed. After four days of fighting, Dollfuss and the *Heimwehr* were victorious. The Social Democratic Party was declared illegal and driven underground. In the course of the same year, all political parties were abolished except the Vaterländische Front (Fatherland Front), which Dollfuss had founded in 1933 to unite all conservative groups. In April 1934, the rump of the parliament was brought together and accepted an authoritarian constitution. The executive was given complete control over the legislative branch of government; the elected assemblies disappeared and were replaced by advisory bodies, appointed in a complicated and futile fashion. The rights of man guaranteed under the democratic constitution were also swept away. "Republic" was removed from the official name of the state, which became merely the Federal State of Austria.

Dollfuss murdered by Nazis

On July 25, 1934, a group of Nazis seized the chancellery and attempted to proclaim a Nazi government. Dollfuss, whom they had taken prisoner, was murdered. The plan, however, miscarried: the Nazi party in the chancellery was compelled to surrender, and its leaders were executed; a Nazi rising in Styria was suppressed; and, faced with the mobilization of an Italian army on the Brenner Pass, Hitler repudiated his Austrian followers. Franz von Papen was sent as German ambassador to reduce Austria by other means. Kurt von Schuschnigg, who became chancellor on the death of Dollfuss, was a man of gentler personality and of less violent political passions. The administration of the authoritarian constitution was in the easygoing Austrian fashion, less oppressive than in Italy and Germany. Schuschnigg had a mild preference for restoring the Habsburgs, but he shrank from the international complications that this would involve. The regime drifted on without popular favour, weakened by the personal rivalries and ambitions of its leaders and sustained only by a guarantee from Italy. The temporary accord of Great Britain, France, and Italy in the "Stresa front" (April 1935) seemed to promise new security, but the Ethiopian crisis soon destroyed the unity of the Western powers, and Austria's isolation was complete when Hitler and Mussolini allied themselves in 1936. Schuschnigg had to negotiate a compromise with Germany, which was signed on July 11, 1936; Germany promised to respect Austrian sovereignty, and in return Austria acknowledged itself "a German state." Schuschnigg promised amnesty to the Nazis and held out a prospect of including some "nationally minded" Austrian in his ministry in the future. The agreement of July 1936 left Austria open to Nazi infiltration. In January 1938, the Austrian police discovered a new Nazi conspiracy. Schuschnigg hoped to defeat this by a meeting with Hitler, but at Berchtesgaden, where Hitler received Schuschnigg on February 12, 1938, Schuschnigg was faced with threats of massive military intervention in support of the Austrian Nazis. He had to agree to give a general amnesty to the Austrian Nazis and to include some leading Nazis in his cabinet; the ministry of the interior had to be entrusted to Arthur Seyss-Inquart, the spokesman of Austrian Nazis. The open agitation of the Nazis threatened to destroy the government's authority, and confidential contacts in the European capitals brought Schuschnigg to realize that he could not count on the support of the great powers. He therefore resolved to challenge Hitler alone. On March 9 he announced that a plebiscite would be held on March 13, 1938, to decide in favour of Austrian independence.

Anschluss and World War II. Hitler could not allow Schuschnigg's plebiscite to be held. Though the Austrian crisis had taken him unaware, he acted with energy and speed. Mussolini's neutrality was assured, there was a ministerial crisis in France, and the British government had made it known for some time that it would not oppose union of Austria with Germany. On March 11, 1938, two peremptory demands were made for the postponement of the plebiscite and for the resignation of Schuschnigg. Schuschnigg gave way, and German troops, accompanied by Hitler himself, entered Austria on March 12. A Nazi government in Austria, headed by Arthur Seyss-Inquart, was established and collaborated with Hitler in proclaiming the *Anschluss* on March 13. France and Great Britain protested against the methods used by Hitler but accepted the *fait accompli*, as did all other governments. A plebiscite on April 10, held throughout greater Germany, recorded a vote of more than 99 percent in favour of Hitler.

Hitler marches into Austria (March 1938)

Austria was absorbed into Germany for all purposes. Immediately after the invasion, the Nazis arrested the leaders of the Austrian political parties. Many Austrians, especially those of Jewish origin, went into exile, but the political antagonism that had previously weakened the status of the republic continued to block cooperation among the emigrés, as well as among the resistance groups that formed inside Austria.

The possibility of re-establishing an independent Austria was, however, far from dead, and, after the outbreak of World War II, the Allied governments began to reconsider their attitude toward the *Anschluss*. In December 1941, Stalin informed the British that the Soviet Union would regard the restoration of an independent Austrian republic as an essential part of the postwar order in central Europe, and, at the meeting of the foreign ministers of Great Britain, the U.S.S.R., and the United States in Moscow (October 1943), a declaration was published that declared the union with Germany null and void and pledged the Allies to restore Austrian independence. Though British Prime Minister Winston Churchill continued to make various proposals for setting up a central European federation comprising the former Habsburg lands and even southern Germany, the European Advisory Commission in London prepared its plans for the liberation and occupation of Austria on the assumption that Austria would return to sovereignty within the borders of 1937. When Soviet troops liberated Vienna on April 13, 1945, representatives from the resistance movement and the former political parties were allowed to organize.

THE SECOND REPUBLIC

The Allied occupation. On April 27, 1945, Karl Renner set up a provisional government composed of Social Democrats, Christian Socialists, and Communists and proclaimed the re-establishment of Austria as a democratic republic. The Western powers, afraid that the Renner government might be an instrument of Communist

expansion, withheld full recognition until the autumn of 1945. Because of similar suspicions, agreement on the division of Austrian zones of occupation was delayed until July 1945. Shortly before the Potsdam Conference (which stipulated that Austria would not have to pay reparations but assigned the German foreign assets of eastern Austria to the Soviet Union), control machinery was set up for the administration of Austria, giving supreme political and administrative powers to the military commanders of the four occupation armies. In September 1945, a Länderkonferenz of representatives of all provinces extended the authority of the Renner government to all parts of Austria. A general election held in November 1945, in which former Nazis were excluded from voting, returned 85 members of the Austrian People's Party (corresponding to the Christian Socialists of the prewar period), 76 Socialists (corresponding to the Social Democrats and Revolutionary Socialists), and 4 Communists. Renner was elected president of the republic; Leopold Figl, leader of the Austrian People's Party, became chancellor of a coalition cabinet. The government decided not to draft a new constitution but to return to the constitution of 1920, as amended by the laws of 1929. In June 1946, the control agreement of July 1945 regulating the machinery of Allied political supervision was modified by restricting Allied interference essentially to constitutional matters. Denazification laws passed in 1946 and 1947 eliminated Nazi influence in the public life of Austria. In the postwar years, the Austrian People's Party and the Socialists were the sole partners in a coalition government that was formed in proportion to the parties' strength in parliament. This principle of proportional representation, originally introduced in 1919, was to be an important factor in Austrian political life after 1945.

General elections of November 1945

From 1945 to 1952, Austria had to struggle for survival. After liberation from Nazi rule, the country faced complete economic chaos. Aid provided by the United Nations Relief and Rehabilitation Administration (UNRRA) and, from 1948, support given by the United States under the Marshall Plan made survival possible. Heavy industry and banking were nationalized in 1946, and, by a series of wage–price agreements, the government tried to control inflation. Reconstruction was especially difficult in the Soviet zone of occupation, where military commanders continuously interfered in political and economic affairs. The result was a migration of capital and industry from Vienna and Lower Austria to the formerly purely agricultural western provinces. This brought about a far-reaching transformation of the economic and social structure of the country. Austria remained occupied by U.S., British, French, and Soviet forces until 1955. A treaty restoring Austrian sovereignty was expected early, but the atmosphere of the Cold War made agreement among the former Allied powers impossible. In 1953, however, a heavy burden was removed from the Austrian economy when the Soviet government declared that it would pay its own occupation costs (as the United States had done since 1947). Thereupon, the British and the French followed suit.

Party alignments

In 1949 former Nazis were allowed to participate in the general election. The Union of Independents (later renamed as the Freedom Party), corresponding to the former German Nationalist group but free from ideological ties, won 16 seats in parliament. In the subsequent elections (1953, 1956, 1959, 1962), the relationship of the three parties remained stable.

When Renner died (December 31, 1950) Theodor Körner, the Socialist mayor of Vienna, was elected president of the republic by direct popular vote. On his death (January 4, 1957) the leader of the Socialist Party, Adolf Schärf, was elected president. After being re-elected in 1962, Schärf died in office in 1965. He was succeeded by Franz Jonas, former mayor of Vienna, who began a second term in April 1971.

The influence of the Socialists in the coalition government, which had been relatively strong under Leopold Figl's chancellorship, was reduced when the Austrian People's Party replaced Figl with Julius Raab in the spring of 1953 and had Reinhard Kamitz appointed minister of finance. The subsequent economic reconstruction and the advance to a prosperity unknown to Austrians since the years before World War I is generally identified with the so-called Raab–Kamitz course, based on a modified free-market economy. The nationalized steel industry, electrical-power plants, and oil fields, together with the privately owned lumber and textile industries and the tourist traffic, were the major economic assets.

The Berlin conference of the foreign ministers of the Big Four (France, Great Britain, the Soviet Union, and the United States; January 1954) raised Austrian hopes for the conclusion of a peace treaty. For the first time, Austria was admitted as an equal conference partner, but the failure of the foreign ministers to agree on the future of Germany again prejudiced Austria's chances. It appeared that the Soviet government was not prepared to forego its strategic advantages in Austria so long as Germany was not "neutralized." In February 1955 the Soviet government suddenly contacted the Austrian government and extended an invitation to bilateral negotiations in Moscow. An Austrian delegation visited Moscow in April 1955, and a memorandum was agreed on, according to which the Soviet government declared itself ready to restore full Austrian sovereignty and to evacuate its occupation troops in return for an Austrian promise to declare the country permanently neutral.

Restoration of sovereignty. The treaty was signed in Vienna on May 15, 1955, by the representatives of the four powers and Austria. It formally re-established the Austrian republic in its pre-1938 frontiers as a "sovereign, independent and democratic state." It prohibited *Anschluss* between Austria and Germany as well as the restoration of the Habsburgs. It guaranteed the rights of the Slovene and Croatian minorities in Carinthia, Styria, and Burgenland. The United Kingdom, the United States, and France relinquished to Austria all property, rights, and interests held or claimed by them as former German assets or war booty. The U.S.S.R., however, obtained tangible payment for the restoration of Austrian freedom. This included $150,000,000 for the confiscated former German enterprises which Austria had to buy back from the Administration of Soviet Property in Austria; $2,000,000 for the confiscated German assets of the First Danube Steamshipping Company; and 10,000,000 metric tons of crude oil as the price of Austrian oil fields and refineries which had been Soviet war booty. The state treaty came into force on July 27, 1955. By October 25 all occupation forces were withdrawn, and Austria recovered its freedom and sovereignty, lost on March 13, 1938. On October 26, 1955, a constitutional law of perpetual Austrian neutrality was promulgated. The Austrian government had never left any doubts that the pledge to neutrality could be interpreted only as a military one and never as an ideological one. Throughout the Soviet occupation, the Austrians had proved their anti-Communist attitude, and the spontaneous reaction of the Austrian people during the suppression of the Hungarian Revolution in 1956 demonstrated their sympathy with Western democratic ideas. Austria preserved political stability; changes in the personal and ideological structure of the government and political parties were effected without major political crisis.

Austria becomes a formal neutral

From 1962 disagreement over economic problems generated friction between the coalition parties. The annual budget led to grave disunity in the coalition, and in the autumn of 1965 the government resigned and called new elections. The election, held on March 6, 1966, brought a setback for the Socialist Party, and the People's Party was returned to parliament with an absolute majority. Negotiations for a new coalition government failed. The Socialists, led by former foreign minister Bruno Kreisky, went into opposition, and Josef Klaus formed the first one-party cabinet of the Second Republic. Contrary to widespread misgivings, the political stability of the country was not disturbed, and parliament was given new vigour and influence. In ensuing provincial elections, the Socialist Party demonstrated recovery from the setback of 1966, and in the national elections of 1970 the So-

cialist Party managed to win a plurality of votes, becoming the strongest party in parliament, with 81 seats, but falling short of a majority. After negotiations for a new coalition cabinet failed, in May 1970 Bruno Kreisky was appointed chancellor; he formed the first Austrian all-Socialist cabinet. Sensing increased support for the Socialists he called for new elections in October 1971. Now the Socialists won a clear majority of 93 seats.

Participation in European organizations

Austria became a member of the United Nations in 1955 and of the European Council in Strasbourg in 1956. Major problems of Austrian foreign relations were the conflict with Italy over South Tirol (Alto Adige) and the problem of association with the European Economic Community (EEC). During the Paris Peace Conference of 1946 an agreement was signed guaranteeing the rights of the German-speaking population of South Tirol. The Austrian government, claiming that the Italians had not lived up to their obligations, initiated bilateral talks. In the early 1960s, acts of terrorism committed by German-speaking chauvinists blocked the progress of the negotiations, but in 1969 agreement was finally reached on implementation of the guarantees provided in the 1946 agreement. In 1958 Austria joined the European Free Trade Association, but a special arrangement with the EEC was regarded as essential to the Austrian economy. Membership in the EEC would, however, present special difficulties, as Austria's neutral status might be compromised. Negotiations to resolve this problem were continuing in the early 1970s. (F.Fe.)

BIBLIOGRAPHY

General works: KARL and MATHILDE UHLIRZ, *Handbuch der Geschichte Österreichs und seiner Nachbarländer Böhmen und Ungarn*, 4 vol. (1927–44; vol. 1, 2nd ed., 1963), with excellent bibliography; H. HANTSCH, *Die Geschichte Österreichs*, 4th ed., 2 vol. (1959–68), a conservative, scholarly, pro-Habsburg account; E. ZOLLNER, *Geschichte Österreichs von den Anfängen bis zur Gegenwart*, 4th ed. (1970), a standard Austrian text from ancient times to the present; O. SCHULMEISTER (ed.), *Spectrum Austriae* (1957), a comprehensive, positive evaluation; R. RICKETT, *A Brief Survey of Austrian History* (1966); *Austrian History Yearbook*, published by Rice University, devoted exclusively to Austrian history.

Prehistory: R. PITTIONI, *Urgeschichte des österreichischen Raumes* (1954).

Roman period: A. BETZ, *Aus Österreichs römischer Vergangenheit* (1956).

Middle Ages to 1246: A.W.A. LEEPER, *A History of Medieval Austria* (1941); K. LECHNER, *Die Babenberger und Österreich* (1947); H. FICHTENAU, *Von der Mark zum Herzogtum: Grundlagen und Sinn des "Privilegium minus" für Österreich*, 2nd ed. (1965).

1246–1526: A. LHOTSKY, *Geschichte Österreichs seit der Mitte des 13. Jahrhunderts (1281–1358)* (1967).

1526–1648: J. LOSERTH, *Die Reformation und Gegenreformation in den innerösterreichischen Ländern* (1898); G. MECENSEFFY, *Geschichte des Protestantismus in Österreich* (1956); V. BIBL, *Maximilian II., der rätselhafte Kaiser* (1929); H. STURMBERGER, *Georg Erasmus Tschernembl* (1953); *Kaiser Ferdinand II. und das Problem des Absolutismus* (1957).

1648–1740: O. REDLICH, *Weltmacht des Barock*, 4th ed. (1961); *Das Werden einer Grossmacht, Österreich von 1700–1740*, 4th ed. (1962); T.M. BARKER, *Double Eagle and Crescent: Vienna's Second Turkish Siege and Its Historical Setting* (1967); A. CORETH, *Österreichische Geschichtsschreibung in der Barockzeit, 1620–1740* (1952); B. GRIMSCHITZ, R. FEUCHTMULLER, and W. MRAZEK, *Barock in Österreich* (1960); R.A. KANN, *A Study in Austrian Intellectual History: From Late Baroque to Romanticism* (1960).

1740–92: F. MAASS (ed.), *Der Josephinismus, Quellen zu seiner Geschichte in Österreich 1760–1790*, 5 vol. (1951–57); R.A. KANN (*op. cit.*).

1792–1848: A.H. SPRINGER, *Geschichte Österreichs seit dem Wiener Frieden 1809–1849*, 2 vol. (1863–65); V. BIBL, *Der Zerfall Österreichs*, 2 vol. (1922–24), see vol. 1, *Kaiser Franz und sein Erbe*; C.A. MACARTNEY, *The Habsburg Empire, 1790–1918* (1968).

1848–67: HEINRICH FRIEDJUNG, *Österreich von 1848–1860*, 2 vol. (1908–11); *Der Kampf um die Vorherrschaft in Deutschland, 1856 bis 1866*, 10th ed. (1916–17; abridged Eng. trans., *The Struggle for Supremacy in Germany*, 2 vol., 1935, reprinted 1966); R. KISZLING, *Die Revolution im Kaisertum Österreich*, 2 vol. (1848–49); R.A. KANN, *The Multinational Empire: Nationalism and National Reform in the Habsburg Monarchy, 1848–1918*, 3rd ed., 2 vol. (1964), rev. and enlarged German ed., *Das Nationalitätenproblem der Habsburgermonarchie*, 2nd ed., 2 vol. (1964), the standard work on the nationality problem, heavily documented; ADOLF BEER, *Die österreichische Handelspolitik im 19. Jahrhundert* (1891).

1867–1918: H. WICKHAM STEED, *The Hapsburg Monarchy*, 4th ed. (1919; 1914 ed. reprinted 1969), a critical contemporary view; A.J.P. TAYLOR, *The Habsburg Monarchy, 1809–1918*, new ed. (1948), prematurely dated but still provocative; FRITZ FELLNER, *Der Dreibund*, 2nd ed. (1963), a reappraisal of Austro-Hungarian Alliance systems; A.J. MAY, *The Hapsburg Monarchy, 1867–1914* (1951), a comprehensive and balanced study; *The Passing of the Hapsburg Monarchy, 1914–1918*, 2 vol. (1966); C.A. MACARTNEY (*op. cit.*), a scholarly work particularly strong on Hungary (excellent bibliography); B. JELAVICH, *The Habsburg Empire in European Affairs, 1814–1819* (1969), concentrates on foreign relations; R.A. KANN (*op. cit.*).

1918 to the present: O. BAUER, *Die Öesterreichische Revolution* (1923; Eng. trans., *The Austrian Revolution*, 1925), Socialist interpretation from one of the leading protagonists; C.A. MACARTNEY, *The Social Revolution in Austria* (1927), the only social history of Austria available; M. MACDONALD, *The Republic of Austria, 1918–1934* (1946), a concise and objective account; C.A. GULICK, *Austria from Habsburg to Hitler*, 2 vol. (1948), the most comprehensive work on the interwar period (pro-Socialist); R. HISCOCKS, *The Rebirth of Austria* (1953), the first attempt to write the history of post-1945 Austria; KARL R. STADLER, *Austria* (1971), a scholarly history of Austria in the 20th century, especially strong on the period since 1938; E. WEINSIERL and K. SKALNIK (eds.), *Österreich. Die Zweite Republik, 1945–1970*, 2 vol. (1972), the most comprehensive account to date.

(E.Zo./R.A.K./F.Fe.)

Austro-Asiatic Languages

Austro-Asiatic languages are spoken by approximately 40,000,000 people scattered throughout Southeast Asia and eastern India. The family comprises about 150 languages, most of them having numerous dialects. Khmer, Mon, and Vietnamese are culturally the most important and have the longest recorded history. The rest are languages of non-urban minority groups written, if at all, only recently. The family is of great importance as a linguistic substratum for all Southeast Asian languages.

Superficially, there seems to be little in common between a monosyllabic tone language like Vietnamese and a polysyllabic toneless Munda language like Mundari of India; every recent study, however, confirms the underlying unity of the family. The date of separation of the three main Austro-Asiatic subfamilies—Munda, Nicobarese, and Mon-Khmer—has never been estimated and must be placed well into prehistory. Within the Mon-Khmer subfamily itself, 12 main branches are distinguished; glottochronological estimates of the time during which specific languages have evolved separately from a common source (derived from a comparison of similarities in selected vocabulary items) indicate that these 12 branches all separated at approximately 1000 to 2000 BC.

Munda, Nicobarese, and Mon-Khmer subfamilies

Relationships with other language families have been proposed, but, because of the considerable time depths involved and the scarcity of reliable data, it is very difficult to present a solid demonstration of their validity. In 1906 Wilhelm Schmidt, a German priest and anthropologist, classified Austro-Asiatic together with the Austronesian family (formerly called Malayo-Polynesian) to form a larger family called Austric. More recently, Paul K. Benedict, a U.S. scholar, accepting the Austric theory, extended it to include the Tai-Kadai family of Indochina and Burma and the Miao-Yao family of China, together forming an "Austro-Tai" superfamily.

Regarding subclassification within Austro-Asiatic, there have been several controversies. Schmidt, who first attempted a systematic comparison, included in Austro-Asiatic a "mixed group" of languages containing "Malay" borrowings and did not consider Vietnamese to be a member of the family. On the other hand, some of his critics contested the membership of the Munda group of eastern India. The "mixed group," recently called Chamic, is now considered to be Austronesian. It includes Cham, Jarai, Rhade, Chru, Roglai, and Hroy, and repre-

sents an ancient migration of Indonesian peoples into southern Indochina. As for Munda and Vietnamese, the recent works of the German linguist Heinz-Jürgen Pinnow on Kharia and of the French linguist André Haudricourt on Vietnamese tones have shown that both language groups are Austro-Asiatic.

Classification of the Austro-Asiatic languages. The work of classifying and comparing the Austro-Asiatic languages is still in the initial stages. In the past, classification has been done mainly according to geographic location. For instance, Khmer, Pear, and Stieng, all spoken on Cambodian territory, were all lumped together, although they actually belong to three different branches of the Mon-Khmer subfamily.

The main lines of the following classifications have been established by the glottochronological method, which involves the statistical comparison of similarities.

Munda subfamily (East India, *c.* 6,000,000 speakers)
- A. North Munda:
  - 1. Korku
  - 2. Kherwarian:
    - a. Santali
    - b. Mundaric
      - i. Mundari
      - ii. Asuri
      - iii. Bhumij
      - iv. Birhor
      - v. Ho
      - vi. Koda
      - vii. Korwa
      - viii. Turi
- B. South Munda:
  - 1. Central Munda:
    - a. Kharia
    - b. Juang
  - 2. Koraput Munda:
    - a. Sora, Gorum
    - b. Geta, Gutob, Remo
- C. (?) West Munda: Nahali

Nicobarese subfamily (Nicobar Islands, *c.* 7,000 speakers)
- A. North Nicobar:
  - 1. Car
  - 2. Chowra
  - 3. Teressa, Bompaka
- B. Central Nicobar:
  - 1. Camorta
  - 2. Nancowry
  - 3. Trinkut, Katchall
- C. South Nicobar:
  - 1. Coastal Great Nicobar
  - 2. Little Nicobar
- D. Inland Great Nicobar: Shompe

Mon-Khmer subfamily (continental Southeast Asia, *c.* 35,000,000 speakers)
- A. Khasi branch (Assam, *c.* 200,000 speakers)
  - 1. Standard Khasi
  - 2. Lyngngam
  - 3. Synteng
  - 4. War
- B. Palaungic branch (Burma, Thailand, Yunnan Province, over 1,000,000 speakers)
  - 1. Palaung
  - 2. Wa
  - 3. Riang-Lang
  - 4. Danaw
  - 5. Lawa
  - 6. Kawa
  - 7. (?) Khamet
  - 8. (?) Mang
  - 9. Bulan
  - 10. Angku
- C. Monic branch (Burma, Thailand, *c.* 500,000 speakers)
  - 1. Mon
  - 2. Niakuol
- D. Khmuic branch (Laos, Thailand, number unknown)
  - 1. Khmu'
  - 2. Mal
  - 3. Mrabri
  - 4. Yumbri
  - 5. Khao
  - 6. Tayhat
  - 7. (?) Puôc
  - 8. (?) Lamet
  - 9. Tin
  - 10. (?) Kha Kon Ku'
  - 11. Kha Kwang Lim
  - 12. (?) Kha Doi Luang
- E. Viet-Muong branch (Vietnam, [?] 23,000,000 speakers)
  - 1. Vietnamese:
    - a. North upland
    - b. Lowland (several major dialects)
  - 2. Muong: Pi, Thàng, Tông, Wang
  - 3. Mày
  - 4. Arem
  - 5. Tày Pong
  - 6. (?) Sach
  - 7. (?) Nguôn
  - 8. (?) Hung Khong Khêng
- F. Katuic branch: (Vietnam, Laos, Cambodia, over 200,000 speakers)
  - 1. Katu
  - 2. Kantu, High Katu
  - 3. Phuâng
  - 4. Brũ
  - 5. Pacoh
  - 6. Taoih
  - 7. Ngeq, Nkriang
  - 8. Kataang
  - 9. Kuy
  - 10. Lor
  - 11. Leu
  - 12. Ir
  - 13. Tong
  - 14. Souei
  - 15. So
  - 16. Alak
  - 17. (?) Kasseng
  - 18. (?) Tiari
- G. Bahnaric branch (mostly South Vietnam, over 550,000 speakers)
  - 1. South Bahnaric:
    - a. Stieng
    - b. Chrau
    - c. Srê
    - d. Mnong: South Central, East,
  - 2. West Bahnaric:
    - a. Loven
    - b. Nyahöñ, Prouac
    - c. Oi, The
    - d. Lave
    - e. (?) Brao, Kru'ng, Kravet
    - f. Sok
    - g. Sapuan
    - h. Cheng
    - i. (?) Suq
  - 3. North Bahnaric:
    - a. Bahnar
    - b. Rengao
    - c. Sedang
    - d. Jeh, Halăng
    - e. Mo'no'm
    - f. Kayo'ng
    - g. Hrê
    - h. Cua
    - i. Takua
    - j. To'drah
    - k. (?) Duan
- H. Pearic Branch (Cambodia, [?] 5,000 speakers)
  - 1. Pear
  - 2. Chong
  - 3. Samre
  - 4. Angrak
  - 5. Sa'och

I. Khmer (Cambodia, Thailand, South Vietnam, [?] 5,500,000 speakers)
J. Jahaic branch (Malaya and Thailand, *c.* 2,000 speakers)
1. Tonga
2. Kensiu, Kintak
3. Jahai
4. Menriq
5. Mintil, Batek
6. Che Wong
K. Senoic branch (Malaya, *c.* 30,000 speakers)
1. Temiar, Lanoh, Semnam
2. Semai
3. Jah Hut
L. Semelaic branch (Malaya, *c.* 5,000 speakers)
1. Mah Meri
2. Semelai, Temoq, Semaq Bri

Khmer and Vietnamese: national languages

Numerically the most important, the Khmer and Vietnamese languages are also the only national languages of the Austro-Asiatic group and are regularly taught in schools and used in the mass media and on official occasions. Speakers of most other Austro-Asiatic languages are under strong social and political pressure to become bilingual in the official languages of the national unit in which they live. Most groups are too small or too scattered to win recognition, and, for many, the only chance of cultural survival lies in retreating to a mountain or jungle fastness, an old Austro-Asiatic tradition.

Alternate Names of Languages and Dialects

Alakong	Bahnar	Mal	T'in (Khmuic)
Amok	Angku (Palaungic)	Mos	Tonga (Jahaic)
Attouat	Katu	Menik, Monik	Kintak (Jahaic)
Boda	Gutob	Maa'	Srê (S Bahnaric)
Bahr	Pear	Mapä	Lawa (Palaungic)
Besisi	Mah Meri (Semelaic)	Northern Sakai	Temiar (Senoic)
		Pareng	Gorum (Munda)
Be	Wa	Pray	T'in (Khmuic)
Biat	C Mnong (S Bahnaric)	Phi Tong Luang	Yumbri (Khmuic)
		Pnār	Synteng (Khasi)
Bodo, Bondo	Gutob (Munda)	Porr	Pear (Pearic)
Bondo	Remo	Phnong	Pear (Pearic)
Boloven	Loven (W Bahnaric)	Pangan	Batek, Jahai (Jahaic)
Bo'nâm	Mo'no'm (N Bahnaric)	Peguan	Mon (Monic)
		Ple-Temer	Temiar (Senoic)
Cambodian	Khmer	Rmang	Stieng (S Bahnaric)
Câu	Khmu'		
Chaoban	Niakuol (Monic)	Ro'teâng	Sedang (N Bahnaric)
Chau Ma	Srê (S Bahnaric)		
Chawng, Chong	Pear	Rlâm, Ro'lo'm	East Mnong (S Bahnaric)
Central Sakai	Semai (Senoic)		
Davak, Davach	Hrê (N Bahnaric)	Ruc	Mang (Viet-Muong)
Darang	Palaung		
Dedang	Rengao (N Bahnaric)	Soai	Kuy (Katuic)
		Sabub'n	Lanoh (Senoic)
Didey	Geta' (Munda)	Sakai	Senoic
Dié	Jeh (N Bahnaric)	Semang	Jahaic, Lanoh (Senoic)
Eastern Sakai	Jah Hut (Senoic)		
Ernga	Korwa (Munda)	Sisi	Mah Meri (Semelaic)
En	Wa (Palaungic)		
Gadaba	Gutob (Munda)	Savara	Sora (Munda)
Galler	Brū (Katuic)	Sodia	Gutob (Munda)
Gar	East Mnong (S Bahnaric)	Theng	Khmu'
		Tamun	Chrau (S Bahnaric)
Hasada	Mundari (Munda)		
Hotêang	Sedang (N Bahnaric)	Tri	Brū (N Bahnaric)
		Tembe	Temiar (Senoic)
Jro	Chrau (S Bahnaric)	Tala	Srê (S Bahnaric)
		Talaing	Mon (Monic)
Juru	Loven (N Bahnaric)	Tai Loi	Wa (Palaungic)
		Umpai	Lawa (Palaungic)
Khamuk	Khmu' (Khmuic)	Va, Vü	Wa (Palaungic)
Kohl, Kol	Santali (Munda)	Ve	Lave (W Bahnaric)
Ko'ho	Srê (S Bahnaric)		
Kuoy	Kuy (Katuic)	Xa	Khmu' (Khmuic)
Leu	Brū (Katuic)	Xhá Mang	Mang (Palaungic)
Lat	Srê (S Bahnaric)	Yang Lang, Yang Shek	Riang-Lang (Palaungic)
La-oop	Lawa (Palaungic)		

Phonological characteristics. In order to make general statements about the phonology (sound systems) of Austro-Asiatic languages, it is necessary to exclude Munda and Vietnamese, which, having been under the influence of Indian and Chinese languages, respectively, have acquired very divergent characteristics. The usual Austro-Asiatic word structure consists of a major syllable sometimes preceded by one or more minor syllables. A minor syllable has one consonant and one minor vowel. Most languages have only one type of minor vowel or, sometimes, three (*e.g.*, *a*, *i*, or *u*); others may also have vocalic nasal sounds, which are produced by releasing the breathstream through the nose, or liquids (*l* and *r* sounds) as minor vowels. Major syllables are composed of one or two consonants, followed by one major vowel and usually one final consonant.

Consonants. A typical consonant system for an Austro-Asiatic language would be the following (the symbols used are from the International Phonetic Alphabet):

p	t	c	k	ʔ		
b	d	j	g			
ɓ	ɗ					
m	n	ɲ	ŋ			
w	r	l	s	y	h	

Some languages (*e.g.*, Pearic, Semelaic) have an aspirated series of consonants, p^h, t^h, c^h, k^h, in which the sounds have an accompanying audible small puff of air, and many have no voiced stops at all (Monic, Khmer, Pearic). (Stops are consonants made with complete stoppage of the breathstream at some point in the vocal tract; voiced stops such as *b*, *d*, and *g* are produced with vocal cord vibration, as opposed to voiceless sounds produced without vocal cord vibration.) The imploded ɓ and ɗ are sounds pronounced by briefly drawing the air inward, causing suction, and are not truly voiced; they have sometimes been called preglottalized sounds or "semivoiceless" sounds. These imploded stops are found only in a few branches of Mon-Khmer (*e.g.*, Mon, Khmer, Bahnaric), but it is possible that they existed in the ancestral language, called Proto-Mon-Khmer. Pre-glottalized nasals and liquids (*i.e.*, nasal and liquid sounds preceded by a glottal stop) are also found, sometimes as single distinctive sounds (unit phonemes) and sometimes as consonant clusters. In final position, all consonants except voiced stops can be found, but in several languages (*e.g.*, Mon, Sedang, Palaung) the number of possibilities is more reduced. Final stops are pronounced without release, nasals are often decomposed (*e.g.*, a final *m* becomes pronounced as b^m), and *s* sounds usually tend toward *h* sounds. Palatal consonants (*č* and *ñ*), produced with the blade of the tongue touching the hard palate, are commonly found at the end of words, a feature that sets Austro-Asiatic languages apart from the other languages of South Asia.

Vowels. Also characteristic of the Austro-Asiatic languages is an extraordinary variety of major vowels: systems of 30 to 35 different vowels are not uncommon (Bru has 41 distinctive vowel sounds [phonemes]). Four degrees of height are often distinguished in front and back vowels as well as in the central area. Diphthongs are not rare. Vowel length is usually distinctive: a normal vowel may contrast with an extra-short vowel of the same quality. Nasal vowels are found in several branches of the family, but, in any one language, they do not occur very frequently. A few dialects of Palaung, Wa, and North Bahnaric have a simple tone system, usually high versus low tone, but this is not typical of the Austro-Asiatic family. The Viet-Muong branch is the only one to have developed complex tone systems, probably influenced by non-Austro-Asiatic languages.

Registers, or voice qualities

Much more typical of the Austro-Asiatic family is a contrast between two series of vowels pronounced with different voice qualities, which are called registers. The voice may have a "creaky" register, a "breathy" register, or a normal one. Mon, Khmer, Jeh, Sedang, and some Palaung dialects have a two-way distinction of this sort. There is a controversy regarding the historical origin of the registers. Some believe that they were found in the original Proto-Mon-Khmer language; others, who seem to hold the more likely theory, propose that they are independent innovations in each branch, representing a transitional state from toneless to tonal languages.

Grammatical characteristics. *Morphology.* In morphology (word formation), Munda and Vietnamese again show the greatest deviations from the norm. Munda

languages have an extremely complex system of prefixes, infixes (elements inserted within the body of a word), and suffixes. Verbs, for instance, are inflected for person, number, tense, negation, mood (intensive, durative, repetitive), definiteness, location, and agreement with the object. Furthermore, derivational processes indicate intransitive, causative, reciprocal, and reflexive forms. On the other hand, Vietnamese has practically no morphology.

Between these two extremes, the other Austro-Asiatic languages have many common features. (1) Except in Nicobarese, there are no suffixes. A few languages have enclitics, certain elements attached to the end of noun phrases (possessives in Semai, demonstratives in Mnong), but these do not constitute word suffixes. (2) Infixes and prefixes are common, so that only the final vowel and consonant of a word root remain untouched. It is rare to find more than one or two affixes (*i.e.*, prefixes or infixes) attached to one root; thus, because roots are mostly monosyllabic, the number of syllables per word remains very small. (3) The same prefix (or infix) may have a wide number of functions, depending on the noun or verb class to which it is added. For instance, the same nasal infix may turn verbs into nouns and mass nouns into count nouns. Sometimes, these different functions have similar meanings: for instance, reduplication, the repetition of a word or word element, may indicate plurality in nouns and repetition in verbs. This phenomenon may be widespread enough to make the distinction between basic word classes very unclear and questionable. (4) Many affixes are found only in a few fossilized forms and have often lost their meaning. (5) Expressive language and wordplay are embodied in a special word class called "expressives." These are sentence adverbials that describe noises, colours, light patterns, shapes, movements, sensations, emotions, aesthetic feelings, and so on. Some sort of symbolism, perhaps based on synaesthesia, is often observable in these words and serves as a guide for individual coinage of new words. The forms of the expressives are thus quite unstable, and the additional effect of wordplay can create subtle and endless, sometimes apparently empty, structural variations. For example, in Bahnar one can say /pha:m lɛč həmrɔ:ŋ həmra:ŋ "blood flows həmrɔ:ŋ həmra:ŋ (like a torrent, irregularly)." (The slash marks indicate that the symbols enclosed are phonetic, standing for speech sounds rather than letters of the alphabet. Many Austro-Asiatic languages do not have their own writing systems and are thus recorded here in phonetic transcription.) In Semai /slɯ:č, səslɯ:č, səralɯ:č, səralī:č, srlī:č, shu:č, səsrhu:č, səralī:m/ and many other forms describe "a massage on oily skin, a snake's creeps, shiny fur, noodles in Chinese soup," and so on. The Semai forms /ɲəɲ pəlayə:ɲ, lɛɲ kəlayɯ:ɲ, pəɲ pəlayã:ɲ, puɲ pəlayõ:ɲ/ all describe "oversize hat, opening of parachute, flying disc, ridiculously large ears" and are based on wordplay with the borrowed Malay noun /payuŋ/, meaning "umbrella." Spontaneous expression rather than rational communication seems to be the dominant function of these expressives.

Use of "expressives"

Syntax. In syntax, possessive and demonstrative forms and relative clauses follow the head noun; if particles are found, they will be prepositions, not postpositions (elements placed after the word to which they are primarily related), and the normal word order is subject–verb–object. There is usually no copula equivalent to the English verb "be." Thus, an equational sentence will consist of two nouns or noun phrases, separated by a pause; *e.g.*, in Rengao, /klan, bəs kən/, literally "python, snake large," means "a python (is) a large snake" or "pythons (are) large snakes." Predicates corresponding to the English "be + adjective" usually consist of a single intransitive (stative) verb; *e.g.*, in Khmer, /srəy nuh, lᵖɔ:/, literally "girl that, pretty," means "that girl (is) pretty." Ergative constructions (in which the agent of the action is expressed not as the subject but as the instrumental complement of the verb) are quite common; *e.g.*, in Semai, the ergative sentence /tlɛy ʔadɛh ʔɲ-caaʔya ʔɛɲ/, literally, "banana this I-ate by me," means "I ate this banana" as does the active sentence: /ʔɛɲ ʔɲ-caaʔ tlɛy ʔadɛh/ "I I-ate banana this." Also noteworthy are sentence final particles that indicate the opinion, the expectations, the degree of respect or familiarity, and the intentions of the speaker. Munda syntax, here again, is radically different, having a basic subject–object–verb word order, like the Dravidian languages of India. It is quite conceivable that the complexity of Munda verb morphology is a result of the historical change from an older subject–verb–object to the present subject–object–verb basic structure.

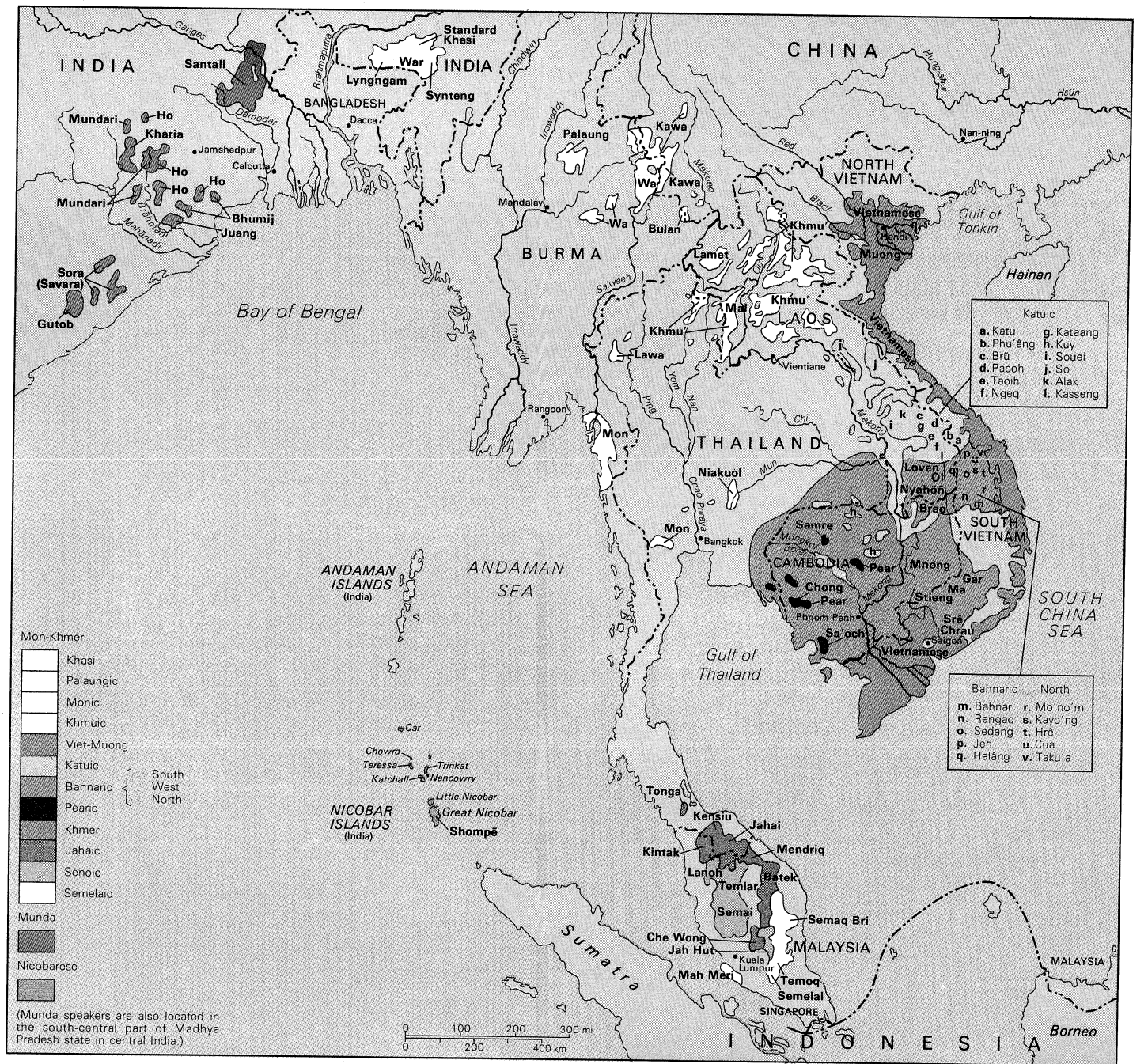
Vocabulary. The composition of the vocabulary of the Austro-Asiatic languages reflects their history. Vietnamese, Mon, and Khmer, the best known languages of the family, came within the orbit of larger civilizations and borrowed without restraint—Vietnamese from Chinese, Mon and Khmer from Sanskrit and Pāli. At the same time, they have lost a large amount of their original Austro-Asiatic vocabulary. It is among isolated mountain and jungle groups that this vocabulary is best preserved. But there, other disruptive forces are at work. For instance, animal names are subject to numerous taboos; and the normal name is avoided in certain circumstances (*e.g.*, hunting, cooking, eating, and so on). A nickname is then invented, often by using a kinship term ("uncle," "Grandfather") followed by a pun or an expressive adverb describing the animal. In the course of time, the kinship term is abbreviated (thus many animal names begin with the same letter), the normal name is forgotten, and the nickname becomes standard. It is then again avoided, and the process is repeated. There are also taboos on proper names; *e.g.*, after a person's death, his name and all words resembling it are avoided and replaced by metaphors or circumlocutions. This may explain why, for instance, the Nicobarese languages, which seem closely related, have few vocabulary items in common. In general, new words and fine shades of meanings can always be introduced by wordplay and from the open-ended set of expressive forms. Borrowings from the nearest majority languages are also common.

Word taboos

Writing systems and texts. Two Austro-Asiatic languages have developed their own orthographic systems and use them to this day. For both scripts, the letter shapes and principles of writing were borrowed from Indian alphabets (perhaps those of the Pallava kingdom in South India) that were in use in Southeast Asia at the time. Both Austro-Asiatic groups modified these alphabets in their own way, to suit the complex phonology of their languages. The most ancient inscriptions extant are in Old Mon (6th century AD), soon followed by Old Khmer in the early 7th century. The monuments of Burma, Thailand, and Cambodia have preserved a large number of official inscriptions in these two languages. Both alphabets were in turn used as models by other peoples for writing their own languages, the Thais using Khmer letters and the Burmese using Mon letters. The religious literature in Old and Middle Mon played a very important role in the spreading of Theravāda Buddhism to the rest of Southeast Asia.

Because Vietnam was a Chinese province for a thousand years, the Chinese language was used and written there for official purposes. In the course of time (perhaps as early as 8th century AD), a system called Chunom (popular writing) was developed for writing the Vietnamese language with partly modified Chinese characters. About 1650, a French missionary, Alexandre de Rhodes, devised a systematic spelling for Vietnamese, based on its distinctive sounds (phonemes). It uses the Latin (Roman) alphabet with some additional signs and several accents to mark tones. In this alphabet, tone was, for the first time in the world, recognized as a functional element and was systematically noted. At first, and for a long time, the use of this script was limited to Christian contexts, but it spread gradually, and in 1910 the French colonial administration made its use official. Now called Quoc-ngu (national language), it is learned and used by all Vietnamese.

Most other Austro-Asiatic languages have been written for less than a century; the literacy rate remains very low



Distribution of the Austro-Asiatic languages.

with a few exceptions (*e.g.*, Khasi). Dictionaries and grammars have been written only for the most prominent languages, with traditional and often insufficient methods. Many languages (*e.g.*, Wa, Kuy, Stieng, Pacoh, Katu, Muong) have only been described briefly in a few articles, and many more (Semelai, Puman, Sa'och, Riang, Lawa, Mrabri) are little more than names on the map.

BIBLIOGRAPHY. H.L. SHORTO, J.M. JACOB, and E.H.S. SIMMONDS (comps.), *Bibliographies of Mon-Khmer and Tai Linguistics* (1963), is an unannotated bibliography of linguistic books and articles from the beginning (1790) to 1960, which does not include the Munda subfamily or the Viet-Muong branch but incorporates the (Austronesian) Cham languages into Mon-Khmer. See W. SCHMIDT, *Grundzüge einer Lautlehre der Mon-Khmer-Sprachen* (1906); *Die Mon-Khmer-Völker: Ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens* (1906; French trans., "Les Peuples Mon-Khmer: trait-d'union entre les peuples de l'Asie Centrale et de l'Austronésie," in *Bulletin de l'École française d'Extrême-Orient*, 7: 213–263 and 8:1–35, 1907–08); and W.W. SKEAT and C.O. BLAGDEN, *Pagan Races of the Malay Peninsula*, 2 vol. (1906). Schmidt's articles for the first time supported the Austro-Asiatic hypothesis with lexical, phonological, and morphological evidence. They remain until today the basic work of Austro-Asiatic studies. Blagden compiled a very large comparative vocabulary but did not attempt any analysis. H.J. PINNOW, *Versuch einer historischen Lautlehre der Kharia-Sprache* (1959), an ambitious project with somewhat uncertain results, contains an analysis and systematic comparison of the phonologies of Munda languages and establishes connections with the rest of the Austro-Asiatic group. The *Mon-Khmer Studies* (Linguistic Circle of Saigon), 4 vol. (1964–), are collections of short technical articles mostly on the Montagnard languages of South Vietnam, with topics varying from basic vocabulary to phonology, morphology, syntax, folk taxonomies, and oral literature.

(G.Di.)

Austronesian Languages

The Austronesian language family, also called Malayo-Polynesian, consists of languages spoken in almost the whole of Malaysia and the Indonesian Archipelago, all of the Philippines, parts of Vietnam, Cambodia, and Taiwan (Formosa), Madagascar, and on all of the main island

groups of the South and Central Pacific (except for Australia and a large part of New Guinea, which contain languages belonging to other stocks). In terms of the number of its languages and of their geographic spread, the Austronesian language family is among the world's largest.

Austronesian languages are generally divided into two primary subgroups: a Western, or Indonesian, branch contains perhaps 200 languages, including such well-known tongues as Malay, Indonesian, Javanese, and Pilipino, the national language of the Philippines (based on Tagalog); and an Eastern branch, more commonly termed Oceanic, comprises about 300 small languages, scattered throughout the South and Central Pacific, the best known of which are the Polynesian group and Fijian. The classification of a small residue of languages is uncertain. While around 150,000,000 people speak languages belonging to the Western Austronesian branch, only about 1,000,000 speak Oceanic languages.

The challenge of piecing together the complex history of a family of languages with such an enormous distribution, and with speakers of such great cultural and physical diversity, has attracted scholars to the study of Austronesian languages. Recently, the Austronesian languages have also received attention because of their structural characteristics and, in addition, have served as a testing ground for subgrouping methods and theories of linguistic change.

HISTORY AND CLASSIFICATION

Resemblances between the languages of Madagascar, the East Indies, and Polynesia were first pointed out in 1706 by Hadrian Reland, a Dutch scholar. It was not until the second half of the 19th century, however, that the languages of the intervening island groups of Melanesia and Micronesia were recognized as belonging to the same family. At that point it became customary to divide Austronesian languages into four groups coinciding with the geographic regions known as Indonesia, Melanesia, Micronesia, and Polynesia. More recently, it has become clear that such a classification is not linguistically valid. Though arguments continue over the position of certain languages, it is generally agreed that most Austronesian languages fall into only two groups. The majority of the languages formerly assigned to the Melanesian and Micronesian divisions, together with the Polynesian group, form a single, although internally very diverse, subgroup: Eastern Austronesian (more often called Oceanic). The Western division coincides fairly closely with the old Indonesian grouping, comprising most, and possibly all, Austronesian languages spoken west of New Guinea, together with two found in Micronesia.

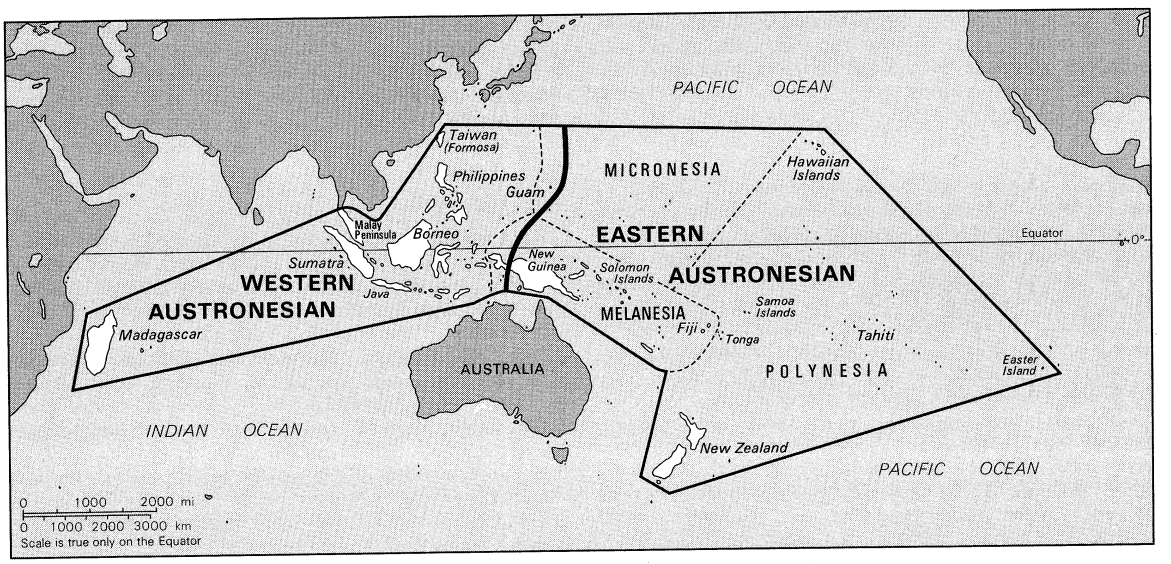
Eastern and Western Austronesian groups

The main areas of disagreement in classification concern the position of the Formosan languages, of certain languages located in eastern Indonesia and on or near the western tip of New Guinea, and of the more divergent languages in Melanesia. The Formosan languages are usually treated as Western Austronesian, though some scholars have suggested that they represent a third primary branch of Austronesian. The languages of east Indonesia, including the western end of New Guinea, more diverse than those of west Indonesia and the Philippines, are sometimes treated as a single subgroup of Western Austronesian, and sometimes regarded as composed of a number of primary branches, each coordinate with Oceanic and with a group comprising the remaining members of Western Austronesian. Similarly troublesome are a number of languages of the north coast of New Guinea and certain regions of insular Melanesia. These languages share very few related words with each other and with other languages in the family, although in grammar they usually show quite strong resemblances to members of the Oceanic subgroup. They are generally regarded as aberrant Oceanic languages, but other theories have been advanced to explain their divergent character.

Proto-Austronesian. A protolanguage is a parent language, or early form of a language or group of languages. It can either be a hypothetical reconstruction or may actually exist in written records, as does Latin, the protolanguage of the Romance tongues. There are no actual records of Proto-Austronesian, but it has been the subject of research and reconstruction by linguists. Because it is clear that the Proto-Austronesian speech community possessed agriculture and may have been responsible for its introduction—along with that of several other important cultural innovations—into the Pacific, the reconstruction of Proto-Austronesian and the history of the dispersal of Austronesian languages has been of considerable concern to culture historians as well as to linguists.

Location. The location of Proto-Austronesian has been the subject of much speculation but little systematic investigation. At various times in the past the parent tongue has been placed somewhere in the Southeast Asian mainland, South China, and even in India and Mesopotamia. There is increasing evidence of an archaeological, geographic, and linguistic nature, however, that the homeland lay in the region of Indonesia and New Guinea.

Chronology. The best evidence for dating and determining the directions of the dispersal of Austronesian languages comes from Oceania, where it is easier to correlate linguistic and archaeological findings than in the west. It is clear that Austronesian languages were already in Fiji and Tonga, near the eastern margin of Austronesian territory, by 1000 BC. Archaeological excavations indicate that the coastal area of Fiji was widely settled by that date, that by then Tonga also had been settled by a people with a material culture essentially iden-



Major divisions of the Austronesian languages.

tical to that of the first inhabitants of Fiji, and that Samoa was inhabited a few centuries later. In each place there is continuity of material culture—and, one may assume, continuity of language—right through to the period of European contact. The first Europeans found Fiji and Polynesia (and indeed the whole of the neighbouring island groups—the New Hebrides, New Caledonia, and Micronesia) to be occupied exclusively by Austronesian-speaking peoples. Glottochronology, a technique for dating the division between languages (known as linguistic splits) based on the assumption that there is a stable rate of basic vocabulary replacement in languages, places the separation of the Fijian and Polynesian subgroups at between 3,000 and 4,000 years ago. The divergence of Proto-Polynesian into separate branches is dated at between 1,800 and 2,500 years ago.

These are relatively recent branchings, far down on the Oceanic limb of the Austronesian family tree, and it follows that the separation of the common ancestor of Fijian and Polynesian from more distant branches of Oceanic must have occurred somewhat earlier. Glottochronological estimates indicate that diversification of Austronesian languages began around 4,000 to 5,000 years ago in the New Hebrides, in New Caledonia, and in the Solomons, and earlier still in the region of New Guinea. Very little archaeological work has been done in these areas, but assemblages of artifacts similar to those found in early Fijian and Tongan sites have been unearthed in New Caledonia and the central New Hebrides and dated at 800 BC and 600 BC respectively. The general trend, at least, is clear. The dispersal of Austronesian languages in Oceania cannot have begun later than around 2000 BC, with 3000 BC appearing to be a more realistic estimate.

Glottochronological computations suggest that the differentiation of the Western Austronesian languages was well advanced by 1000 BC.

Relationships to other families. Many different proposals have been made to link Austronesian with other language groups—Mon-Khmer, Munda, and Vietnamese of the Austroasiatic language family, Tai-Kadai, Sino-Tibetan, and Indo-European, among others. None has been convincing. Ultimately, no doubt, Austronesian languages, like every other family in Oceania, must derive from ancestral stages spoken in Asia at some remote period. Discovery of such distant connections, however, will have little bearing on the question of where the ancestral Austronesian language itself developed.

Reconstructions of Otto Dempwolff

Reconstruction. Many scholars have worked to reconstruct the Proto-Austronesian sound system and word stock. The reconstructions of Otto Dempwolff, a German ethnologist and linguist, published between 1920 and 1938, have remained the point of departure for all subsequent comparative studies. Dempwolff attributed to Proto-Austronesian a four-vowel system, consisting of a low vowel *a* and three higher vowels—*i* (front), *e* (central), and *u* (back). The reconstructed consonants are the voiced stops *b, d, D, j, g* (a stop is a sound made with complete stoppage of the breath from the lungs); the matching voiceless stops *p, t, T, c, k* and the glottal stop *q;* the nasal consonants *m, n, ñ,* ŋ (nasals are pronounced with the breath going through the nose); the semivowels *w* and *y;* plus *l, r, R, h, s, z,* and *Z*. (Phonetic value of phonemes represented by capital letters, and that of certain other reconstructed symbols, cannot be precisely determined.) Most word bases consisted of two syllables, the commonest shape being consonant-vowel-consonant-vowel-consonant or consonant-vowel-consonant-consonant-vowel-consonant. Clusters of consonants were restricted to a few types and occurred only in the middle of words and possibly in initial position. Words with initial vowels or final vowels or both were probably more common than Dempwolff's reconstructions allow.

Although some languages, particularly certain members of the Oceanic group, have changed this sound system drastically, it is still reflected fairly faithfully by many Western Austronesian languages. Indeed, in both Western Austronesian and Oceanic languages, many words seem to have persisted in almost unchanged form, a condition unparalleled in the Indo-European languages, for example. Table 1 compares the forms of some Proto-Austronesian words with the cognate terms in four modern languages.

Table 1: Some Proto-Austronesian Terms and Their Related Forms in Several Modern Languages

	Proto-Austronesian	Tagalog	Malay	Fijian	Samoan
two	*Duwa	dalawa	dua	rua	lua
four	*e(m)pat	apat	empat	vā	fā
five	*lima	lima	lima	lima	lima
six	*enem	anim	enam	ono	ono
bird	*manuk	manok	manu	manu-manu	manu
eye	*mata	mata	mata	mata	mata
road	*Zalan	daan	jalan	sala	ala
pandanus	*panDan	pandan	pandan	vadra	fala
coconut	*niuR	niyog	nior	niu	niu

*Form that is not actually found in any document or living dialect; it is a reconstructed, hypothetical form.

Systematic reconstruction of Proto-Austronesian grammar has scarcely begun. Structural features retained by a majority of languages in both major branches include a fairly constant form for each grammatical and vocabulary element, with boundaries between elements in words being clearly definable, and a relatively simple morphology of verbs and nouns. In addition, in the verb phrase, a number of elements indicating tense, aspect, and voice are present; they were evidently prefixes and infixes (particles inserted within the body of a word) in Proto-Austronesian. Reduplication, the repetition of a word or a portion thereof, occurs in the case of the verb root and has several functions. Most roots are capable of being used either as nouns or verbs. Adjectives, numerals, and markers indicating negatives can act as verbs. Noun subclasses include personal names, marked by a personal article; common nouns, marked by a common article; locatives, place names, and directionals, marked by a locative particle; and temporals, which are not marked. Personal pronouns include distinct forms for 1st person including the hearer and 1st person excluding the hearer; pronouns marking subject or possessor differ in form from those marking object or focus.

Current research. Since World War II the descriptive and comparative study of Austronesian languages has expanded considerably, with centres at universities in the Philippines, Indonesia, Australia, New Zealand, the United States, and Europe. Reasonably good dictionaries and grammars are now available for most of the better known Western Austronesian languages and many of the Polynesian languages. Several large areas remain poorly known, however, including most of Melanesia and much of Borneo and east Indonesia. A good deal of recent research has concentrated on classification and associated problems in the methodology of subgrouping and the theory of linguistic change. A major work has been a lexicostatistical classification of over 200 Austronesian languages prepared by the American linguist Isidore Dyen (1965). A lexicostatistical classification uses statistics to compare the vocabularies of two or more related languages; the method is similar to glottochronology (see above) but assumes a constant rate of change only within a given language family.

Lexicostatistical classification

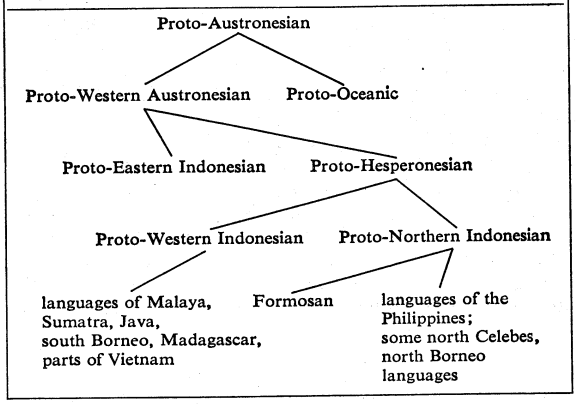
WESTERN AUSTRONESIAN (INDONESIAN)

The Austronesian languages lying west of New Guinea, together with the Chamorro and Palauan tongues of Micronesia, are often called Indonesian. The term is unfortunate in that it does not do justice to the wide geographic distribution of the languages concerned and can be confused with the name of Indonesia's national language. The commonly used alternative, Western Austronesian, is, therefore, employed here.

The classification of Western Austronesian languages is not completely agreed upon. There is, however, fairly general recognition of one very large grouping containing most of the languages of west Indonesia and of Malaysia, all the languages of the Philippines and Madagascar, some languages of the northern Celebes, and the Chamic group of Vietnam and Cambodia. The name Hesperone-

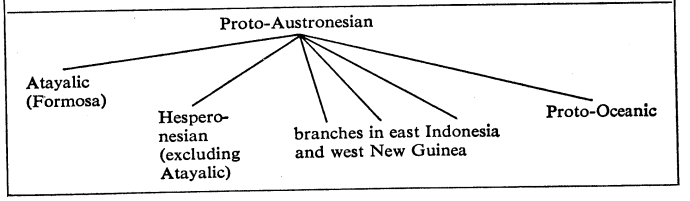
sian has recently achieved some acceptance as a convenient label for this grouping. Disagreement centres around the place of the Austronesian languages of Formosa and east Indonesia, including the western end of New Guinea. Many regard these as most closely related to Hesperonesian, and, in particular, treat the Formosan languages as a branch of Hesperonesian with closest relatives in the Philippines. In this view, Western Austronesian refers to a primary subgroup of Austronesian, contrasting with Oceanic, as indicated in the family tree in Table 2.

Table 2: Austronesian Family Tree (Theory 1)



Other scholars regard the east Indonesian–west New Guinea languages as forming a number of primary subgroups, each coordinate with Hesperonesian and Oceanic. Similarly, a case recently has been made for treating the Atayalic group in Formosa as a primary branch of Austronesian. The Atayalic languages share a very low percentage of basic vocabulary with all other members of the family, including other Formosan languages. There is, then, some support for an alternative family tree, with several primary branchings, representable roughly as in Table 3.

Table 3: Austronesian Family Tree (Theory 2)



Branches of the Hesperonesian group

The Hesperonesian group is generally regarded as dividing into two main branches (see Table 2). One, Western Indonesian, includes the languages of Malaya, Sumatra, Java, Madura, Bali, and Lombok, south Borneo, Madagascar, and probably the Chamic group of Vietnam and Cambodia. The other, Northern Indonesian, encompasses the Philippine languages, some of the languages of north Borneo and the northern Celebes, and possibly the Formosan languages.

Regional division of Western Austronesian languages. In the following sections, the main Western Austronesian languages are discussed in more detail under the appropriate regional headings.

Western Indonesia and Malaysia, Madagascar, Vietnam. The west Indonesian–Malaysian region includes the islands of Sumatra, Java, Madura, Bali, Lombok, and Borneo, and the Malay Peninsula. Except for a few Mon-Khmer languages in the interior of Malaya, its indigenous languages are all Austronesian. There have been important influences on the languages and cultures from mainland Asia, however, and, more recently, from Europe. Hindu culture reached Java and Sumatra in the 1st century AD, and from the 4th century diffused to Borneo and the Celebes. Many languages of western Indonesia contain extensive borrowings from Sanskrit. Islāmic influence began in the 7th century and spread over much of Indonesia and the Philippines. European colonization, beginning in the 16th century, added a third layer of borrowings to the Sanskrit and Arabic.

Malay in its several dialects is the native language of some 6,000,000 people occupying both sides of the Strait of Malacca. Its expansion throughout Malaya is apparently recent; until the 13th century it was confined to east and south Sumatra and the facing coastline of the Malay Peninsula. For some centuries prior to European contact, a simplified form of Malay, Bazaar Malay, had been the lingua franca in coastal regions of the whole Indonesian Archipelago. This status was attained because of the strategic position of Malacca on the trade routes and its function as a dispersal centre of Islām, and also because of the seafaring and trading skills of the Malay speakers. Malay was in turn adopted by the Dutch and British administrations as their lingua franca. In the present century, the language, now the national tongue of the Indonesian state, is called Bahasa Indonesia, or Indonesian. The choice of Malay as a politically acceptable national language was facilitated by the fact that it was the first tongue of only a small minority of Indonesians (in contrast to such languages as Javanese and Sundanese) but was, at the same time, widely used in Indonesia as a second language. The Indonesian variant of Malay has undergone spelling reforms, and a large new vocabulary has been coined to cover modern technical concepts. Many words have also been borrowed from European languages and Javanese.

Bahasa Indonesia

Malay is closely related to most of the other languages of Sumatra, including Minangkabau, Kerintji (Kinchai), Rejang, and Achinese, and to Madurese of Madura Island. Other well-known Sumatran languages are Gajo and Toba-Batak, spoken in the north and somewhat less closely related to Malay. In all, Sumatra contains between 12 and 15 languages (the boundaries between dialects and languages not being clear-cut) spoken by about 9,000,000 people. Engganese, used on a small island off southeast Sumatra, is an extremely divergent Indonesian language that shares only about 10 percent of basic vocabulary with the rest.

Java, the most populous island of Indonesia, is linguistically one of the most uniform, with only three indigenous languages. Javanese, with about 45,000,000 speakers, is numerically the largest Austronesian language after Indonesian (which is known by over 80,000,000 people, but mainly as a second language). Javanese dialects are spoken throughout central and east Java and in sections of western Java. Old Javanese is known from inscriptions dating back to the 9th century AD. The language was extensively influenced by Sanskrit between the 9th and 15th centuries, during the time of the Indianized kingdoms of east Java. Thoroughly documented by Dutch linguists in a series of grammars and dictionaries produced over the last 250 years, Javanese is now generally written in the roman (Latin) alphabet and continues to flourish alongside Indonesian as a literary language and as the daily medium of communication of many Javanese newspapers, magazines, and radio stations. Several registers, or styles, distinguishing degrees of respect and marked by vocabulary differences, have arisen out of the elaborate stratification of Javanese society.

Styles, or registers, of Javanese

Sundanese, with about 13,000,000 speakers, is found throughout west Java, with the exception of small islands of Javanese speakers along the north coast of Banten and in the region of Indramaju. The language has been written in various scripts since the 14th century, roman now being in general use. The dialect spoken around Bandung is considered to be the most prestigious, or the standard form. Madurese is the language of Madura Island and smaller offshore islands, but about half of its 6,000,000 speakers now reside in east Java. There are two principal variants of Madurese—Eastern and Western; an Eastern dialect, Sumenep, has been adopted for educational purposes as the standard dialect. Madurese shows greater similarities to the Malay–Sumatran group than to Javanese or Sundanese. Balinese, with about 2,000,000 speakers, is spoken on Bali and the western part of Lombok; Sasak is the language of eastern Lombok.

Linguistically the most diverse of the islands of the western Indonesian region, Borneo is also the least known. It is sparsely populated, and most coastal areas have been occupied in relatively recent times by speakers of Malay-type languages, such as Iban (Sea Dayak), Brunei Malay, Kutai Malay, Banjarese, and Sambas Malay, all of which are closely related dialects or languages. The most important Bornean language is Ngadju (Ngaju), spoken as the native language in southwest Borneo in the region of the Barito, Kapuas, Kahayan, Kaitingan, and Sampit (Mentaya) rivers; it is more widely used as the lingua franca of most of south Borneo. In northeast Borneo there are a number of languages with close relatives in the Philippines, including Illanum (Lanum), Bajau, Sulu, and the various Murut dialects. Maanyan and closely related languages spoken in south Borneo are apparently the closest relatives of the Malagasy dialects of Madagascar.

The Malagasy dialects

The Malagasy dialects probably derive from Bornean traders who journeyed to Indian, Arabian, and East African ports and settled the previously uninhabited island of Madagascar. A quite close correspondence in basic vocabulary (45 percent agreement) between Maanyan and Malagasy indicates separation around 2,000 years ago. The Malagasy dialects show sufficient internal diversity to justify classification into two or three different languages. The Merina dialect, now official in the Malagasy Republic, has around 5,000,000 speakers. A small group of Austronesian languages spoken in Vietnam and Cambodia—the Chamic languages—probably falls into the Western Indonesian subgroup. Cham, with 150,000 speakers in Vietnam and Cambodia, has been the subject of several studies. Rade, Curu (Chru, Cru), and Jarai, spoken in Vietnam, may be closely related to Cham.

The Philippines and Formosa. At least 70 languages are spoken in the Philippines in a land area a little larger than that of Great Britain. There are two main subgroups. A Central (Mesophilippine) division includes many of the languages of central and southern Luzon, Mindoro, Palawan, and the Visayan Islands. The most important is Tagalog, the first language of about 5,000,000 people in central and southwest Luzon, including the region of the capital city, Manila, and, parts of coastal Mindoro. A standardized form of Tagalog was selected as the national language after the Philippines gained independence in 1946. This language, officially called Pilipino, is now spoken as a first or second language by a considerable proportion of the 38,000,000 Filipinos. Although it faces competition from English and other important local languages, it is increasingly used in education and government and in the popular press, radio, and literature of the Philippines. Closely related to Tagalog are Bikol or Bicol (over 2,000,000 speakers in southeast Luzon and smaller adjacent islands) and Subanon (western Mindanao, southern Philippines). Also in the Central branch are Cebuano, or Sugbuhanon, with 6,000,000 speakers, and Hiligaynon, or Ilongo, with 2,000,000 speakers, the two main members of the Bisayan subgroup. Spoken primarily in Cebu, Bohol, west Leyte, and the eastern third of Negros, Cebuano is used as a trade language throughout Mindanao and Hiligaynon in Panay and on smaller adjacent islands and in Negros. Pampangan (Kapampangan) is another important language, spoken in Luzon on the northwest flank of the Tagalog speech area.

Northern Philippine subgroup

The second major Philippine group, called Northern or Cordilleran, contains most of the languages of northern Luzon. Chief among them is Ilocano (Iloko), the lingua franca in northern Luzon, which has some 3,000,000 speakers in much of northwest and northern Luzon. Other members of this branch include the languages of the Igorot mountain tribes, including Bontoc, Ifugao, and Tinggian, the Gaddang group, Isneg, Isinay, and Kalinga. Pangasinan, spoken by 750,000 in the province of the same name, probably belongs to the Northern subgroup.

A number of Philippine languages fall outside the two large subgroups. Maranao, Tiruray, Bilaan, and Tagabili are languages of important groups in the southern Philippines, while Luzon contains several unclassified tongues—Ilongot, Casiguran, and others. These languages share around 30 percent of basic vocabulary and many structural characteristics with other Philippine languages but are not closely related to any large subgroup. The Sulu Archipelago has Tau Sug and Samal; both are also spoken on Borneo, where Tau Sug is known as Sulu. The major Philippine languages—Tagalog, Cebuano, and Ilocano—have extensive literatures, and together with other important languages, are widely used in broadcasts and newspapers.

Prior to the arrival of the Chinese, Austronesian languages were probably spoken over most of Formosa. The surviving languages are now confined to small communities in less accessible regions. The Formosan languages are internally fairly diverse, but there is some evidence for treating them as a single subgroup of Western Austronesian. The Atayalic or Northern group contains the Atayal dialects, Seedik, and probably Saisiyat. Although some central, eastern, and southern languages share less than 25 percent of basic vocabulary with each other, most of them are generally assigned to a single group (Central or East Formosan). Its members include Ami, Paiwan, Bunun, and Thao. Tsou, Saaroa, and Kanabu (central Formosa), though they may form a third subgroup, are usually treated as members of the Central division.

Two languages of Micronesia, Chamorro (Mariana Islands) and Palauan (Palau, western Carolines), are generally regarded as deriving from the Philippines or Indonesia.

The Celebes, eastern Indonesia, west New Guinea. Eastern Indonesia is centred in the Lesser Sunda Islands and the Moluccas. The large island of the Celebes occupies a position intermediate between Borneo, the Philippines, and eastern Indonesia. Approximately 100 different Austronesian languages are spoken in eastern Indonesia, in places on the coast of west New Guinea, and on Aru Island south of the Doberai (Vogelkop) Peninsula, west New Guinea. Papuan languages (*q.v.*) are spoken in parts of Timor in the Lesser Sundas and Halmahera in the Moluccas.

Austronesian tongues in eastern Indonesia and west New Guinea

Dutch and Indonesian linguists have long recognized that this region contains many disparate Austronesian groups but have generally assigned them to a relatively small number of divisions: Bima–Sumba, Ambon–Timor, Sula–Batjan, south Halmahera–west New Guinea, and several Celebes groups. Very thorough lexicostatistical comparisons provide little support for most of these larger groupings and indicate instead that east Indonesia–west New Guinea harbours many small, genetically diverse subgroups, most of which are members of or have their closest relationships with the Hesperonesian group and with each other, with a residue not at present assignable to any large subgroup of Austronesian.

The Celebes, with more than 40 languages, is the most diverse area. Buginese, with about 2,500,000 speakers, is the most important language of the group. Gorontalo and Suwawa, in the north Celebes, and Sangir (Sangir Islands, north of the Celebes) show closer agreement with Philippine languages than with other Celebes tongues. Other numerically large speech communities in or near the Celebes are the Sidjai, Duri, and Mandar (Andian), Kendari, Muna (Muna Island), and Butung (Butung Island). Sikka of Flores Island and Solor of nearby Solor Island appear to be quite closely related to each other, with a 37 percent common basic vocabulary. Endeh of Flores is more distant. Havunese, spoken on Sawu and Raidjua, between Flores and Timor, and Sumba (Sumba Island) appear to be fairly distant from all other eastern Indonesian languages. A large number of similar isolated languages exists on smaller islands in this region; one such small group includes Ambonese of Ambon Island and the adjacent coastal area of Ceram. Buli and Minyafuin, of Halmahera, may belong to a group that also includes As, spoken on the western tip of New Guinea, and Biga, from Wakde Island in northwest New Guinea. The Bomberai Peninsula group in west New Guinea has close relatives on east Ceram. Most of the Austronesian

languages on the north coast of west New Guinea, as far east as Sarera (Geelvink) Bay, however, show no close relationship to the Hesperonesian, eastern Indonesian, or Oceanic languages.

Great internal diversity of western Austronesian

Characteristics of Western Austronesian. The Western Austronesian languages are so diverse internally that little can be said about their common characteristics apart from what has already been said of Proto-Austronesian. Most of the members of this division maintain distinctions in the Proto-Austronesian sound system that are lost in Oceanic; *e.g.*, the distinction between **b* and**p*. (An asterisk indicates an unattested, hypothetical, reconstructed form.) The original verb morphology is also generally retained more completely than in the Oceanic division, but it is difficult to isolate common innovations that mark off Western Austronesian languages in the absence of detailed reconstructions of Proto-Austronesian grammar.

The Philippine languages, typologically the most uniform of the large subgroups of Western Austronesian, have been the best described. No doubt owing to their long coexistence in a single region, as well as to their relatively close relationship, members of this subgroup of Hesperonesian are more alike in their sound systems (though not in vocabulary) than the languages of the West Indonesian subgroup. Consonants usually include three voiceless stops, *p*, *t*, *k*, and the glottal stop *q*; three voiced stops, *b*, *d*, and *g*, and the nasals *m*, *n*, and ŋ; plus *l*, *r*, *h*, *s*, *w*, and *y*. In many languages, borrowings from Spanish have resulted in the addition of two new vowels, *e* and *o*, to the original four (*a*, *i*, *u*, and ɨ or ə) retained from Proto-Austronesian. The morphology of the verb and, to a lesser extent, of the noun is fairly complex. The use of affixes and reduplication distinguishes several tenses (past, future, and present or general), various aspects (progressive, distributive, causative, etc.), and two modes (obligatory and indicative).

Philippine voice and case systems

The grammar of Philippine languages is most famous for its complex voice and case systems. (Voice, in grammar, indicates the relationship of the subject of the verb to the action that the verb expresses. A grammatical case is a form of a word that shows the relationship of the word to other words.) Besides distinguishing so-called active (*e.g.*, "John washed the dishes") and passive (*e.g.*, "The dishes were washed by John") constructions, Philippine languages possess at least three kinds of passive. The noun phrase occurring as "topic" or "focus" in the sentence can represent the actor (subjective or active constructions), the goal or object (goal or object-focus passives), the referent (benefactive or location-focus passives), or the instrument (instrumental-focus passives). The distinctions are expressed by three devices. One uses affixes to mark voice in the verb: subjective voice is usually indicated by an infix *-um-* and the prefixes *ma-*, *mag-*, *maka-* (the actual choice varies with the different verb classes); objective voice is shown by the suffix *-in* or *-en*; referential voice by *-an*; and instrumental voice by the prefix *i-*. The other methods of distinguishing passives include the marking of case with nominal (noun) particles and transposing phrases and changing stress. An example from Tagalog is: "The child bought the mango," *b-um-ili, ang bata, n-ang mangga*; *bili* is the verb "buy" with the subjective infix *-um-*, *bata* is "child," and *mangga* is "mango"; *ang* marks the focus common-noun phrase ("the child"), and *n-ang* indicates the nonfocus common-noun phrase ("the mango"). To say "The mango was bought by the child," the sentence construction is rearranged to *b-in-ili nang bata ang mangga*, which in actual Tagalog order is "buy-objective voice, the child, the mango." A number of other arrangements of the sentence elements are possible to cover the instrumental and referential voices.

Members of the Western Indonesian and Formosan groups share a great deal of morphological and syntactic structure with the Philippine languages, although they are typologically less homogeneous. Most of the Indonesian languages have developed more elaborate vowel systems than that of Proto-Austronesian, with five (Havunese), six (Malay, Balinese, Buginese), seven (Sundanese), eight (Javanese, in one analysis), or nine (Cham) contrasting vowels. Maanyan and the Malagasy dialects retain the four original vowels. Most Western Indonesian languages preserve the distinction between four voiced and four voiceless stops (the voiced *b*, *d*, *j*, *g*, and the voiceless *p*, *t*, *c*, *k*) and four nasals (*m*, *n*, *ñ*, ŋ), plus *l*, *r*, *w*, *y*, *s*, and *h*. Only Javanese, however, also distinguishes retroflex stops and alveolar stops (made with the tip of the tongue touching the ridge behind the upper teeth). This distinction seems likely to have arisen from the later influence of Sanskrit, which possessed retroflex consonants, on Javanese. As in all branches of Austronesian, most word bases may be used in the function of nouns or verbs. There are numerous transformative and formative affixes in Western Indonesian that derive adverbs, adjectives, nouns, comparative forms, ordinal forms, and various verb constructions, in addition to several kinds of verb reduplication; these features also appear in the Philippine languages.

Possible influence of Sanskrit on Javanese

The large Hesperonesian group, in general, appears to be the most conservative branch of Austronesian. A high proportion of the 2,207 words attributed to Proto-Austronesian, for example, are retained by Tagalog (1,125), Toba–Batak (1,299), Javanese (1,446), Malay (1,627), and Ngaju (1,170). Much lower proportions are present in Oceanic languages (Bauan Fijian 461, Tongan 328, Samoan 385).

Eastern Indonesian languages in general appear to be morphologically simpler than members of the Hesperonesian group, in which regard they resemble Oceanic languages. They have, however, retained several distinctions in sound that have been lost in the Oceanic group.

Although books and articles on Western Austronesian languages run into several thousands, documentation of the speech traditions of many of the smaller, minority communities is still poor, and many gaps remain in the understanding of the history of the group as a whole. Indonesian and Filipino linguists are now working extensively on their own languages, while several Dutch and American universities, and the Summer Institute of Linguistics in the Philippines (an association of Protestant missionary linguists that specializes in studying unrecorded languages) are active in both descriptive and comparative studies.

European languages contain a number of words borrowed from the better known languages of the Indonesian region. From Malay derive such English words as sarong, (to run) amuck, mandarin (through Portugese), the kris (or creese) dagger, and the names of several animals (orang-utan, pangolin, dugong, kalong) and plants or plant products (kapok, paddy, nipa). Javanese contributions include junk (sailing vessel) and batik. The American English word boondocks is from Tagalog *bundok* "mountain."

EASTERN AUSTRONESIAN (OCEANIC)

Roughly 300 Austronesian languages are spoken east of Sarera (Geelvink) Bay in New Guinea and on the islands of Melanesia, Micronesia, and Polynesia.

Classification. The classification of these languages remains somewhat controversial. Dempwolff found evidence to indicate that all known Austronesian languages in the Oceanic region, apart from Chamorro and Palauan, belong to a subgroup that excludes Western Austronesian languages. The names Oceanic and, less commonly, Eastern Austronesian are now applied to this grouping. The evidence consists of a number of common innovations in the treatment of Proto-Austronesian sounds. All Oceanic languages agree in having lost the original final consonants of most word bases in most contexts, and in simplifying the initial and medial consonants as shown in Table 4. In addition, Oceanic languages reflect a five-vowel system in which Proto-Austronesian *a*, *i*, *u* were retained, *ay* became *e*, and *e* and *aw* became *o* in Proto-Oceanic.

Though the evidence cited above is not challenged, the conclusions drawn from it have been. Scholars remain puzzled by the great differences among Oceanic languages, especially in vocabulary items. Comparisons of

Table 4: Correspondences Between Proto-Austronesian (PAN) and Proto-Oceanic (POC) Consonants

PAN	p	b	mp	mb	t	nt	d	D	nd	nD	l	r		(n)s (n)z	(n)j (n)Z
POC	p		mp		t	nt	d		nd		l	r		s,z	
PAN	k	g	ŋk	ng	m		n	ñ	ŋ	w	q	R	h		y
POC	k		ŋk		m		n		n	w	q	R	Ø (zero)		y

Puzzling differences among Oceanic languages

basic vocabulary show that several New Guinea–West Melanesian groups share less than 15 percent of a common basic vocabulary with all other members of Austronesian, indicating a divergence from the parent language at least 5,000 years ago. It has been suggested that the Oceanic languages may not constitute a single subgroup but may divide into several primary divisions within the Austronesian family, and indeed that West Melanesia may be the primary dispersal centre for Austronesian. A quite different theory, with a few modern adherents, explains the diversity among the languages of Melanesia by deriving them from relatively recent mixtures of Indonesian and Papuan languages, which results in disparate varieties of pidginized Indonesian with Papuan substrata in each island group.

There are increasing grounds for accepting Dempwolff's theory of the Western Austronesian–Oceanic division as correct, while still allowing for a more dramatic time depth. It is also evident that, after a period of development as a single language, most of the descendants of Proto-Oceanic that dispersed in the New Guinea–West Melanesian area were influenced by nearby Papuan languages; this, together with other factors, such as the small size of the speech communities, word taboos, and several millenia of separate development, probably account for the extreme diversity found in this part of Oceania. It is not clear whether the New Guinea–West Melanesia region was the primary dispersal centre for Austronesian languages, but it is certain that it was a very early one. From there Oceanic-speaking groups moved fairly rapidly into the unoccupied islands of East Melanesia, Micronesia, and Polynesia. As has been noted, both Fiji and Tonga were inhabited by 1000 BC, almost certainly by speakers of the branch of Oceanic ancestral to Fijian and Polynesian, while the Solomons, New Hebrides–Banks Islands, and New Caledonia were probably settled by 2000 BC or earlier. In New Guinea and the western Melanesian islands, Oceanic-speaking peoples encountered Papuan populations and in many places intermarried with them and borrowed from their languages. One result of this contact is that many Austronesian languages in that region have changed faster than those situated further east.

Austronesian encounter with Papuan languages

Characteristics of Proto-Oceanic. Among the additional evidence for Dempwolff's hypothesis that has come to light in recent years is the sharing by the Oceanic languages of many words and grammatical features that are not characteristic of Western Austronesian languages. More than 30 percent of Proto-Oceanic words from a standard list of 215 "basic vocabulary" meanings are not represented at all in Western Austronesian. Oceanic languages also agree in having peculiar forms for certain common Austronesian words; *e.g.*, the Proto-Oceanic forms are **au* "I," **mai* "come," **suRi* "bone horn," **pati* "four," **moze* "sleep," **katoluR* "egg," where Western Austronesian languages reflect **aku*, **maRi*, *duRi or **DuRi*, **e(m)pat*, **peZem*, **teluR*. In grammar, a number of common simplifications and elaborations are observable. The general features of the following discussion of Proto-Oceanic grammar apply generally to present-day Oceanic languages, although some languages retain the parent system more completely than others. Most languages of the central and eastern Solomons, some languages of the central and northern New Hebrides and Banks Island, the Fijian dialects, the Polynesian group, and certain languages in New Guinea–West Melanesia appear to have preserved the Proto-Oceanic system more completely than others.

Some morphological simplifications occurred in Proto-Oceanic: many verb and noun affixes were lost, some prefixes lost their prefix quality and were assimilated as nonproductive elements into the word base, and some affixes normally appended to a word became reinterpreted as free-form particles. On the other hand, Proto-Oceanic developed several new prefixes and changed the shape or function of certain old ones.

Whereas tense, aspect, mode, and probably person and number of the verb were marked in Proto-Western Austronesian by affixes (*e.g.*, as in English "walk," "walked"), in Proto-Oceanic they were indicated by separate particles. Many verb bases also functioned as postverbal particles indicating direction or tense-aspect—*e.g.*, Proto-Oceanic **zake* "ascend," **zipo* "descend," **mai* "come," **nopo* "stay" also occurred as modifiers meaning "upward," "downward," "toward," and "progressive aspect," respectively.

Nominal (noun) phrases were of at least four types. Common nouns were marked by the "common article" **na*, personal pronouns and possibly personal names were indicated by personal articles (**i*, probably **a* or zero—unmarked—in certain contexts), locative nouns were shown by **(q)i* immediately preceding the base word, and temporal nouns were marked by a temporal prefix or zero (the lack of a prefix). One subclass of locatives, called relationals, used personal suffixes, as **(q)i lalo-na* "at inside-his" (*i.e.*, "inside him").

Pronouns in Oceanic

As in Proto-Austronesian, personal pronouns fell into two sets of variants: subjective and possessive forms, and focus and objective forms. Examples from the Wayan dialect of Fijian that preserve the Proto-Oceanic forms are: "I go," *qu laka* (*qu* is the subjective pronoun "I," *laka* is "go"); "my arm," *qu-lima* (*qu* is the possessive pronoun "my," *lima* is "arm"); "It is I," *o au* (*o* is the personal article, *au* is the focal or emphatic pronoun "I"); "give (it to) me," *vaganī au* (*vaganī* is "give," *au* is the objective pronoun "me"). Several new distinctions appeared in Proto-Oceanic, however. One was the development of dual (two-person) and trial (three-or-more-person) suffixes. The forms **ru(a)* "two" and **tolu* "three" were suffixed when two or a few people, rather than an unlimited plurality, were referred to. Plurals were indicated by the simple nonsingular base, as Fijian *keda* "we plural (including hearer)" contrasting with *keda-ru* "we two (including hearer)" and *keda-tou* "we three (including hearer)."

It was in possessive constructions that Proto-Oceanic underwent some of the most distinctive elaborations. Common nouns in such structures fell into at least three "genders": edible (foods, drinks, etc.), familiar or inalienable (kinship terms, parts of a whole), and neutral. The possessor was linked to the possessed noun by a particle of variable form. When the possessed noun was of edible gender, the consonant of the possessive particle was **k-*; in instances of neutral gender it was **n-*; and when of familiar gender, the consonant was omitted. The vowel or vowels of the possessive particle varied according to whether the possessor was a personal pronoun, a common noun, or a proper noun. Similarly, word order depended on the class of the possessed and possessor nouns: the particle plus possessor normally followed the head noun, as in Fijian *na uvi kei Manu* "Manu's yam," *na vosa nei Manu* "Manu's speech," *na tina i Manu* "Manu's mother," but, when the possessor was a pronoun and the possessed noun was of other than familiar gender, then the particle and pronoun preceded, as Fijian *na ke-mu uvi* "your yam," *na no-mu vosa* "your speech" (but *na tina-mu* "your mother," *na mata-mu* "your eye").

Proto-Oceanic interrogative pronouns included several retained from Proto-Austronesian: **zapa* "what?", **zai* "who?", **piza* "how much? how many?", and **kuya* "how? in what state?" together with **pai* "where?", and **ŋ(a)iza* "when?". Demonstratives distinguished three positions or persons: **[i,e]ni* "this, here, near me, now," **ina* "that, there (by hearer), then," and a form marking remote position in space and time.

Oceanic words functioning as nouns and verbs

As in Proto-Austronesian, most word bases, other than proper nouns, functioned as both verbs and nouns, and adjectives, numerals, and negatives played the role of verbs in sentences.

Subgroups of Oceanic languages. Many aspects of the subgrouping of Oceanic languages are still obscure. In the west, many small divisions are evident, but no large subgroup has been attested. In the east, there is some evidence for a wider subgroup—Eastern Oceanic—comprising the languages of the southeast Solomons, much of the New Hebrides–Banks Islands, Rotuma, Fiji, and Polynesia, and possibly also most of the Micronesian languages. Within Eastern Oceanic, there is fairly clear support for a subgroup consisting of Fijian and Polynesian, whose closest immediate relatives may lie in the central New Hebrides. Table 5 presents a classification that has some degree of general acceptance.

LANGUAGES OF MELANESIA

At a conservative estimate, 250 different Austronesian languages (not counting dialectal variants) are found in Melanesia. Most of them are spoken by communities of a few hundred to a few thousand people; with several exceptions they remain documented only by brief word lists and sketches. The term Melanesian is often applied to the Austronesian languages of Melanesia but has validity only as a geographical label, for these languages are not a linguistic unity comparable to the Polynesian and Nuclear Micronesian groups. Rather, Polynesian and Micronesian are each branches of particular Oceanic subgroups whose other members are in Melanesia (see Table 5).

Papuan and Austronesian languages in New Guinea

In New Guinea, Austronesian languages occur on the Doberai (Vogelkop) Peninsula, extending southwest as far as Kaimana, and in patches along the north coast, and along the southeast coast of Papua as far west as Cape Possession (100 miles west of Port Moresby). Only in the Markham Valley and parts of the Central District of Papua are Austronesian languages found any distance inland. The interior and large stretches of the coast of New Guinea are occupied by Papuan languages. Apart from those tongues situated around Sarera (Geelvink) Bay and on the Doberai (Vogelkop) Peninsula, all the New Guinea Austronesian languages appear to belong to the Oceanic division. The best known of them is Motu, spoken by about 10,000 people who occupy the coastal strip around Port Moresby. Police Motu, a simplified form of Motu, is used as a lingua franca throughout Papua. Motu is quite closely related to the eight other languages of Papua's Central District. This group, in turn, appears to form a larger grouping with many of the languages spoken around the southeast tip of Papua and the offshore islands, including Suau (Suau Island and adjacent coastal area) and Dobu (Dobu and Normanby Islands). These two languages, together with Kiriwina from the Trobriand Islands and Wedau spoken around Dogura, have some currency as lingua francas in their respective regions of the Milne Bay District.

Austronesian languages are found in coastal pockets in the Northern District of Papua, in the Morobe, Madang, and Sepik Districts of New Guinea, and on many offshore islands in these regions. Most of them have been strongly influenced by Papuan languages; *e.g.*, Yabêm and several closely related languages in the Morobe District have developed tone, like their Papuan neighbours. Yabêm and Graged (Gedaged) of Kranket Island, in the Madang District, have gained wider currency as the lingua francas of the Lutheran Mission in their regions. Some of the Morobe and Madang languages appear to fall into a subgroup with Kove (Kombe) and certain other languages of southwest New Britain.

Austronesian tongues of New Britain

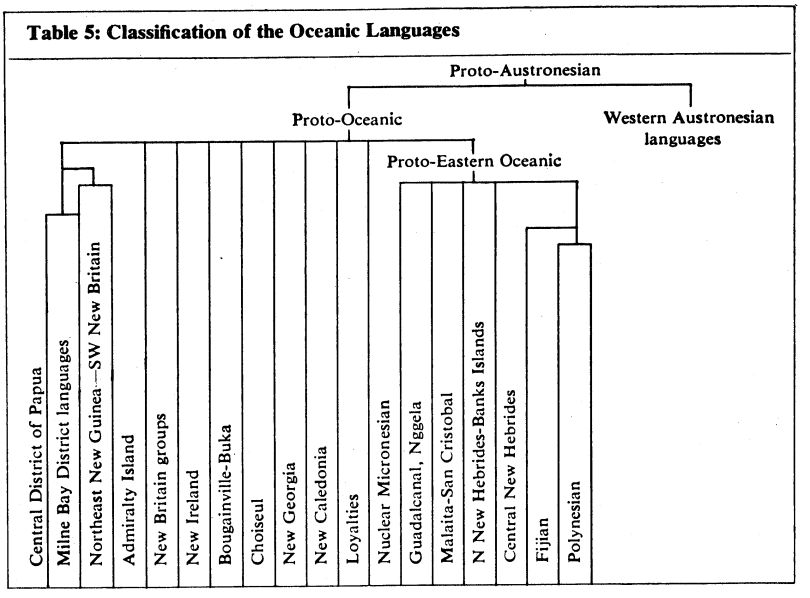
The large island of New Britain contains around 25 extremely diverse Austronesian languages, the most important being Tolai (Kuanua, Tuna, Raluana), spoken around Rabaul on the Gazelle Peninsula and used extensively by missions in New Britain and New Ireland. New Ireland's languages are poorly known but appear to be of only moderate diversity. On the small Admiralty Islands group at least 20 (in some estimates as many as 50) different languages are found.

The Solomon Islands contain more than 60 languages that belong to several divergent groups. Bougainville is predominantly Papuan speaking, but several Austronesian languages are present on the north coast, and a few in the south and on offshore islands. The adjoining island of Buka contains several more, most of them closely related to the Bougainville languages.

On Choiseul Island are found an indeterminate number of languages and dialects closely related to each other but not to any outside group. Babatana is the literary language of a Methodist Mission there. Santa Isabel (Ysabel Island) is sparsely peopled by communities speaking perhaps 10 languages; Kia in the northwest and Bugotu in the southeast are the best known. Bugotu, the lingua franca of the Melanesian mission on Santa Isabel, is closely related to the Guadalcanal and Nggela languages. The New Georgia group contains more than a dozen languages—*e.g.*, Roviana (southwest New Georgia) and Marovo (east New Georgia and Vangunu Island)—most of which belong to a single division. Roviana is used by the Methodist Mission everywhere in the British Solomons except on Vella Lavella. The languages of the southeast Solomons divide into two subgroups. Almost all those on Malaita and San Cristóbal belong to a single group whose best documented members are Sa'a, Kwara'ae (Fiu), Lau (all of Malaita), and Arosi (of San Cristóbal). The other group includes most of the Guadalcanal languages, together with Bugotu and Nggela, of Florida Island.

Polynesian languages spoken on several islands near the main Solomons group include: Nukuria, Nukumanu, Taku, Luangiua (Ontong Java), and Sikaiana to the

Table 5: Classification of the Oceanic Languages



north; Rennell–Bellona to the south; and Tikopia, Anuta, and Pileni–Taumako (Duff Islands) to the east. The latter is spoken in the Santa Cruz group, which also contains one Papuan and several little-known Austronesian languages.

Languages of the New Hebrides–Banks–Torres archipelago

Approximately 60 different languages, many of them dialectally very diverse, are spoken in the New Hebrides–Banks–Torres archipelago. The best known are Mota (of Mota or Sugarloaf Island, in the Banks group), used in the 19th century throughout the northern islands and in the Solomons as the lingua franca and literary language of the Melanesian Mission, and the Nguna–Tongoa dialects, spoken on north Efate and adjacent islands in the central New Hebrides. Many of the central and northern languages are closely related to Fijian and Polynesian. Further south, Eromanga, Tana, and Aneityum harbour three languages that show no very close relationship to other members of the Oceanic group. Polynesian languages are spoken on Emae, Futuna, and Aniwa and on Mele and Vila islands in the New Hebrides.

Both New Caledonia and the Loyalty Islands are internally fairly diverse. New Caledonian languages appear to form a subgroup characterized by far-reaching sound changes. Languages of the central and far southern regions of the island are tonal. A Polynesian language, West Uvean, is spoken on Heo Island in the Loyalties.

Easily the best described, and politically the most important, language in Melanesia is Fijian, or Bauan Fijian, spoken as a first or second language by over 200,000 Fijians and also by other ethnic groups in the Fiji group. The Fijian dialects are sharply differentiated into Western and Eastern Divisions. Bauan, an Eastern dialect, became Fiji's lingua franca in the 19th century and is now widely used in the popular press, broadcasting, and administration. Rotuman, spoken on Rotuma Island, 200 miles north of Fiji, has borrowed heavily from Polynesian, but its classification as a language closely related to Fijian and Polynesian has been disputed.

LANGUAGES OF POLYNESIA

A triangle-shaped area, the apexes of which are Hawaii, Easter Island, and New Zealand, embraces almost all the islands of the central and eastern Pacific. Early European explorers found the inhabited islands in this region occupied by culturally homogeneous peoples speaking some 16 closely related languages. The name Polynesian was first applied to them by 19th-century scholars, who also observed that another 12 to 14 languages belonging to the Polynesian group are situated to the west of the Polynesian Triangle, on islands in Melanesia, and on the southern fringes of Micronesia; these are the Polynesian "Outliers."

Origins of the Polynesians

The question of Polynesian origins has been the subject of much romantic speculation. Various writers have pictured the ancestral Polynesians as migrating from such far-flung homelands as India, Egypt, Mesopotamia, and the Americas. It is now quite clear that, linguistically at least, the Polynesians developed their distinctive characteristics within Polynesia itself. Their languages form a subgroup of the Oceanic branch of Austronesian, whose immediate relatives lie in eastern Melanesia (see Table 5). Fijian, Polynesian, and certain languages of the New Hebrides were a single language until approximately 1000 BC, when some speakers of this early eastern Oceanic language appear to have moved from the New Hebrides into Fiji. There a dialect developed that subsequently split into Fijian and Polynesian branches. During several centuries of isolation somewhere in west Polynesia, probably in Tonga, the Polynesian branch underwent those further modifications that characterize all present-day Polynesian languages.

Characteristics of Proto-Polynesian. The general features of the following sketch of Proto-Polynesian are applicable to all its descendants. Each has changed some of the details for the Polynesian languages show an internal diversity comparable to the Romance or Germanic subfamilies of the Indo-European language family.

Phonology. Proto-Polynesian retained the five-vowel system of Proto-Oceanic but added a contrast between long, or geminate (double), vowels and short vowels. Several simplifications occurred in the consonant system, with the loss of the nasal element in stops (Proto-Oceanic *p* and *mp* both became *p*; *t* and *nt* became *t*; *k* and *ŋk* became *k*) and the loss of retroflex *R*. Proto-Polynesian consonants thus comprised four stops, *p, t, k,* and ʔ (the glottal stop); three nasal sounds, *m, n, ŋ*; three fricatives, *f, s, h*; plus *l, r,* and *w*. No consonant clusters or final consonants were permitted, but vowel sequences were common. Most word bases had two syllables, but longer forms were not uncommon, while some prebasic particles were monosyllabic—*e.g.*, **fale* "house," **waka* "canoe," **wai* "water," **uaua* "vein," **fafine* "woman," **tuakana* "older sibling of same sex," **maanifinifi* "thin," **ki* "to," **ma* "and," and **o* "of."

Grammar. Proto-Polynesian retained most features of the Proto-Oceanic grammar, the main changes being in possessive constructions and in the development of distinctive passive and nominal (*i.e.*, verbless) sentence structures. The smallest natural phonological or pause unit was the phrase rather than the word, a phrase consisting of a head word flanked by affixes and particles. The verb phrase contained several classes of preposed particles: (1) tense-aspect particles (marking past, prospective, subjunctive, imperative, negative subjunctive, hypothetical, nonpast, and progressive); (2) subject-marking personal pronouns occurring between tense-aspect particle and verb; (3) manner particles, indicating the "manner" of the action; and (4) negatives. Verbal prefixes included **fe-* reciprocal, **faka-* causative, **ma-* abilitative (*i.e.*, "able to," "capable of"), **tua-* ordinal, **taki-* distributive, **toko-* human number. Verbal suffixes included a passive **-(C)ia*, instrumental **-aki*, and derivative or transformative **-(C)i*, **-(C)aki*.

Significance of the phrase in Polynesian

Noun phrases divided into common, personal, locative and temporal, each class of noun being marked by a distinctive article or other syntactic marker. Proto-Polynesian probably distinguished only two common articles, a definite **(t)e* and an indefinite **ha*, but several of its descendants have more elaborate series. Most interrogative pronouns were retained from Proto-Oceanic, but manner and temporal interrogatives were innovations. In personal pronouns the distinction between trial (three) and plural was lost, the old trial forms becoming the new plurals. Demonstratives included **eni* "this, these, here," **ena* "that, those, there (near hearer)," and **ia* "that, those (particular ones)."

In possessive constructions the three-way gender distinction of Proto-Oceanic was replaced by a two-way distinction between nouns subordinate to the possessor and nouns not subordinate to the possessor.

Case relations between noun phrases and the verb were indicated by particles placed at the beginning of the phrase. Active sentence structures normally consisted of verb phrase plus subject noun phrase plus complement, and contrasted with passive, nominal, and topicalized structures.

Samoan, Tahitian, and Hawai'ian

Languages and subgroups. The best known Polynesian languages are Samoan, spoken by about 200,000 people in Western and American Samoa and by sizable migrant communities in New Zealand and the United States; Tongan, with about 80,000 speakers; Tahitian, spoken as the native language in the Society Islands and as a second language by Marquesans, Tuamotuans, Magarevans, Austral Islanders (Îles Tubuai), and other ethnic groups in French Polynesia; New Zealand Maori, with about 100,000 speakers; and Hawai'ian, once the language of perhaps 100,000 people, but now used as a daily medium of communication by only a few hundred Hawai'ians. Samoan and Tongan are the national tongues of Western Samoa and the Kingdom of Tonga. Hawai'ian, Maori, and Tahitian, together with the languages of the Marquesas, Tubuai (Austral), Tuamotuan, and Cook island groups, and Magareva and Easter Island, form a well-marked subgroup known as Eastern Polynesian. Samoan with Futunan, Uvean, Tokelauan, Ellice, Pukapukan (in the northern Cooks), and all the Outlier languages spoken in Melanesia and Micronesia form a second division known as the Samoic-Outlier group. A third subgroup

consists of Tongan and Niuean. Eastern Polynesian and Samoic-Outlier languages possess certain innovations that suggest they shared a period of common development after separating from the language ancestral to Tongan and Niuean. This wider grouping is known as Nuclear Polynesian.

The numerous subgroupings and language divisions in Polynesia are the result of the geography of this region. Around 2,000 years ago, following the breakup of Proto-Polynesian and the build-up of populations on the main islands in the west Polynesia area, there was a fairly rapid settlement of all the main uninhabited islands to the immediate east and west. Distances between most of the island groups were such that regular contact between them was impossible, and a distinct language evolved for each group. The last islands to be settled appear to have been the marginal east Polynesian islands, such as Hawaii, New Zealand, Mangareva (all settled about AD 1000), and some of the Outliers. Surprisingly, Easter Island was one of the first eastern islands to be settled, a fact attested by archaeological remains and by the considerable differences between the Easter Island language and those of other members of the Eastern Polynesian group.

The oral literature of Polynesia

The oral literature of Polynesia is extremely rich. Much of it has been recorded by Polynesian and European scholars and by missionaries, who in the 19th century developed excellent orthographies—using the roman alphabet—for Polynesian languages. In the precontact period, Easter Island was the only community to have developed a script of its own; it was an ideographic form of writing with very restricted uses. Missionary scholars also produced many excellent grammars and dictionaries, and intensive linguistic research continues, chiefly at universities in Hawaii and New Zealand.

European languages have borrowed many words from Polynesian. Among the English borrowings are taboo (Polynesian *tapu*), tattoo (Tahitian *tatau* "to mark"), mana, ukulele, hula, Kanaka (South Sea Islander), tapa cloth, kava (drink and plant), and the names of many plants and animals native to Pacific regions; *e.g.*, the New Zealand birds kiwi, moa, tui, huia, kaka, and takahe, the tuatara lizard, the kauri family of trees, and many others.

LANGUAGES OF MICRONESIA

The widely dispersed groups of small islands comprising Micronesia lie to the north of Melanesia, between the Philippines and the Polynesian Triangle. Micronesia has received linguistic infusions from each of the surrounding regions. Of the 13 or so languages native to Micronesia, two are intruders from the Philippines–Indonesia region, and two are Polynesian. The rest have probably been there longer but almost certainly derive ultimately from some part of Melanesia. Most or all of this last group fall into a single subgroup of Oceanic that has been called Nuclear Micronesian.

The two Indonesian-type languages are Chamorro, spoken in the Mariana Islands, and Palauan, spoken on Palau in the western Carolines. Chamorro, with about 50,000 speakers, most of them living on Guam, shows close resemblances to Philippine languages. The Palauan-speaking community numbers about 12,000. Palauan is Hesperonesian, but its immediate affiliations are uncertain. The two Polynesian languages in Micronesia are Nukuoro and Kapingamarangi; they are spoken on atolls lying south of Truk and Ponape by about 400 and 700 people respectively.

Nuclear Micronesian languages

Nuclear Micronesian languages include Marshallese (Marshall Islands, 19,000 speakers), Gilbertese (Gilbert Islands, 44,000 speakers), Trukese (Hall Islands, Truk, Mortlock Islands, 26,000 speakers), Ponapean (15,000 speakers on Ponape, Ngatik, Mokil, and Pingelap), Kusaiean (Kusaie Island, 3,600 speakers), Carolinean (4,000 speakers, with a migrant community on Saipan in the Marianas and a source community in the eastern Carolines speaking a number of fairly diversified dialects), and Ulithian (originally applied to the dialect of Ulithi Island, but now used for the language or dialect continuum that extends from Tobi and Sonsoral through Ulithi and Woleai to Lamotrek, in the western Carolines). In addition, there are two languages that are possibly Nuclear Micronesian but their divergent structure and vocabulary make their membership doubtful at present. Yapese, spoken on Yap Island by more than 4,000 speakers, is one of them, and Nauruan, with about 3,000 speakers on Nauru Island, is the other.

Nuclear Micronesian languages show a variability in vocabulary and structure exceeding that of the Polynesian group, thus indicating that their immediate common ancestor began to divide into several languages at least 3,000 years ago. Gilbertese and Marshallese, for instance, share only about 21 percent of related words in their basic vocabularies, while Kusaiean and its neighbouring languages, Ponapean and Marshallese, share only about 26 percent and 24 percent, respectively. The figures for Yapese and Nauruan and other languages are still lower.

The greatest diversity lies in east Micronesia, suggesting that the homeland of Proto-Nuclear Micronesian may have been in that region. Further west, the languages are more alike; a continuum of mutually intelligible dialects or very closely related languages extends from Truk through the main body of the Caroline Islands southwest to Tobi Island, and includes the Trukese, Carolinean, and Ulithian languages.

The following features appear to be characteristic of Nuclear Micronesian languages (with the partial exception of Gilbertese): velarized consonants (made with movement of the tongue toward the soft palate), long or geminate consonants, the assimilation of vowel pronunciations to neighbouring sounds, and a number of other fairly complex phenomena that modify the pronunciation of particular sounds in particular contexts. Grammatical features include verb phrases introduced by subject prefixes marking person and number, attachable to the verb or to one of the many preverbal tense-aspect particles; many subclasses of nouns distinguished in numerical and possessive constructions; elaborate sets of demonstrative pronouns paralleling many of the meaning distinctions found in personal pronouns; complex numeral and classifier systems; and several kinds of reduplication of base words carrying several grammatical–semantic functions.

In general, the Nuclear Micronesian languages have been among the most innovative of the Oceanic languages. They have considerably modified the original Oceanic sound system and lost a number of grammatical distinctions made in Proto-Oceanic, while elaborating others and developing several new features. Grammatical sketches exist for most of the languages, but as yet few dictionaries have been published. Micronesian communities are almost 100 percent literate. Because of the phonological complexity of many of the languages, however, few of the orthographic systems, all of which are in roman script, are as uniform or satisfactory as those for the Polynesian or Indonesian languages.

BIBLIOGRAPHY. The classic comparative work is OTTO DEMPWOLFF, *Vergleichende Lautlehre des austronesischen Wortschatzes*, 3 vol. (1934–37), in German, which followed a series of important earlier essays by RENWARD BRANDSTETTER, some of which appear in *An Introduction to Indonesian Linguistics*, trans. by C.O. BLAGDEN, (1916). ISIDORE DYEN, *The Proto-Malayo-Polynesian Laryngeals* (1953), modifies some of Dempwolff's reconstructions, while the same author's *A Lexicostatistical Classification of the Austronesian Languages* (1963), presents a subgrouping of 214 languages. G.W. GRACE, *The Position of the Polynesian Languages Within the Austronesian (Malayo-Polynesian) Language Family* (1959), also contains an excellent discussion of previous comparative work. R.H. CODRINGTON, *The Melanesian Languages* (1885), and S.H. RAY, *A Comparative Study of the Melanesian Island Languages* (1926), are now important chiefly for the many grammatical sketches they contain. A. CAPELL, "Oceanic Linguistics Today," *Current Anthropology* 3:371–396, 422–428 (1962), reviews comparative work up to 1960. *A Pacific Bibliography*, 2nd ed. by C.R.H. TAYLOR (1965), and H.R. KLIENEBERGER (comp.), *Bibliography of Oceanic Linguistics* (1957), contain excellent bibliographies of earlier works. More-up-to-date information may be found in *Oceanic Linguistics* (semi-annual), chief among the several journals that

publish work on Austronesian; in T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 8, *Oceania* (1971); and in C.F. and F.M. VOEGELIN, *Languages of the World: Indo-Pacific Fascicles* (1964–).

(A.K.Pa.)

Automata

Usually designed to be decorative or amusing or both, automata are mechanical objects that are relatively self-operating once they have been set in motion. They often imitate natural phenomena, both in form and in movement. An automaton may be large enough to be watched by a concourse of people or small enough to be carried in the pocket. Although some automata, such as robots, are functional and utilitarian, this article is concerned only with automata as art objects, that is, with those that delight the aesthetic sense and fancy.

Aesthetic considerations. Automata are designed to arouse interest through their visual appeal and then to inspire surprise and awe through the apparent magic of their seemingly spontaneous movement. In ancient and medieval times, this ability to stimulate wonder was used in magico-religious ceremonies. Most automata, however, were made either for ornamentation of functional objects and architecture or as gifts and items of personal whimsy. As decorative objects their appeal lay not only in their realism but also in the beauty of their precision craftsmanship and technology. The more ingenious and involved the action of their concealed power source and mechanisms, the greater was their ability to intrigue the spectator. Elaborate movements demand thought and skill on the part of both designer and craftsman. Design limitations are imposed by the materials available and the level of technological knowledge. Complex actions involve many precisely moving parts, some of delicate construction; by nature, therefore, automata are generally fragile and require frequent repair.

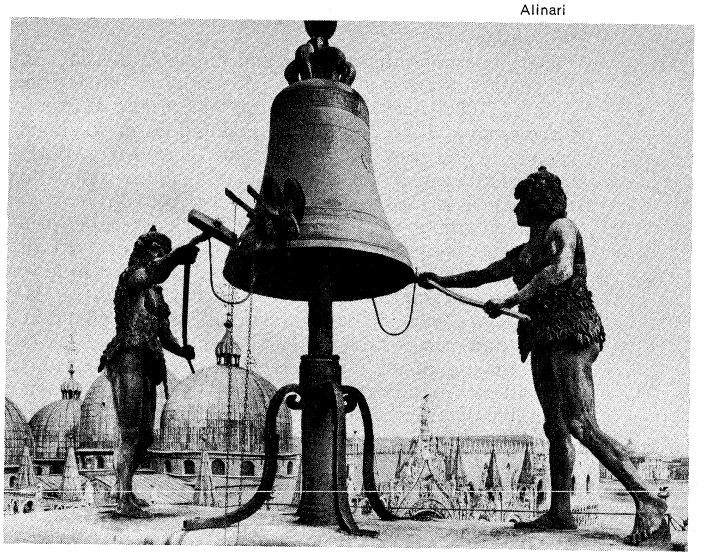
The majority of automata are direct representations of creatures and plants or of kinetic aspects of natural phenomena. Imitations of such natural phenomena as the moving water of streams and waterfalls, for instance, can be simulated with twisted rods of glass. A mechanical device can be used to make a flower open its petals to imitate blooming or to make a figure walk. Some purely capricious automata consist of complete scenes in which caricature personages perform in a humourous manner. Not all automata, however, are mimetic. Some offer only visual fascination, such as spinning roundels set with gems to make flashing patterns of color and light.

Automata classified

Automata can be generally classified into two groups: those that are ancillary to a functional article and those that in themselves are fanciful objects, solely for decoration and pleasure. Clocks and watches, which lend themselves to displays of motion, are the most common type of functional object with automata. With or without the accompanying music of bells or other instruments, clock automata can take a variety of forms. In 1599, for example, Queen Elizabeth I of England sent to the Turkish Sultan a 16-foot-high (4.877-metre) clock that included not only an organ and bells but also mechanical figures of trumpeters and a holly bush with blackbirds and thrushes that sang and moved.

Early mechanical clocks, dating from the 14th century, were set up in public places and equipped with human-looking figures that marked the hours by striking bells with hammers. Known as jaquemarts or jacks, these figures replaced live watchmen or bell ringers who had performed this task. Increased mechanical knowledge and technical skill resulted in more complicated displays. Public clocks showed joustings, processions, or religious scenes, such as the scene of the adoration of the Magi begun in 1352 for the Cathedral of Strasbourg in France and the famed procession of puppets made in 1356–61 for the clock of the Frauenkirche at Nürnberg in Germany. One of the most famous public clocks with automata is located in Venice's Piazza di S. Marco. A three-level clock with jacks, it was built in 1496–99 from the designs of the Renaissance architect Mauro Coducci (1440–1504). Clocks for the home performed comparably to these elaborate works, but on a smaller scale. From *c.* 1775 automata, especially tiny jacks that beat out the time, became increasingly popular for watches.

Alinari



The two "Mori" or jacks by Mauro Coducci, 1496–99. The figures strike the hour on the Torre dell'Orologio in the Piazza di S. Marco, Venice.

Frequently, functional objects became overburdened with decorative mechanisms, and the objects' primary utilitarian function was subordinated. The simple snuff box, for example, eventually became so encumbered with clockwork mechanisms for telling time, providing music, and operating moving figures that only a tiny space remained for holding snuff. Thus, the automaton itself became the centre of interest, and the container was often designed around it, rather than vice versa. The majority of automata have been objects of fancy that are purely decorative in concept and function. The most complicated are the androids: figures in human form that can be made to walk about, play music, write, or draw. They are mostly of fairly large size and intended for public display. At the other end of the scale are exquisitely finished, pocket-sized objects such as trick pistols that were the speciality of the Rochat brothers, Ami-Napoléon and Louis, both of whom were among the finest automata designers and craftsmen of the early 19th century. These pistols are made of gold or gilded silver, enamelled, and set with pearls and precious stones. When the trigger is pressed, a tiny bird appears at the end of the barrel and proceeds to sing, simultaneously moving its wings and turning its head from side to side, or a flower shoots forth in full bloom and sprays perfume.

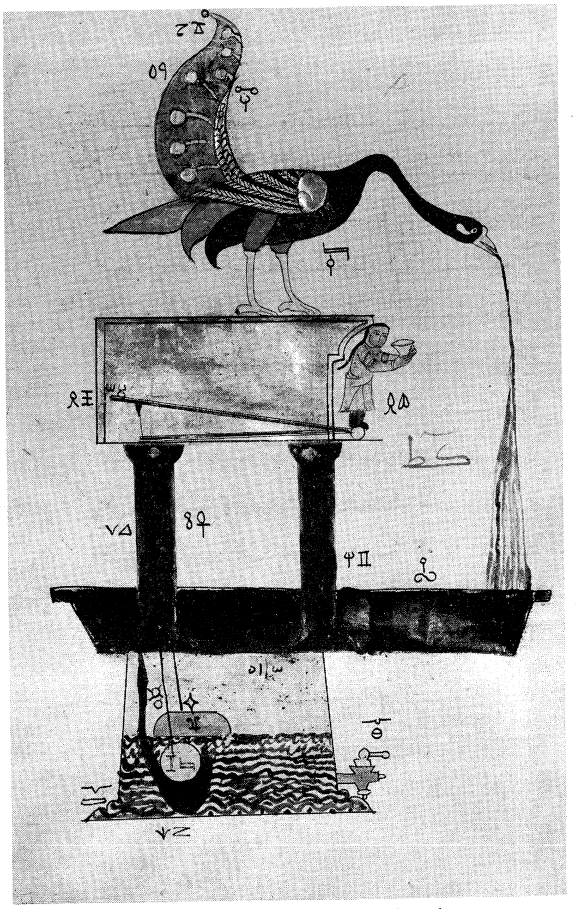
Automata as decorative objects

Historical development. Few examples of automata made prior to the 16th century remain, but numerous documents record their one-time existence. Among the earliest references is one to a wooden model of a pigeon constructed by Archytas of Tarentum (flourished 400–350 BC), a Greek friend of Plato. The bird was apparently suspended from the end of a pivoted bar, and the whole apparatus revolved by means of a jet of steam or of compressed air. More complete information about other devices is found in the writings of Hero of Alexandria (flourished 1st century AD), who described devices actuated by water, falling weights, and steam.

Some of the most elaborate examples of medieval automata were designed by the early Byzantines. Most renowned of these was a great clock erected in about 500 AD at Gaza, Syria. It was operated by water and centred on a figure of Hercules that struck the hours with its club. Later in date was a gold throne made *c.* 835 for the Byzantine emperor Theophilus (reigned 829–842), who is said to have employed it to impress foreigners. When required, figures of lions supporting the royal seat roared realistically, while mechanical birds sang in imitation trees at either side.

During the Middle Ages in the Islāmic world, there were a number of inventors active from about the 9th century. Best documented are the water-operated automata, many of moving peacocks, invented and made by al-Jazarī, who worked in the 13th century for princes of the Urtugid dynasty in Mesopotamia.

By courtesy of the Museum of Fine Arts, Boston



An automated peacock fountain from the treatise of al-Jazarī, Mesopotamian, 13th century AD. In the Museum of Fine Arts, Boston.

References to automata devised by Western Europeans in the Middle Ages cite such distinguished names as Roger Bacon (*c.* 1220–*c.* 1292) and Albertus Magnus (*c.* 1200–1280), both of whom are credited with constructing androids—Bacon, a talking head, and Albertus, an iron man. Decorative mechanical objects for ecclesiastical use are illustrated by the Gothic architect Villard de Honnecourt in his famed sketchbook (1235). He drew, for example, a movable eagle to be operated when scripture was read. A sensational medieval automaton was a "magic fountain" designed in 1242 by Guillaume Boucher, a Parisian goldsmith, for the Mongol Möngke Khan (1208–1259), who held Boucher captive in Belgrade. Late Gothic examples included the flying iron eagle and fly of the German astronomer and mathematician Johann Müller (1436–76), known as Regiomontanus. When demonstrated for the emperor Maximilian, the fly was said to have alighted on the imperial arm. Perhaps the inspiration for this construction was the legendary mechanical fly of the medieval magician Vergil, which was reputed to have kept the Neapolitan meat market free of other flies.

In the early 16th century there was renewed interest in the manufacture of automata, largely stemming from the influence of Eastern examples brought to Europe through trade with the Orient and the translation from the ancient Greek of the 1st century AD writings on mechanical objects by Hero of Alexandria. Intricate fountains emphasizing spectacular and trick effects became highly fashionable among the wealthy. Notable among them were the mid-16th century fountains and waterworks built for the gardens of the Villa d'Este at Tivoli, Italy.

With the use of coiled tempered steel spring from the mid-15th century, a truly portable source of motion became available in the Renaissance. It was used, for instance, in some of the nefs, table ornaments in the form of sailing ships. Largely dating from the second half of the 16th century, nefs probably originated in the gold and silversmithing centres of Germany, namely, Augsburg and Nürnberg, with such important masters of mechanical construction and the jeweler's craft as Hans Schlottheim (1547–1625). Among the most celebrated nefs is the "Ship of Charles V" (Musée de Cluny, Paris). When the hour strikes, a miniature organ plays, and the approximately 2½-foot-long by 3¼-foot-high (.7620-metre by 1-metre) boat rocks, with musicians playing instruments on the deck and sailors moving in the rigging; ten courtiers pay their respects to Charles, who bows in acknowledgment.

Androids were also popular in the Renaissance. One of the best preserved examples is a mandolin-playing lady (*c.* 1540; Kunsthistorisches Museum, Vienna) attributed to Giannello Torriano of Cremona (*c.* 1500–85), clockmaker to Charles V (of the Holy Roman Empire).

It was in the later 18th and early 19th centuries that the most intricate automata made their appearance. Typical are the objects made by the Rochat brothers, who specialized in the manufacture of miniature singing birds. The mechanical songbirds were devised to appear suddenly from beneath hinged panels in snuffbox tops or to operate in cages that were suspended so that a clock under the base was visible. Such birds might even be concealed in a bracelet or a pocket watch. Perhaps the most intriguing of small-size automata were the so-called magician boxes. A disk engraved with a question is inserted in a slot in the box, upon which the tiny figure of a magician comes to life and points with his wand at a space where the answer appears.

Among the more elaborate mechanical devices popularized in the 18th century were *tableaux mecaniques*, or mechanical pictures. These framed painted landscapes, in which figures, windmills, and so forth spring to life by means of hidden clockwork, remained popular through the 19th century. A tableau designed for Mme de Pompadour (1759; Conservatoire National des Arts et Métiers, Paris) is a prime example of this type of automaton. Closely related to the *tableaux mécaniques* are mechanical theatres, the most extravagant of these having been built in the gardens of Hellbrunn near Salzburg. Consisting of 113 hydraulically operated figures, it was assembled between 1748 and 1752.

The most intricate mechanisms of all were those employed for the androids. The first famed maker of them was a Frenchman, Jacques de Vaucanson (1709–82), who in 1738 created a flute player capable of playing a dozen songs. His most distinguished followers were the Swiss Jaquet-Droz, father and son: Pierre (1721–90), who worked in conjunction with Jean-Frédéric Leschot (1746–1824), and Henri-Louis (1752–91). They constructed figures to write and draw and others to play musical instruments. Later came the Maillardet family, who worked on into the 19th century, when simplified versions of clockwork-operated automata began to be produced in tin and cardboard.

From the late 16th century onwards, automata featured among exports from West to East. Ambassadors and representatives of trading companies were agreeably received upon presenting such objects. Invariably complex in action and sumptuously cased, they were displayed with pride in the palaces of the Far East and India. The most prominent European supplier was James Cox (died 1788), of London, whose 18th-century clients included the East India Company. Turkey was another profitable market for English-made clocks with automata. Markwick Markham of London supplied (*c.* 1750–75) many clocks for sale. All have Turkish numerals on their faces, and some are embellished with automata.



(Top left) Gold and enamel snuffbox with automated figures in a garden scene and a carillon musical movement, Geneva, *c.* 1820. In a private collection. Length 7.62 cm. (Top centre) Silver swan that swims in water of glass rods, London, 18th century. In the Bowes Museum, England. Length 1.7 m. (Top right) Android of a child writing by P. Jaquet-Droz, *c.* 1772. In the Musée d'Art et d'Histoire, Neuchâtel, Switzerland. Height 75 cm. (Bottom left) Nef, the so-called "Ship of Charles V," 16th century. In the Musée de Cluny, Paris. Length 76 cm. (Bottom right) *Tableau mécanique* designed by Desmares for Mme de Pompadour, 1759. In the Conservatoire National des Arts et Métiers, Paris. 60 × 67 cm.

By courtesy of (top centre) the Bowes Museum, Barnard Castle, County Durham, England, (top right) Musee d'Art et d'Histoire, Neuchatel, Switzerland, (bottom right) Musee des Techniques, Paris; photographs, (top left) A La Vieille Russie, New York/Paris, (bottom left) Giraudon

Automata in China and India

Accounts of automata in China date from as early as the 3rd century BC, during the Han dynasty, when a mechanical orchestra was made for the emperor. By the Sui dynasty, in the 6th and 7th centuries AD, automata had become widespread, and a book entitled the *Shui shih t'u Ching* ("Book of Hydraulic Elegancies") was published. In the T'ang period, from the 7th to the 10th centuries AD, automata continued to be popular with imperial circles. There are records of flying birds, an otter which caught fish, and figures engaged in numerous activities ranging from a monk begging to girls singing. After the Yüan period (1279–1368), the creation of automata seems to have waned. In the 18th and 19th centuries during the Ch'ing period, however, automata were again made and were often inspired by those in clocks imported from Europe. The imperial Summer Palace in Peking was the principal repository of the great collection of automata formed by the emperor Ch'ien-lung in the second half of the 18th century. Much of the contents of the palace was destroyed or looted by European soldiers in 1860 and again in the Boxer Rebellion of 1900, leaving little evidence of the artistic and mechanical merit of Chinese automata.

In India there is little more than references in legends and myths to mechanical flying chariots, animals, and dancing figures. With the exception of a few brief entries in various treatises, most of the knowledge of construction seems to have been kept secret and passed orally from one craftsman to another. There are no examples, therefore, of automata before the Mughal period of the 16th to the 18th centuries. Most extant automata in India were imported from the Near and Far East and Europe, especially England and France.

20th century

With the exception of a few works by Peter Carl Fabergé (1846–1920), the production of costly artistic automata virtually ceased in the late 19th and early 20th centuries because of the diminishing number of skilled craftsmen, as well as rich patrons to support them. Collecting, therefore, is reserved for only the most wealthy. This expensive hobby is still served by the dealer who locates increasingly rare examples of historic automata and by a small corps of highly skilled craftsmen whose dearly priced services keep the objects in working order.

BIBLIOGRAPHY. Splendidly illustrated monographs on this subject include: A. CHAPUIS and E. GÉLIS, *Le Monde des Automates,* 2 vol. (1928), with an extensive bibliography; A. CHAPUIS and E. DROZ, *Les Automates* (1949; Eng. trans., *Automata: A Historical and Technological Study,* 1958), a re-assessment of the earlier work by Chapuis and Gélis; and E. MAINGOT, *Les Automates* (1959). A description with illustrations of some examples of 13th- to 18th-century automata may be found in W. BORN, "Early European Automatons," *Connoisseur,* vol. 100, pt. 1–2, pp. 123–129, 246–252 (1937).

(G.Wi.)

Automata Theory

In simple terms an automaton represents a formalization of a set of rules for a computation, and automata theory, studied as part of the foundations of mathematics, is used in the building of such machines as all-purpose computers. An example of a typical automaton is a pendulum clock. In such a mechanism the gears can assume only one of a finite number of positions, or states, with each swing of the pendulum. Each state, through the operation of the escapement, determines the next succeeding state, as well as a discrete output, displayed as the discrete positions of the hands of the clock. As long as such a clock is wound and its operation is not interfered with, it will continue to operate unaffected by outside influences except the effect of gravity on the pendulum.

More general automata are designed to respond to changes in external conditions or to other inputs. For example, thermostats, automatic pilots of aircraft, missile guidance systems, telephone networks, and controls of certain kinds of automatic elevators are all forms of automata.

The internal states of such devices are not determined solely by their initial state, as is the case of the pendulum clock, but may be determined by an input from a human operator, from another automaton, or by an event or series of events in the environment. A thermostat, for instance, has an "on" or "off" state that depends on the temperature. Perhaps the most widely known example of a general automaton is the modern electronic computer, the internal states of which are determined by the data input and which operates to produce a certain output.

The components of automata consist of specific materials and devices, such as wires, transistors, levers, relays, gears, and so forth, and their operation is based on the mechanics and electronics of these parts. The principles of their operation as a sequence of discrete states can, however, be understood independently of the nature or arrangement of their components. In this way, an automaton may be considered, abstractly, as a set of physically unspecified states, inputs, outputs, and rules of operation, and the study of automata as the investigation of what can be accomplished with these. This mode of abstraction yields mathematical systems that in certain respects resemble logical systems. Thus, an automaton can be described as a logically defined entity that can be embodied in the form of a machine, with the term automaton designating both the physical and the logical constructions.

The Turing machine

In 1936 an English mathematician, Alan Mathison Turing, in a paper published in the *Proceedings of the London Mathematical Society* ("On Computable Numbers with an Application to the Entscheidungsproblem"), conceived a logical machine the output of which could be used to define a computable number. For the machine, time was considered to be discrete and its internal structure, at a given moment, was described simply as one of a finite set of "states." It performed its functions by scanning an unbounded tape divided into squares, each of which either contained specific information in the form of one of a finite number of symbols or was blank. It could scan only one square at a time, and, if in any internal state except one called "passive," it was capable of moving the tape forward or backward one square at a time, erasing a symbol, printing a new symbol if the square was blank, and altering its own internal state. The number it computed was determined by symbols (the "program") provided on a finite portion of the tape and the rules of operation, which included stopping when the passive state was reached. The output number was then interpreted from the symbols remaining on the tape after the machine stopped.

Automata theory since the middle of the present century has been extensively refined and has often found practical application in civilian and military machines. The memory banks of modern computers can store large (though finite) amounts of information. The original Turing machine had no limit to the memory bank because each square on the unbounded tape could hold information. The Turing machine continues to be a standard reference point in basic discussions of automata theory, and many mathematical theorems concerning computability have been proved within the framework of Turing's original proposal.

Part of automata theory lying within the area of pure mathematical study is often based on a model of a portion of the nervous system in a living creature and on how that system with its complex of neurons, nerve endings, and synapses (separating gap between neurons) can generate, codify, store, and use information. The "all or none" nature of the threshold of neurons is often referred to in formulating purely logical schemata or in constructing the practical electronic gates of computers. Any physical neuron can be sufficiently excited by an oncoming impulse to fire another impulse into the network of which it forms a part, or else the threshold will not be reached because the stimulus is absent or inadequate. In the latter case, the neuron fails to fire and remains quiescent. When several neurons are connected together, an impulse travelling in a particular part of the network may have several effects. It can inhibit another neuron's ability to release an impulse; it can combine with several other incoming impulses each of which is incapable of exciting a neuron to fire but that, in combination, may provide the threshold stimulus; or the impulse might be confined within a section of the nerve net and travel in a closed loop, in what is called "feedback." Mathematical reasoning about how nerve nets work has been applied to the problem of how feedback in a computing machine can result in an essential ingredient in the calculational process.

Original work on the neurophysiological aspect of automata theory was done by Warren S. McCulloch and Walter Pitts at the Research Laboratory of Electronics at the Massachusetts Institute of Technology starting in the 1940s.

While the automata themselves are prototypes of deterministic machines, the U.S. mathematician Norbert Wiener (*q.v.*) showed that they may be programmed in such a way as to extrapolate certain types of random data that are introduced as input. A prediction of data that are not yet received as input can be accomplished, provided the data are what will later be defined to constitute a stationary time series and provided the prediction is restricted according to a well-defined optimization procedure. In this way a logically defined robot, or automaton, may be placed in an environment that evolves according to both deterministic and random processes (the bifurcation of the environment into deterministic and random processes being mathematically postulated by the designer of the robot) and may be seen to respond to the advantage of its designer: The robot can control a ship's rudder, guide an airplane to its landing, reorient a rocket on its course, predict weather, and so forth. The programming of an automaton so that it will react in a suitable way when placed in a naturalistic environment falls under the heading of prediction theory. (B.R./R.J.Ne)

Use of automata for prediction

NEURAL NETS AND AUTOMATA

The finite automata of McCulloch and Pitts. The definitions of various automata as used here are based on the work of two mathematicians, John von Neumann (*q.v.*) and Stephan Cole Kleene, and the earlier neurophysiological researches of McCulloch and Pitts, which offer a mathematical description of some essential features of a living organism. The neurological model is suggested from studies of the sensory receptor organs, internal neural structure, and effector organs of animals. Certain responses of an animal to stimuli are known by controlled observation, and, since the pioneering work of a Spanish histologist, Santiago Ramón y Cajal, in the latter part of the 19th and early part of the 20th century, many neural structures have been well-known. For the purposes of this article, the mathematical description of neural structure, following the neurophysiological description, will be called a "neural net." The net alone and its response to input data are describable in purely mathematical terms.

The neural net

A neural net may be conveniently described in terms of the kind of geometric configuration that suggests the physical structure of a portion of the brain. The component parts in the geometric form of a neural net are named (after the physically observed structures) neurons. Diagrammatically they could be represented by a circle and a line (together representing the body, or soma, of a physiological neuron) leading to an arrowhead or a solid dot (suggesting an endbulb of a neuron). A neuron may be assumed to have either an excitatory or an inhibitory effect on a succeeding one; and it may possess a threshold, or minimum number of unit messages, so to speak, that must be received from other neurons before it can be activated to fire an impulse. The process of transmission of excitation mimics that which is observed to occur in the nervous system of an animal. Messages of unit excitation are transmitted from one neuron to the next, and excitation is passed along the neural net in quantized form, a neuron either becoming excited or remaining non-excited, depending on the states (excitatory or quiescent) of neurons whose endbulbs impinge upon it. Specifically, neuron N, with threshold h, will be excited at time t, if and only if h or more neurons whose excitatory endbulbs impinge upon it are excited at time $t - 1$ and no neuron whose inhibitory endbulb impinges upon it is excited at time $t - 1$. A consistent picture can be made of these conditions only if time and excitation are quantized (or pulsed). It is assumed conventionally that a unit of time is required for the transmission of a message by any neuron.

Certain neurons in the configuration mathematically represent the physiological receptors that are excited or left quiescent by the exterior environment. These are called input neurons. Other neurons called output neurons record the logical value, excited or quiescent, of the whole configuration after time delay t and transmit an effect to an exterior environment. All the rest stimulate inner neurons.

Any geometric or logical description of the neural structure of an organism formulated as the basis of physical construction must be sufficiently simple to permit mechanical, electric, or electronic simulation of the neurons and their interconnections.

The basic logical organs. The types of events that can excite the automaton and the kinds of responses that it can make must next be considered. By stripping the description down to the most simple cases, the basic organs from which more complicated robots can be constructed may be discovered. Three basic organs (or elementary automata) are necessary, each corresponding to one of the three logical operations of language: the binary operations of disjunction and conjunction, leading to such proposition as $A \cup B$ (read "A or B"), $A \cap B$ (read "A and B"), and the unary operation of negation or complementation, leading to such propositions as A^c (read "not A" or "complement of A"). First to be considered are the stimulus–response pattern of these elementary automata.

Binary operations of disjunction and conjunction

Assuming that a neuron can be in only one of two possible states—*i.e.*, excited or quiescent—an input neuron at a given instant of time $t - 1$ must be either excited or nonexcited by its environment. An environmental message transmitted to two input neurons N_1 and N_2 at time $t - 1$ can then be represented numerically in any one of the four following ways, in which binary digit 1 represents excitation and binary digit 0 represents quiescence: (0, 0), (0, 1), (1, 0), (1, 1). The disjunction automaton must be such that a single output neuron M correspondingly registers at time t the response: 0, 1, 1, 1. The conjunction automaton must be such that a single output neuron M correspondingly registers at time t the response: 0, 0, 0, 1. The negation automaton considered as having two input neurons N_1 and N_2, of which N_1 is always excited, must respond to the environmental messages (1, 0) and (1, 1) with 1, 0, respectively, at the output neuron M.

The generalized automaton and Turing's machine. The construction of more complicated robots from these basic building blocks constitutes a large part of the theory of automata. The first step in the direction of generalization is to define the neural nets that correspond to formal expressions in n variables of the propositional calculus—that is, the formal system that concerns "or," "and," "not," and "implies." A single output automaton (of which the above three are simple examples) is a neural net with n input neurons, one output neuron, and with interconnections between neurons that conform to the rule that no neuron stimulated at time t can impinge upon a neuron that could have experienced its first stimulation at the same or an earlier time. The latter rule is the requirement of no feedback. Given this concept of a single output automaton, it is possible to examine the output response at time $t + s$, considered as a function of the configuration of stimuli at the n input neurons at time t. This response can be compared with the truth value of a logical statement (polynomial) from the propositional calculus. A logical statement is formed from n component propositions, each of which can assume the truth value either true or false. The comparison between automaton and logical statement is accomplished by matching response at the output neuron at time $t + s$ with truth value of the statement for every one of the 2^n cases in which the configuration of stimuli conforms to the configuration of truth values of the component propositions. If, in this sense of comparison, the functional response of the automaton is identical to the functional value of the logical statement (polynomial), the automaton is then said to compute the statement (polynomial) or the statement is said to be computable. A wider class of computable statements is introduced with the general automaton, yet to be defined, as with the more general Turing machine.

Response of automata and truth value of statements

The important distinction between the logical statement and the automaton that computes it is that the first is free of any time ingredient while the second is defined only with reference to a time delay of length s.

A basic theorem states that for any polynomial P of the propositional calculus, there exists a time delay s and a single output automaton A, such that A computes P with time delay s. The proof of the theorem rests on the fact from the propositional calculus that all statements are composed from component propositions with the operations of disjunction, conjunction, and negation and the fact from the automata theory that all single output automata can be composed by interconnecting elementary automata of the disjunctive, conjunctive, and negative types.

A second step of generalization in the construction of robots proceeds from the single output automata to the neural net that possesses more than one output neuron and in which the internal connections may include feedback. Such a construction is called a "general automaton." The class of general automata includes all-purpose, electronic digital computers the memory-storage units of which are of fixed, though possibly of very considerable, size. It is within the context of the general automaton that the purely automated decision-making, computing, controlling, and other sophisticated neural functions so suggestive of the mental ability of man may appropriately be discussed.

The Turing machine can be defined not only as it was in the introduction (following roughly Turing's approach) but as a general automaton to which an unbounded memory unit (such as an unbounded tape) is added. Thus, the general automaton and the Turing machine differ in logical design only with respect to the extent of memory storage.

Extent of memory in a Turing machine

The distinction is critical, however, for Turing proposed that the class of numbers computable on his machine (a wider class than can be obtained by general automata) coincide with those that are effectively computable in the sense of constructive logics (see MATHEMATICS, FOUNDATIONS OF). A simple convention also makes it possible to interpret the output of a Turing machine as the computation of a function. The class of functions so computed, called "Turing computable" or "computable," are of basic importance at the foundations of mathematics and elsewhere. It can also be stated that a useful class of functions that are definable without

reference to machines, namely, the so-called partial recursive functions, has the same membership as the class of computable functions. For the present purposes, then, no effort need be made to define the partial recursive functions.

Turing's approach admitted mathematical formalization to the extent that a finite list of symbols q_1, q_2, q_3, $\cdots$, q_n could be used to denote internal states and a finite list of symbols a, b, c, $\cdots$, λ could designate abstractly what is called "the alphabet"—that is, the list from which individual members could be chosen and printed into the squares of the machine's tape. If the symbols R and L, respectively, designate a move of the tape one square to the right and one square to the left, it remains only to list in some orderly fashion the alternative possible steps in the machine's operation in order to define it completely. Turing himself chose to list alternate steps, or instructions, in the form of quintuples of the above symbols. It is also possible to use quadruples to define a machine. Such a list, then, of, say, quadruples of instructions is equivalent to a Turing machine, and it is significant that the list is finite.

The finiteness of the list of quadruples of instructions leads to the idea that all Turing machines can be listed —that is, they are at most countable in number. This being the case, it can be proved that there is what Turing called a "universal" machine capable of operating like any given Turing machine. For a given partial recursive function of a single argument, there is a corresponding integer, called the "Gödel number," that identifies the Turing machine capable of computing the given function. The Gödel number and the argument value of the function to be computed can be given as input data on the tape of the universal machine. From the Gödel number, the list of instructions, defined in the form of quadruples, that are necessary for the computation of the given recursive function at the specific argument value can be encoded by the universal machine on its own tape, and, from that point on, the universal machine will duplicate the required Turing machine.

The "universal" machine

Input: events that affect an automaton. Once having reached the definition of the general automaton and the more general universal Turing machine, a general definition of the events in the environment that stimulate it may be introduced. The automaton, which computes logical statements, is not defined without reference to time, a characteristic that distinguishes the machine itself from the logic. In the same way, stimuli are not definable, in general, without reference to time. These facts are indicative of the simulation features that the computing machine bears with respect to man.

For an automaton with n input neurons, $N_1, N_2, \cdots, N_n$, an individual history of stimulation, starting with the present moment, $t = 0$, and continuing to the remote past, can be recorded as a sequence of n-tuples, $(\beta_1, \beta_2, \cdots, \beta_n)$, in which each binary digit, β_k, is either a 0 or a 1. Thus, the beginning of one such individual history for an automaton of four neurons might be recorded in tabular form as an unending list of quadruples of the type (1, 0, 1, 1) (see Box, display 1).

Events as collections of individual histories

An event is a collection of individual histories. This is a generalization of the idea already used to characterize an environmental message transmitted to the two input neurons of an elementary automaton at time $t - 1$. As an example, the stimulus (0, 1) is the same as the collection of all individual histories in which neuron N_2 was stimulated at time $t - 1$ and neuron N_1 was not. As another example, the event that neuron N_2 (of a two-neuron automaton) is presently stimulated and has always been stimulated on alternate second can be represented as the collection of two individual histories (see 2). While some events require an infinite tabulation, others that specify the states of each neuron over a finite past (allowing that anything might have occurred before) permit a finite tabulation. Events of the second kind are called definite events, or stimuli.

The construction (either actual or theoretical) of a general automaton with the help of the logical components and interconnections of a neural net results in an entity that responds in reproducible ways to stimuli. A response becomes recorded as a configuration of binary digits, corresponding to the states of the finite number of output neurons at a specified time t in the future, while a stimulus is a collection of individual histories extending over the past and including the present. The logical construction implies a behaviour in the guise of a listing of responses to all possible stimuli. Reciprocally, for a given behaviour of the type defined, the possible structure of a machine that could produce such behaviour can be investigated.

PROBABILISTIC QUESTIONS

It was traditional in the early treatment of automata theory to identify an automaton with an algorithm, or rule of computation, in which the output of the automaton was a logically determined function of the explicitly expressed input. From the time of the invention of the all-mechanical escapement clock in Europe toward the end of the 13th century, through the mechanistic period of philosophy that culminated in the work of the French mathematician Pierre-Simon Laplace (*q.v.*), and into the modern era of the logically defined Turing machine of 1936, an automaton was a mechanical or logical construction that was free of probabilistic components. It was also understood to be immersed in an environment (that is, activated or supplied with input data) that could be logically specified without the concept of chance. The strength of the mechanistic philosophy as it influenced science prior to the 20th century is illustrated in a quotation from Laplace in PROBABILITY, THEORY OF: *Foundations of 20th-century probability theory*.

Automata and the idea of chance

After the middle of the 20th century, mathematicians explicitly investigated questions concerning automata that included in their formulation the idea of chance, and in doing so they drew upon earlier applicable mathematical results.

Of the types of probabilistic questions considered, four (which will be listed in arbitrary order) were predominant. The first, that of Norbert Wiener, was broached in 1948. It concerned the use of mathematically expressed algorithms or physically constructed computers to predict the future of a system, such as the weather, that includes random components—*i.e.*, an automaton in Turing's logical sense immersed in a random environment. The second, of von Neumann, was concerned with the reliability of large computing machines with many components and sought methods of design, called "multiplexing," that would reduce the chance for unwanted error during the machine calculation of a problem. In this context, the automaton was interpreted as a randomly operating device that in practice approximates the operation of a Turing machine under the influence of better and better design. The third, considered by various researchers, concerned the possibility of computing a wider class of sets than are accessible to Turing machines by adding a random component to the machine itself. In this context, the automaton was being interpreted as a Turing machine modified with the potentiality for injecting the output of a random number generating device into one or more of its operational steps. The fourth concerned the logical possibility of an automaton, such as a Turing machine, actually yielding as output a sequence of random numbers. In this context, the automaton was considered to be simultaneously a Turing machine and a generator of numbers that are indistinguishable from measurements on random phenomena.

Some results that have been achieved from examination of each of these four types of questions will constitute the remainder of this section.

The automaton and its environment. It must first be observed that, just as an automaton is an acceptable description (or model) of a neural structure, an automaton, though frequently thought of as a computing machine, is in general a response mechanism that produces output (or behaviour) as a consequence of the input (or environmental stimuli). "Environment" is then another

name for the input and output of an automaton. Some poetic license in identifying automata with living things may justify the use of the term.

During his researches on cybernetics, Wiener recognized that, if computers could be programmed to solve certain mathematical equations, then the data read from physically generated time series (or numerical values indexed consecutively in time and related through a transformation) could be extrapolated. He saw that, if this process could be accomplished with sufficient speed, as would be possible with modern electronic circuits, then the extrapolated values would be obtained faster than the actual physically evolving process that produced the time series, and a prediction of the future would result. Errors would be inevitable because a complete history of data and adequate measurements would be unobtainable. For this reason, the mathematical equations that would be at the heart of such an extrapolation could be deduced, in part, from the objective of minimizing the errors. Thus, the matching of an automaton, or computer, with a real physical environment could result in the anticipation of the future, if certain mathematical equations were derived that minimized prediction error.

Matching an automaton with an environment

Control and single-series prediction. A derivation of the mathematical equations of prediction had been accomplished in a limited sense some years before Wiener's work on cybernetics. In 1931 Wiener had collaborated with an Austrian-born U.S. mathematician, Eberhard Hopf, to solve what is now called the Wiener–Hopf integral equation, an equation that had been suggested in a study of the structure of stars but later recurred in many contexts, including electrical-communication theory, and was seen to involve an extrapolation of continuously distributed numerical values. During World War II, gun- and aircraft-control problems stimulated further research in extrapolation, and Wiener composed a purely mathematical treatise, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, which was later published in 1949. As early as 1939, a note on extrapolation by a Russian mathematician, A.N. Kolmogorov (*q.v.*), had appeared in the French journal *Comptes Rendus*. Although the Wiener–Hopf work was concerned exclusively with astronomy and done without the guiding influence of computers, it was recognized in World War II that high-speed computations could involve input information from a moving object and, through prediction or extrapolation, provide output data to correct its path. This recognition was the seed of the concept of the guided missile and radar-controlled aircraft. Weather prediction was also possible, as was computerized research on brain waves whose traces on the electroencephalograph offered another physical realization of the time series that are predictable. The mathematics that was necessary for a complete understanding of prediction included the concept of a stochastic process, as described in the article PROBABILITY, THEORY OF.

The Wiener and Kolmogorov research on extrapolation of time series became known as single-series prediction and owed much to the studies (1938) of a Swedish mathematician named H. Wold, whose work was predicated on the assumption that, if X_1, X_2, $X_3 \cdots$ are successive values of a series identified with discrete points in time $t = 1$, $t = 2$, $t = 3$, $\cdots$, then the successive values are not entirely unrelated (for if they were, there would be no way for an algorithm or an automaton to generate information about later members of the sequence—that is, to predict). It was assumed, with anticipation that there is frequently such a thing in nature, that a transformation T relates members of the series by successively transforming an underlying space of points ω according to a rule. The rule states that the kth member of the time series is a function of an initial point ω that has migrated in the underlying space $X_k = X(T^k\omega)$. It was also assumed that, if sets of points $\{\omega\}$ constituted a region (of sufficient simplicity called "measurable") in space, then when the set was transformed under the influence of T its volume would not be changed. The last assumption had, in fact, been proved by a French mathematician, Joseph Liouville, a century earlier for a wide class of physical processes whose behaviour is correctly described by the so-called Hamiltonian equations. The clearly stated assumptions of Wiener and Kolmogorov, referred to as the stationarity of the time series, were supplemented with the idea (the linearity restriction) that a solution $S_k(\omega)$ for the predicted value of the series, displaced in time k steps into the future, should be restricted to a linear combination of present and past values of the series (see 3).

With the help of one other mathematical assumption, it was then possible to solve the single-series prediction problem by specifying an algorithm that would determine the coefficients in the linear combination for $S_k(\omega)$, in which k is a positive integer (see 4). It was possible also to solve for the error of prediction (see 5)—that is,

The assumptions of prediction theory

(1)
$$\begin{array}{l}(1,0,1,0)\\(1,0,1,1)\\(0,0,0,1)\\(1,0,1,0)\\\cdots\\\cdots\\\cdots\end{array}$$

(2)
$$\begin{array}{ll}(0,1) & (1,1)\\(0,0) & (1,0)\\(0,1) & (1,1)\\(0,0) & (1,0)\\\cdots & \cdots\\\cdots & \cdots\\\cdots & \cdots\end{array}$$

(3) $$S_k(\omega) = \sum_{n=0}^{\infty} X(T^{-n}\omega)\, P_n^{(k)},$$
with convergence defined in the L_2-metric.

(4)
An algorithm for computing the coefficients $P_n^{(k)}$ in the prediction $S_k(\omega)$ is the following:
From the auto-correlation (X_{-n}, X_0) of the time series compute $|\psi(\theta)|$:
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\psi(\theta)|^2 e^{in\theta}\, d\theta = (X_{-n}, X_0).$$
From $|\psi(\theta)|$ compute $\psi(r, \theta)$:
$$\log \psi(r,\theta) = \frac{1}{2\pi}\left\{\sum_{n=1}^{\infty} r^n e^{in\theta}\int_{-\pi}^{\pi} \log|\psi(x)|^2 e^{-inx}\, dx + \int_{-\pi}^{\pi} \log|\psi(x)|\, dx\right\}.$$
From $\psi(r, \theta)$ compute (X, h_{-n}):
$$\psi(r,\theta) = \sum_{n=0}^{\infty} r^n e^{in\theta}(X, h_{-n}), \qquad 0 \leq r < 1.$$
From (X, h_{-n}) compute a_n:
$$(X, h)a_0 = 1$$
$$\sum_{n=0}^{m} (X, h_{n-m})a_n = 0; \qquad m > 0.$$
From a_n and (X, h_{-n}) compute $P_n^{(k)}$:
$$P_n^{(k)} = \sum_{m=0}^{n} a_{n-m}(X, h_{-m-k}), \qquad k > 0.$$

(5) $$\sigma_k^2 = \sum_{n=0}^{k-1} |(X, h_{-n})|^2$$

a measure of the discrepancy between the value predicted and the true value of the series that would occur at time k in the future. This meant that for a variety of circumstances, such as the prediction of atmospheric pressure measured at one weather station, or the prediction of a single parameter in the position specification of a particle (such as a particle of smoke) moving according to the laws of diffusion, an automaton could be designed that could sense and predict the chance behaviour of a sufficiently simple component of its environment.

Multiple-prediction theory. Generalizations of the above limited accomplishments are tantalizing to mathematicians. If animals, and man in particular, are viewed, even in part, as automata with varying degrees of accomplishment and success that depend on their abilities to cope with their environment, then man himself could be better understood and his potentialities could be further realized by exploring a generalized version of an automaton's ability to predict. Success in generalizations of this kind have already been achieved under the heading of what is called multiple-prediction theory. A reference to the problem of multiple prediction without a complete solution was made as early as 1941 by a Russian mathematician, V. Zasuhin. The first major step forward, after Zasuhin, was taken by Wiener in 1955 under the title "On the Factorization of Matrices." Many significant results soon followed.

If multiple-prediction theory is identified with part of automata theory (which is not always done), it is possible to consider the construction of a computing machine, or automaton, capable of sensing many interdependent elements of its environment at once and, from a long history of such data, of predicting a future that is a function of the same interdependent elements. It is recognized that multiple prediction is the most general approach to the study of the automaton and its environment in the sense that it is a formulation of prediction free of the linearity restriction earlier mentioned with reference to single series (see 3). To express a future point $S_k(\omega)$, for example, as a linear function of its present and past values as well as first derivatives, or rates of change, of its present and past values is to perform a double prediction or prediction based on the two time series $X_1, X_2, X_3, \cdots; X'_1, X'_2, X'_3, \cdots$, in which primes indicate derivatives with respect to time. Such double prediction is a first step toward nonlinear prediction.

Automata with unreliable components. In 1956 with the continuing development of faster and more complex computing machines, a realistic study of component misfiring in computers was made. Von Neumann recognized that there was a discrepancy between the theory of automata and the practice of building and operating computing machines because the theory did not take into account the realistic probability of component failure. The number of component parts of a modern all-purpose digital computer was in the mid-20th century already being counted in millions. If a component performing the logical disjunction (A or B) misfired, the total output of a complex operation could be incorrect. The basic problem was then one of probability: whether given a positive number δ and a logical operation to be performed, a corresponding automaton could be constructed from given organs to perform the desired operation and commit an error in the output with probability less than or equal to δ. Affirmative results have been obtained for this problem by mimicking the redundant structure of parallel channels of communication that is frequently found in nature—*i.e.*, rather than having a single line convey a pulse of information, a bundle of lines in parallel are interpreted as conveying a pulse if a sufficient number of members in the bundle do so. Neumann was able to show that with this redundancy technique (multiplexing) "the number of lines deviating from the correctly functioning majorities of their bundles" could with sufficiently high probability be kept below a critical level.

Discrepancy between automata theory and practice

Automata with random elements. The term algorithm has been defined to mean a rule of calculation that guides an intelligent being or a logical mechanism to arrive at numerical or symbolic results. As discussed above under *Neural nets and automata*, a formalization of the intuitive idea of algorithm has led to what is now called an automaton. Thus, a feature of an automaton is the predictability, or the logical certainty, that the same output would be obtained for successive operations of an automaton that is provided with the same input data. If, as a substitute for the usual input data, random numbers (or results due to chance) are provided, the combination of input data and automaton is no longer completely predictable. It is notable, however, that unpredictable results that might be obtained with the use of uncertain input are not without their practical application. Such a method of combining the operation of a computer with the intentional injection of random data is called the "Monte Carlo method" of calculation and in certain instances (such as in the numerical integration of functions in many dimensions) has been found to be more efficient in arriving at correct answers than the purely deterministic methods (see MATHEMATICS AS A CALCULATORY SCIENCE).

Quite apart from the questions of efficiency that might bear upon the addition of an element in a computing machine (automaton) that could produce numbers due to chance, the purely logical question has been asked: "Is there anything that can be done by a machine with a random element that cannot be done by a deterministic machine?" A number of questions of this type have been investigated, but the first clear enunciation and study of such a question was accomplished in 1956 by the U.S. engineer Claude E. Shannon and others. If the random element in the machine is to produce a sequence of digits 0 and 1 in a random order so that the probability is p for a digit 1 to occur, then (assuming that the number p is, itself, obtainable from a Turing machine as a computable number) the machine can do nothing new, so to speak, as compared to the unmodified Turing machine. This result is precisely expressed in the language of automata theory by saying that the sets enumerated by the automaton with random elements can be enumerated also by the unmodified automaton. The computability of p, however, is critical and is necessary for the result. It is also important to emphasize, in order to distinguish this result from what follows, that the computability of p is under discussion, not the computability of the sequence of random digits.

The computability condition for random elements

Computable probability spaces. Finally, it is to be observed that the concept of chance or random number, wherever it has occurred in the above discussion, submits to the interpretation of result of observation of an experiment or physical phenomenon. The chance ingredients in the weather data to which prediction theory applies could be due to molecular disturbances in the atmosphere that are of diverse and minute origin. The chance failure that might cause a component breakdown in a computing machine is due to the physical structure and environment of the defaulting part. The source of chance that could be used to augment the input of a computer for the purposes of the Monte Carlo method of calculation may be chosen as the erratic emission of electrons from the cathode of an electronic tube, and is frequently so chosen.

An entirely distinct question is involved in relating chance and computers. It would be important to know whether an automaton in the sense defined by Turing can generate random numbers. The question is tantamount to asking whether a Turing machine can logically describe the behaviour of those sources of chance that are found in nature and are the subject of the study of probability theory. Because there are many points of view—too many to consider here—more tightly phrased questions may serve as an introduction to the subject, and a few conclusions that can be brought as answers will be mentioned. At the outset, one limited question can be affirmatively answered: Can a computable sequence of numbers, $S = (a_1, a_2, a_3, \cdots)$, serve as the basic ingredient of a probability theory by providing all of the necessary points in a probability space? In this question the term computable sequence is defined to mean that the

numbers a_k are real and there is a Turing machine that, for any pair of positive integers A, B, will print out in order the first A digits of all a_k, for k ranging from 1 to B, in a finite number of steps. The term probability space is precisely defined in the article PROBABILITY, THEORY OF. It might appear that an affirmative answer to the above question is not striking if simple probability theory alone is considered—that is, a theory of events in which the number of possible outcomes is finite, as in the theory of dice, coins, roulette, and the like. On the other hand, it was shown in 1960 that, although a computable sequence can serve as a set of points in a simple probability space, the mathematical expectations of all random variables X defined on the space can be computed according to an explicit algorithm (see 6) that makes use of the sample values, $X(a_1)$, $X(a_2)$, $X(a_3)$, $\cdots$, which themselves are computable if X is computable. In this algorithm it is evident that the potential number of values to be calculated is infinite, though the number of possible outcomes (distinct values of X) might be finite.

Random numbers that a Turing machine can compute

In the language of the limited question considered, a listing of all sample values (random numbers) of an infinite sequence of statistically independent random variables can be printed out by a Turing machine, at least in the simple case, with strict adherence to the definition of all probabilistic terms as based on measure theory, the theory that generalizes the concept of length.

Extension of such constructions beyond the simple case has also been shown to be possible, provided the concept of a random variable can be extended to a class of functions that are more general than the measure–theoretic class. The most explicit formulation of a suitable generalization was given in 1966, and on the basis of this work it is possible to answer affirmatively a second question: For any sequence of probability distributions, is there a sequence of statistically independent random variables with these respective distributions, each of whose sample values can be computed on a Turing machine and whose mathematical expectations are also attainable by algorithm? The implications are discussed in PROBABILITY, THEORY OF: *The additivity assumption.*

Such results would seem to affront the intuition that tends to divide phenomena into deterministic (or computable) and random (or uncomputable) parts. It is to be observed, however, that in probabilistic matters, passage to the limit and infinite collections are essential ingredients, and such entities are unfamiliar objects in the world in which intuitions are formed. (B.R.)

CLASSIFICATION OF AUTOMATA

All automata referred to from this point on may be understood to be essentially Turing machines classified in terms of the number, length, and movement of tapes and of the reading and writing operations used. The term discrete state automaton is sometimes used to emphasize the discrete nature of the internal states. The principal classes are transducers and acceptors. In automata theory, a transducer is an automaton with input and output; any Turing machine for computing a partial recursive function, as previously described, can stand as an example. An acceptor is an automaton without output that, in a special sense, recognizes or accepts words on the machine alphabet. The input of an acceptor is written on tape in the usual way, but the tape is blank at the end of the computation, and acceptance of the input word is represented by a special state called a final state. Thus, a word x, or sequence of symbols from an alphabet denoted by the letter S, is said to be accepted by an acceptor A if A computes, beginning in an initial state q_0 with x on tape, and halts in a final state with tape being entirely blank. A subset designated U of the set of words S^* on an alphabet S is called an accepted set if there is an automaton A that accepts any word $x \in U$.

Acceptors. An elementary result of automata theory is that every recursively enumerable set, or range of a partial recursive function, is an accepted set. In general the acceptors are two-way unbounded tape automata.

A useful classification of acceptors has been introduced in conjunction with a theory of generative grammars de-

(6) $$EX = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} X(a_k)$$

(7)

1st rule:	$\overline{S} \rightarrow \overline{Pr}\ \overline{VP}$
2nd rule:	$\overline{NP} \rightarrow \overline{Art}\ \overline{Adj}\ \overline{N}$
3rd rule:	$\overline{Adj} \rightarrow \overline{Adv}\ \overline{Adj}$
4th rule:	$\overline{VP} \rightarrow \overline{V}\ \overline{NP}$
5th rule:	$\overline{Pr} \rightarrow$ she
6th rule:	$\overline{Art} \rightarrow$ a
7th rule:	$\overline{N} \rightarrow \overline{Adj}\ \overline{N}$
8th rule:	$\overline{Adj} \rightarrow$ pretty
9th rule:	$\overline{Adj} \rightarrow$ little
10th rule:	$\overline{Adv} \rightarrow$ pretty
11th rule:	$\overline{N} \rightarrow$ girl
12th rule:	$\overline{V} \rightarrow$ is

(8)

Step (i):	$\overline{S}$, initial symbol
Step (ii):	$\overline{Pr}\ \overline{VP}$, using 1st rule
Step (iii):	$\overline{Pr}\ \overline{V}\ \overline{NP}$, using 4th rule
Step (iv):	$\overline{Pr}\ \overline{V}\ \overline{Art}\ \overline{Adj}\ \overline{N}$, using 2nd rule
Step (v):	$\overline{Pr}\ \overline{V}\ \overline{Art}\ \overline{Adj}\ \overline{Adj}$ N, using 7th rule
Step (vi):	she $\overline{V}\ \overline{Art}\ \overline{Adj}\ \overline{Adj}\ \overline{N}$, using 5th rule
Step (vii):	she is a pretty little girl, using rules: 12, 6, 8, 9, 11

veloped in the United States by a linguist, Noam Chomsky. A generative grammar is a system of analysis usually identified with linguistics. By its means a language can be viewed as a set of rules, finite in number, that can produce sentences. The use of a generative grammar, in the context of either linguistics or automata theory, is to generate and demarcate the totality of grammatical constructions of a language, natural or automata oriented. A simple grammar for a fragment of English, determined by 12 rules (see 7), can serve to introduce the main ideas.

Generative grammars

In this simple grammar, each rule is of the form $g \rightarrow g'$ (read, "g' replaces g") and has the meaning that g' may be rewritten for g within strings of symbols. The symbol $\overline{S}$ that appears in the rules may be understood as standing for the grammatical category "sentence," $\overline{Pr}$ for "pronoun," $\overline{VP}$ for "verb phrase," $\overline{NP}$ for "noun phrase," and so forth. Symbols marked with a vinculum ($\bar{}$) constitute the set V_N of nonterminal symbols. The English expressions "she," etc., occurring in the rules constitute the set V_T of terminal symbols. $\overline{S}$ is the initial symbol.

Beginning with $\overline{S}$, sentences of English may be derived by applications of the rules. The derivation begins with $\overline{S}$; the first rule allows $\overline{Pr}\ \overline{VP}$ to be rewritten for $\overline{S}$, yielding $\overline{Pr}\ \overline{VP}$; the fourth rule allows $\overline{V}\ \overline{NP}$ to be rewritten for $\overline{VP}$, yielding $\overline{Pr}\ \overline{V}\ \overline{NP}$; and so forth (see 8). A last step yields a terminal string or sentence; it consists solely of elements of the terminal vocabulary V_T. None of the rules apply to it; so no further steps are possible.

The set of sentences thus generated by a grammar is called a language. Aside from trivial examples, grammars generate denumerably infinite languages.

Recursively enumerable grammars and Turing acceptors. An elementary result of automata theory is that every recursively enumerable set, or range of a partial recursive function, is an accepted set. In general, the acceptors are two-way unbounded tape automata. On the other hand, a grammar consisting of rules $g \rightarrow g'$, in which g and g' are arbitrary words of $(V_T \cup V_N)^*$ is an unrestricted rewriting system and any recursively enumerable set of words—*i.e.*, language in the present sense—is generated by some such system. These very general grammars thus correspond to two-way ac-

ceptors, called Turing acceptors, that accept precisely the recursively enumerable sets.

Finite-state grammars and finite-state acceptors. Acceptors that move tape left only, reading symbol by symbol and erasing the while, are the simplest possible, the finite-state acceptors. These automata have exactly the same capability as McCulloch-Pitts automata and accept sets called regular sets. The corresponding grammars in the classification being discussed are the finite-state grammars. In these systems the rules $g \rightarrow g'$ are restricted so that g is a nonterminal v of V_N (as exemplified above) and g' is of the form us, $u \in V_N$ and $s \in V_T$. The languages generated by finite-state grammars, owing to this correspondence, are called regular languages. 09

Comparison with McCulloch-Pitts automata

Although these simple grammars and acceptors are of some interest in information theory and in neural network modelling, they are not descriptively adequate for English or for such standard computer languages as Algol (see COMPUTERS) because they are not able to account for phrase structure. In particular, finite-state grammars cannot generate self-embedded sentences such as "the man the dog bit ran away," nor can they produce sentences with several readings such as "she is a pretty little girl."

Context-free grammars and pushdown acceptors. Context-free, or phrase-structure, grammars, although apparently not affording completely adequate descriptions of vernacular languages, do have the desirable properties just noted. For this family, the rules $g \rightarrow g'$ contain single nonterminals on the left, as in the finite-state grammars, but allow g' to be any word of $(V_T \cup V_N)^*$. The example discussed above is a context-free grammar.

These grammars can account for phrase structure and ambiguity (see 9). For further discussion, see LINGUISTICS; *Methods of synchronic linguistic analysis.*

Pushdown acceptors, which play a key role in computer-programming theory, are automata corresponding to context-free grammars. A pushdown acceptor is a finite-state acceptor equipped with an added two-way storage tape, the so-called pushdown store. At the beginning of operation this tape is blank. As the automaton computes, the store is used to analyze the syntactical structure of the sentence being read. The store moves left when printed, and only the last symbol printed may be read, then the next to the last, and so forth. The input is accepted if both the (one-way) input and storage tapes are blank when the automaton halts in a final state.

The representation of Turing machines in quadruple form may be replaced here by a somewhat clearer list of rules that simulate tape action in their application. Rules can be formulated for a pushdown acceptor P for a context-free language L of items xcx^{-1}, in which x is a word on an abstract alphabet $\{a, b\}$ and x^{-1} is x written in reverse. A first such rule can be formulated to mean that, if P is in state q_0 scanning a on input and any (defined) symbol on the pushdown store, it moves tape left, erases a from the input, prints a on the store, and goes into state q_1. A symbolic expression for the rule might be: $q_0a \rightarrow aq_1$. Another rule might be of the form: if P is in state q_1 scanning c on input and anything on store, it moves input left, erases c, and does nothing with respect to the store—briefly, $q_1c \rightarrow q_2$. Another requires that, if P is in q_2 scanning a on input and a on store, then it moves input left, erases a, moves store right, and erases a (see 10). An example is easily constructed to show that under certain rules a set, say, *abcba* is accepted (see 11). If q_0abcba indicates the outset of a computation with P in the initial state q_0 scanning the first a in *abcba* on input tape and blank on store tape, and if q_2 is a final state, then the computation is determined by the rules given above (see 10). At the end of the computation the automaton is in a final state q_2, both tapes are blank, and there is no rule with q_2 alone on the left; P halts and hence *abcba* is accepted.

Accepted sets in specific calculations

Reflection on the example and on others easily constructed shows that a pushdown acceptor is able, in effect, to parse sentences of context-free languages.

Context-sensitive grammars and linear-bounded acceptors. A fourth type of acceptor, which is mainly of mathematical rather than applied interest, is the two-way acceptor with bounded tape—*i.e.*, tape the length of which never exceeds a linear function of the input length. These are the linear-bounded acceptors. They correspond in the present classificatory scheme to context-sensitive grammars. Unlike the context-free grammars, these latter systems use rules $g \rightarrow g'$, in which the nonterminal symbol $v \in V_N$ in g may be rewritten only in a context xvy; thus $g \rightarrow g'$ is of the form $xvy \rightarrow xwy$, $x, y, w \in (V_T \cup V_N)^*$. An example of a context-sensitive language accepted by a linear-bounded automaton is the copy language xcx.

The family of recursively enumerable languages includes the context-sensitive languages, which in turn includes the context-free, which finally includes the regular, or finite-state, languages. No other hierarchy of corresponding acceptors has been intensively investigated.

Finite transducers. The most important transducers are the finite transducers, or sequential machines, which may be characterized as one-way Turing machines with output. They are the weakest transducers with respect to computing power, while the universal machine (see above *The generalized automaton and Turing's machine*) is the most powerful. There are transducers of intermediate power, but they are not considered here.

Algebraic definition. Because the tape is one-way with output, a finite transducer T may be regarded as a "black box" with input coming in from the right and output being emitted from the left. Hence, T may be taken to be a quintuple $\langle S, Q, O, M, N \rangle$, in which S, Q, and O are finite, nonempty sets of inputs, states, and outputs, respectively, and M is a function on the product $Q \times S$ into Q and N is a function on the same domain into O. The values are written in the usual functional notation $M(q, s)$, and $N(q, s)$, $s \in$ S and $q \in$ Q. M and N may be extended to the domain $Q \times S^*$ by four relations (see 12).

Equivalence and reduction. The most natural classification is by equivalence. If two machines (finite trans-

(9)

At step (v) of the derivation in (8), the 3rd rule of (7) could have been applied to (iv), yielding

(va) $\overline{\text{Pr}}\ \overline{\text{V}}\ \overline{\text{Art}}\ \overline{\text{Adv}}\ \overline{\text{Adj}}\ \overline{\text{N}}$.

Successive steps would again result in (vii), but now to be read with a different meaning than before because of an immediate constituent structure differing from (v). To emphasize this divergence the replacing word g' flanked by parentheses, (g'), may be inserted into a previously obtained word in a derivation. Thus at (ii),

(ii)′ $(\overline{\text{Pr}}\ \overline{\text{VP}})$

would be written; and at (iii),

(iii)′ $(\overline{\text{Pr}}(\overline{\text{V}}\ \overline{\text{NP}}))$.

Proceeding in this way, (v) and (va) would appear as

(v)′ $(\overline{\text{Pr}}\ (\overline{\text{V}}\ (\overline{\text{Art}}\ \overline{\text{Adj}}\ (\overline{\text{Adj}}\ \overline{\text{N}}))))$

and

(va)′ $(\overline{\text{Pr}}\ (\overline{\text{V}}\ (\overline{\text{Art}}\ (\overline{\text{Adv}}\ \overline{\text{Adj}})\ \overline{\text{N}})))$.

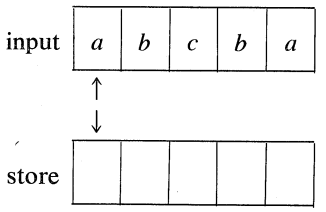
Sentences in which phrase structure is indicated by parentheses are *phrase markers*. Making the further substitutions (vi) and (vii) within (v)′ and (va)′ produces two phrase markers, which represent two readings of (vii).

(10)

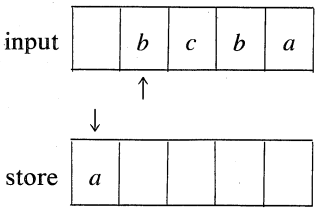
1st rule: $q_0a \rightarrow aq_1$
2nd rule: $q_0b \rightarrow bq_1$
3rd rule: $q_1a \rightarrow aq_1$
4th rule: $q_1b \rightarrow bq_1$
5th rule: $q_1c \rightarrow q_2$
6th rule: $aq_2a \rightarrow q_2$
7th rule: $bq_2b \rightarrow q_2$

(11)

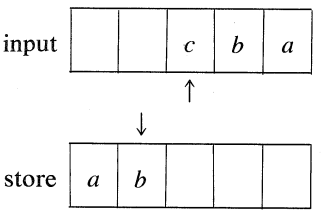
(i) q_0abcba initial tape
This represents the following situation



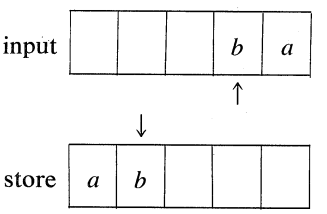
(ii) aq_1bcba by 1st rule (see 10)



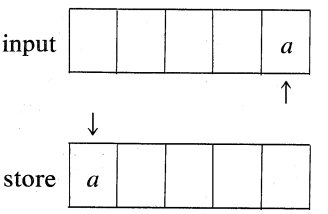
(iii) abq_1cba by 4th rule (see 10)



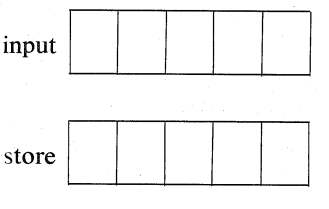
(iv) abq_2ba by 5th rule (see 10)



(v) aq_2a by 7th rule (see 10)



(vi) q_2 by 6th rule (see 10)



(12)

$$
\begin{cases}
M(q, \Lambda) = \Lambda \\
M(q, xs) = M(M(q, x), s) \\
N(q, \Lambda) = \Lambda \\
N(q, xs) = N(M(q, x), s),
\end{cases}
$$

in which the empty word $\Lambda \in S^*$, Λ is adjoined to 0, $x \in S^*$, and $q \in Q$.

ducers) share the same inputs, then representative states from each are equivalent if every sequence x belonging to the set of words on the alphabet causes the same output from the two machines. Two finite transducers are equivalent if for any state of one there is an equivalent state of the other, and conversely. Homomorphisms between transducers can also be defined (see 13). If two automata are onto homomorphic they are equivalent, but not conversely. For automata that are in a certain sense minimal, however, the converse holds.

Homomorphisms between transducers

Each equivalence class of transducers contains a smallest or reduced transducer—that is, one having the property that equivalence between its states implies equality. There is an algorithm for finding the reduced transducer of a class, which proceeds in a natural way from equivalence classes or blocks of states of a given transducer, each such block being defined as a state of the reduced transducer. Reduced equivalent finite transducers are unique up to an isomorphism—that is to say, if two finite transducers are reduced and equivalent, they differ only in the notations for their alphabets.

Classification by semi-groups. A mathematically significant classification of transducers may be obtained in terms of the theory of semi-groups. In outline, if the transducer T is reduced, the functions ϕ_s given in terms of M, for fixed input, as maps from and to the space of states Q constitute a semi-group termed the semi-group of T (see 14). By a certain procedure these semi-groups and their associated transducers T may be decomposed into more elementary systems called serial-connected and parallel-connected transducers. In explanation, the next state (starting from state $\langle q_a, q_b, \rangle$ in the serially connected machine $T_A \rightarrow T_B$ is the pair of states made up of the next state in T_A from q_a with input s and the next state in T_B from q_b with input $N_a(q_a, s)$—which latter is the output of T_a (see 15). Schematically, the connection may be depicted, indicating that in a serial connection the output of T_A is the input to T_B.

The parallel connection of two transducers is a system that may be rigorously defined (see 16) and that may be schematically depicted with input leading in parallel to both machines and output leading in parallel out of both

(13)

Let $\psi: Q_a \rightarrow Q_b$
be a map from the states of T_a into those of T_b. The map ψ is a homomorphism if furthermore,

$$\psi M_a(q, s) = M_b(\psi(q), s)$$

and

$$N_a(q, s) = N_b(\psi(q), s) \text{ for all } s \in S.$$

(14)

$$
\begin{cases}
\phi_s(q) = q' & \text{if and only if} \\
M(q, s) = q' & (q, q' \in Q \text{ of } T)
\end{cases}
$$

(15)

The *serial connection* of two transducers $T_a = \langle S, Q_a, O_a, M_a, N_a \rangle$ and $T_B = \langle O_a, Q_b, O_b, M_b, N_b \rangle$ is the transducer

$$T = T_A \rightarrow T_B = \langle S, Q_a \times Q_b, O_b, M, N \rangle.$$

$Q_a \times Q_b$ is the Cartesian product of states, and M, N are given by

$$M(\langle q_a, q_b \rangle, s) = \langle M_a(q_a, s), M_b(q_b, N_a(q_a, s)) \rangle$$

and

$$N(\langle q_a, q_b \rangle, s) = N_b(q_b, N_a(q_a, s)).$$

(16)

The *parallel connection* of T_A and T_B is the system

$$T = T_A \times T_B$$

in which

$$M(\langle q_a, q_b \rangle, s) = \langle M_a(q_a, s)\, M_b(q_b, s) \rangle$$

and

$$N(\langle q_a, q_b \rangle, s) = \langle N_a(q_a, s), N_b(q_b, s) \rangle.$$

Serial and parallel connections between transducers

machines. It has been shown that any finite transducer whatsoever can be decomposed into a system of series-parallel-connected automata, such that each element is either a two-state automaton or one whose semi-group is a simple group (see ALGEBRAIC STRUCTURES). This affords a classification of machines that depends ultimately on the determination of the simple groups of finite order.

An earlier decomposition scheme was based on a generalization of the concept of congruence relations over sets of states, but discussion of it is omitted here.

Post machines. Types of automata have been investigated that are structurally unlike Turing machines though the same in point of computational capability. The mathematician E.L. Post (U.S.) proposed in 1936 a kind of automaton (or algorithm) that is a finite sequence of pairs $\langle 1, a_1\rangle$, $\langle 2, a_2\rangle$, $\cdots$, $\langle m, a_m\rangle$, such that a_i is either an instruction to move an associated two-way tape one square right or left, an instruction to print a symbol, including a blank from a finite alphabet, or an integer. A Post machine begins at 1 and at step n obeys the instruction a_n and then goes to step $n + 1$, unless a_n is an integer m, in which case it goes to step m if the square scanned at n is marked or to step $n + 1$ if that square is blank. Post machines are prototypes of the program schemes developed 10 years later by von Neumann and his associates. For any partial recursive function a Post machine can be found that is capable of computing it.

Generalizations to automata or information processors in which the restriction to finiteness on sets is dropped or in which additional information from arbitrary sets is available to a machine during computation continue to be considered in the literature. (R.J.Ne.)

BIBLIOGRAPHY. MICHAEL L. MINSKY, *Computation: Finite and Infinite Machines* (1967); and RAYMOND J. NELSON, *Introduction to Automata* (1967), are the most comprehensive elementary introductions. MICHAEL A. ARBIB, *Theories of Abstract Automata* (1969), is an advanced introduction. MARTIN DAVIS, *Computability and Unsolvability* (1958); and HARTLEY ROGERS, JR., *Theory of Recursive Functions and Effective Computability* (1967), are concerned with the concepts of Turing computability and the theory of recursive functions. C.E. SHANNON and JOHN MCCARTHY (eds.), *Automata Studies* (1956), contains some of the original basic material concerning neural nets and automata with unreliable components or with random elements. BAYARD RANKIN (ed.), *Differential Space, Quantum Systems, and Prediction* (1966), discusses the automaton and its environment in the sense of prediction theory and gives reference to other literature in this area as well as the area of computable probability spaces. A good account of automata theory and its relations to switching theory is MICHAEL A. HARRISON, *Introduction to Switching Automata and Theory* (1965). The best introduction to machine decomposition theory is J. HARTMANIS and R.E. STEARNS, *Algebraic Structure Theory of Sequential Machines* (1966). NOAM CHOMSKY, "Formal Properties of Grammars," in R. DUNCAN LUCE, ROBERT R. BUSH, and EUGENE GALANTER (eds.), *Handbook of Mathematical Psychology*, vol. 2 (1963), is still the best survey of the field of automata and generative grammars. Articles presenting approaches to languages and automata from very general mathematical points of view are SEYMOUR GINSBURG and SHEILAH GREIBACH, "Abstract Families of Languages," *Mem. Am. Math. Soc.*, no. 87, pp. 1–32 (1969); and GENE F. ROSE, "Abstract Families of Processors," *J. Comput. & Syst. Sci.*, 4:193–204 (1970).

(B.R./R.J.Ne.)

Automation

The term *automation* was coined in the early 1940s to describe those processes in which mechanisms are used to perform tasks that previously required the attention and control of humans. Since then the term has been applied to a wide variety of automatic machinery and automatic systems, and is commonly used to describe any operation in which there has been a substantial substitution of controlled mechanical, chemical, or electrical action for human effort or intelligence. Thus it became common to call an operation "automated" if it was substantially more automatic than its predecessor.

With the development of computers and the concepts of machine logic and control, a more sophisticated concept has evolved. Automation may be said to be the performance of automatic operations directed by programmed commands combined with automatic measurement of action, feedback, and decision-making control.

HISTORY

The history of automation begins with prehistoric man, the toolmaker. The first simple stone tools represented a conscious extension of man's physical effort under the control of human intelligence. The next advance was the development of simple machines such as the wheel, the lever, and the pulley, by which the power of human muscle could be multiplied. These still required the conscious and complete control of the human mind. The next extension was the development of powered machines. More than 2,000 years ago the Chinese developed trip-hammers powered by flowing water and waterwheels. The early Greeks experimented with simple reaction motors powered by steam. Windmills, with automatic devices for turning the sails, were developed in the Near East and Europe in the Middle Ages. The culmination of these machines, which derive their energy from moving fluids and burning fuels, was the steam engine, elaborated over the past 200 years into engines and machines deriving their energy from chemical, mechanical, and finally nuclear sources.

As man learned to harness energy sources he found that increasingly complex control devices were necessary. The earliest steam engines required a man to open and close valves first to admit steam into the piston chamber and then to exhaust it. Later, a slide valve mechanism was devised, coupled to the piston shaft, which automatically performed these functions. It was then necessary only for the operator to regulate the amount of steam to control the engine's speed and power. Next, James Watt's flying-ball governor took over that function; a weighted ball on a hinged arm, coupled to the output shaft of the engine, was moved outward by centrifugal force as the rotational speed of the shaft increased; this motion controlled a valve that decreased the steam fed to the engine, thus slowing the engine down. Watt's device remains an elegant early example of "negative feedback," by which the increasing output of a mechanism is used to decrease —*i.e.*, negate—the activity of the mechanism, reducing its output.

Negative feedback has proven one of the most effective means of automatically controlling a mechanism to achieve a constant operating level. A simple negative-feedback control is the thermostat used in most home heating systems. In this instrument a rise in room temperature causes a metallic strip to flex, opening a switch that turns off the heating source. As the room cools down the metallic strip flexes in the opposite direction, closing the switch and turning on the heat.

In 1801 J.-M. Jacquard, a French inventor, devised an automatic loom capable of producing complex patterns in textiles by automatically controlling the motions of many shuttles of differently coloured threads. The machine was "programmed" to produce the desired patterns by steel cards in which holes were punched. These cards were the ancestors of the paper cards and tapes that control modern automatic machinery. In the latter part of the 19th century, Charles Babbage (1792–1871), English mathematician, proposed a complex, mechanical "analytical engine" that embodied a primitive decision-making ability.

Building blocks of automation

These four elements—(1) an energy source and its controls, (2) sensing and feedback mechanisms, (3) programming techniques, and (4) decision-making devices—are the building blocks of automation. Elaborated and combined, they have produced machines and systems that perform without human intervention. As early as 1784 Oliver Evans of Philadelphia built an "automatic" flour-making factory in which mechanical devices conveyed the raw material through all of the flour-making steps, including the filling of the flour bags at the end. Though human operators attended the machines to aid certain critical operations, the basic objectives of a continuous automatic process were achieved, making the mill an early example of thorough "mechanization." Its lack of automatic sen-

sing and feedback devices and of decision-making mechanisms, however, remove it substantially from today's automated systems. Indeed, even today, much of the automatic machinery in industry represents a high degree of mechanization rather than complete automation in the sense defined above. To an increasing degree, however, steps are being taken to achieve truly automated operation. The invention and development of the electronic computer has been an essential step.

GENERAL PRINCIPLES OF AUTOMATION

Five components of a system. The five principal components of an automated system are (1) program elements, (2) action elements, (3) sensing elements, (4) decision elements, (5) control elements.

Program elements. Program elements determine *what* the automated system shall do and *how* the parts of the system must function in order to accomplish the desired result. The program may consist of mechanical parts such as gears, cams, and linkages that connect the various portions of a machine together and determine the timing of the various actions of the machine. In the most modern computer-controlled systems the program may be stored in the form of punched paper cards, paper tape, magnetic tape, ferrite cores, electronic circuits, or other mechanical or electronic "memory" devices.

The program can be divided into two parts: the *command program* and the *process program*. The *command program* is a series of instructions that directs the other system elements through a sequential series of steps required to complete the desired operation. The *process program* contains the instructions that tell the system elements how each step of the operation is performed. It may also contain the information that is required for the decision elements described below.

Action elements. Action elements are generally of two kinds: (1) energy application and (2) transfer and positioning. In an automated manufacturing system the transfer and positioning elements control valves and conveyors or otherwise move and align the material being processed into proper position relative to the energy-application elements that shape, heat, chemically treat, or perform other process steps.

Sensing elements. Sensing elements represent one of the principal differences between a mechanized process and an automated process. Their function is to detect and measure a specific property of the processed item and present that measurement in a form upon which the automated system can act. A typical sensing element is the thermocouple, a device capable of measuring temperature and producing an electrical voltage proportional to the temperature. Other sensors may measure a physical dimension, an electrical resistance, the magnetic permeability, the optical properties, or the weight of the processed object. Measurements obtained from the sensors are used to determine if the process is going according to plan and provide necessary corrective changes. The advantages of sensing devices lie in their ability to be absolutely accurate. Fatigue is absent and the speed of operation far exceeds that of maximum human output. Observations can also be made in places which are inaccessible to or unsafe for human beings, such as in a nuclear reactor.

Decision elements. Decision elements use information from sensors that measure how the system is operating and compare these data with information from the process program that describes how the operation should proceed. Based upon this comparison, the decision elements generate instructions, usually in the form of electrical signals, to actuate the control elements. Decision elements are typically electrical circuits, often in an electronic computer, that carry out logical operations in making comparisons and producing command signals.

Control elements. Control elements are the mechanisms by which decisions are carried out. A control element may be a valve that opens and closes on a command signal to add chemical reactant to a process stream. It may be a rheostat that controls electrical power flowing to a heating element or an electric motor. The control elements regulate the action elements described above until the sensors indicate that each step of the process is proceeding according to process program instructions, and that each of the action elements is performing as required by the instructions in the process memory.

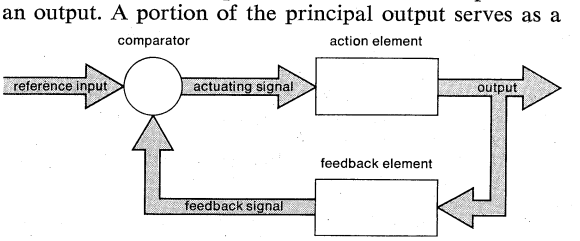
Feedback loops

These five elements provide what is called a closed-loop feedback system, the principal characteristic of modern automation. Analogous closed-loop feedback system controls may be seen in most human functions, for example, the action of reaching for a glass of water. The memory (program) contains a record of the sequence of muscular actions that must be performed and sends control signals along the nervous system that stimulate the arm muscles (action) to start the arm moving in the proper direction. As the hand approaches the glass, the eye (sensing) observes the distance and direction of the hand relative to the glass and feeds back to the brain the progress of the operation. This information is compared with the desired action from memory (program) and if there is a deviation from plan, decision elements in the brain send correction signals to the control elements that cause the muscular action pattern to change. These operations are repeated through many cycles until the glass is finally grasped.

Although there is much about the exact nature and function of the elements in the human that is not understood, the sequence of actions is well defined. The description is helpful in understanding certain diseases of the nervous systems such as palsy, in which simple muscular functions cannot be controlled, because feedback loops do not work properly.

Automation theory. A typical automated system involves a large number of action elements, sensors, and control elements combined in a series of feedback loops and, under the direction of a complex program, combined with sophisticated decision-making equipment. In addition, many of the action elements are interacting so that, for example, a change in one of the process steps may alter the course of other process steps. The principal difficulty in designing an automated system lies in obtaining a complete understanding of each of the action elements and their interactions, and incorporating this information in the program so that proper decision-making routines will be available as required.

In addition to the fundamental understanding of the process elements, a generalized theory has developed to describe behaviour of feedback loops and of systems that contain numbers of interacting elements with feedback. A simple element with feedback is shown below. An actuating signal on the input of the action element produces an output. A portion of the principal output serves as a



Simple element with feedback.

signal on the input to the feedback element. The output of the feedback element is fed to the comparator. A reference input from the process program is also fed to the comparator. If the feedback signal differs from the reference input, the difference produces a new actuating signal, and the new actuating signal acting on the action element will produce a new output. This process will continue until the feedback signal matches the reference input, which will occur only when the output has reached the desired value. Any future change in the reference input or in the output will, through the feedback loop, again cause a change in a direction to drive the output of the action element toward the desired value.

The dynamics of practical feedback systems can be complex. If the system is not properly designed, the feedback action may be too slow to control the action ele-

Results of improper design

ments properly. On the other hand, if the response is too rapid, the system may overcorrect and begin to oscillate. If several feedback elements are interconnected, the dynamics become very complex. A substantial body of theory has developed to aid the system designer in designing automated systems that are both stable and sufficiently responsive to provide adequate control without danger of failure.

Technical and economic advantages and disadvantages. Automation can produce both quantitative and qualitative improvements in human productivity. Automated systems permit a few skilled individuals to obtain results that previously required large numbers of semiskilled and unskilled personnel. In highly automated systems a few skilled people attend to general surveillance of the system, correct for unpredictable breakdowns, maintain process and control elements, and design improvements.

Automated systems also permit the performance of tasks that are effectively beyond human capabilities. The reaction speed of electronic controls and the ability of large data-processing machines to monitor and direct changes at hundreds of control points permit the optimization of complex processes to a degree finer than would be possible with human operators. The launch, tracking, control, and return of space vehicles would be impossible without highly automated systems.

Similarly, in the microminiaturized domain of modern electronic integrated circuitry, and in the inhospitable environment of nuclear reactors, automated systems can perform functions beyond the acuity of humans, in environments in which humans could not survive.

Inflexibility of automation

On the other hand, automated systems lack the flexibility and adaptability of human beings. An automated system is intolerant of error in its design. If incorrectly programmed if will function improperly until corrected by human intervention. Similarly, each element and each step must be completely defined at the beginning. All of the relationships between elements and their interactions must be completely understood and included in the design. Although there have been some experiments with machines that demonstrate a primitive "learning" ability, the practical application of such devices is probably still remote. Some functions that are very simply performed by humans, such as identifying sizes and shapes and aligning two dissimilar objects, are extremely difficult to perform by electrical or mechanical means. Thus, the design and construction of a complex automated system require substantial investment in scientific and engineering effort and in complex process, computation, and control equipment. Automated systems also tend to be relatively inflexible so that significant changes in their desired function may require massive redesigns of equipment. The use of general-purpose digital computers has moderated this problem somewhat in the areas of program and decision elements. Action, sensing, and control elements, however, must still be tailored to specific application.

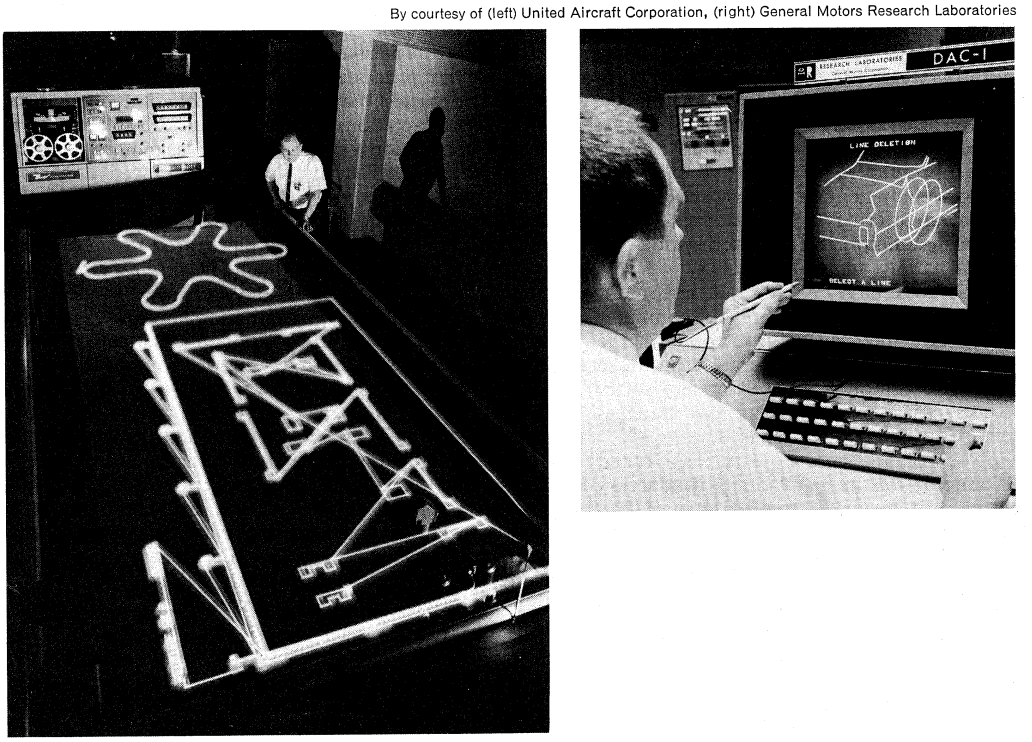
For these reasons the design of an automated system is usually accomplished in stages. First, individual functions are thoroughly investigated and mechanized. Separate stages are automated and their interactions observed and controlled by flexible human intervention. In the final stages increasing portions of the system may be placed under automatic closed-loop control. Human "override" of the automatic functions is provided until the system's basic soundness has been proved and completely automatic control can be instituted.

During these stages there is a careful study of the points of meeting (interfaces) between man and machine. The most effective systems continue to rely upon human action with machine aid in those areas in which computers and machines are inadequate substitutes for human knowledge and flexibility. Thorough understanding of the man–machine interaction remains one of the keys to successful design and operation of automated systems.

Economic requirements are also a key determinant in the evolving applications of automation. The large commitments in research and development and the subsequent large capital investment required to implement an automated system can be a deterrent to application. These large investments, which must often be made years in advance of system use, increase the risks associated with automation.

As an industry becomes more automated, the capital outlays necessary to finance a given expansion increase rapidly. Thus, availability and cost of capital become prime determinants of industry growth rate. Because of

By courtesy of (left) United Aircraft Corporation, (right) General Motors Research Laboratories



Use of automated systems in product design.
(Left) Patterns generated by a computer are traced on an electronically controlled drafting table. (Right) Operator at a graphic console signals a computer about assigned design tasks by touching a "light pen" to a cathode-ray tube display.

these heavy capital requirements, the growth of automation is determined both by technological advances that reduce the cost of automated facilities, and by the availability of large capital funds.

APPLICATION OF AUTOMATION

Automation finds application in almost every area of human endeavour. Although the term initially brings to mind the automatic factory, some of the earliest applications of automation were in communications, especially telephony; important modern examples are found in the utility industries, transportation, distribution, clerical and administrative activities, education, and even in the design of automated systems themselves.

Manufacturing industries. *Chemical processing.* The most highly automated manufacture is found in the chemical process industries. The processing of gases and liquids is especially amenable to automation because of the ease with which materials can be transported between process stations, the relatively simple design of sensors required for measurement and the existence of fundamental understanding of many chemical processes.

Divisions of a typical petrochemical system

In one advanced petrochemical system that manufactures more than 20 different products, the facility is divided into three areas each containing two to six different chemical units and each with a separate process control computer to provide scan, alarm, and control functions. The three process computers are connected to a central computer that calculates how to obtain maximum yield from each process, provides equipment regulation control, and generates management information reports.

Each process computer scans and monitors up to 2,000 process characteristics such as temperature, pressure, flow rate, liquid levels, densities and compositions. Time between measurements varies from 2 to 120 seconds. Each computer controls up to 400 processes. If process parameters exceed specified normal and safe ranges, a computer actuates a signal light and alarm horn and prints an alarm message for the process technician.

The central computer receives information from the process computers and performs calculations to determine the functioning of each chemical unit. The results of these calculations pass to the process computer in the form of commands to change the various processes in such a manner as to optimize unit operations.

In this kind of application, the integrated control system was not instituted as a manpower reduction matter, for a staff of more than 700 personnel is required. But, substantial economic advantages are realized from advanced and unique process control schemes that can be used and that were not possible with conventional control techniques. Increased productivity and process efficiency are obtained by better equipment regulation and optimization. Also, improved safety is achieved because the computers continuously monitor the entire operation and can initiate corrective actions and sound alarms.

Papermaking. In a large papermaking machine, automated for optimum control, a process control computer monitors more than two hundred critical points in the papermaking process. The collected operating data are analyzed; the control operator is advised of any unacceptable conditions. Journals of operating and management information are automatically prepared.

Iron and steel. The iron and steel industry has adopted the principles of automation in many of its production processes. Direct control has been applied to blast furnaces in which iron ore is reduced to pig iron. Automatic instruments measure the pressure and composition of the gases that leave the blast furnace and thereby follow the course of the reduction reaction. This information is analyzed by computer and the results used to control the blast air volume, temperature, humidity, the stoves that fire the furnace, and other variables that affect the efficiency of the process and the quality of the resulting iron.

Many of the chemical operations in the production of iron and steel are dependent on the detailed shape and construction of the blast furnaces, oxygen converters, and other large process units in which they are carried out. Because of this complexity, theoretical analysis of all the process variations and their interaction is often exceedingly difficult. In these cases the tremendous data storage and analysis capabilities of modern computers have been of special importance. Practical automatic control systems operate by collecting the history of previous operations in a specific process unit, and correlating it with the results. By approximations, an empirical model, or series of equations based on this history, is derived that indicates optimum operation of the specific unit.

In a current example the operation of a converter used to produce steel is controlled by using such an empirical model. Eight variables are continuously monitored, including temperature, quantity of molten steel, and concentration of several chemical elements such as carbon, phosphorus, and manganese. In addition, the computer controls the operating sequences including the charging of the converter, provides operating guides, monitors and reports any abnormal operating conditions, and prepares logs of all operating data.

Automated ingot rolling

One of the earliest steelmaking operations in which the principles of automation were applied was the rolling of steel ingots into their final shapes; *e.g.*, coils and strips. In this process the ingot of steel is passed through a rolling mill consisting of one or more sets of large cylindrical rollers that force the ingot into the desired shape. Usually several passes through the rolls are required. In a typical application, automatic instruments measure dimensions and temperature of the steel sheet after each pass through the rolls, and a control computer calculates and directs the optimum settings for the next pass.

Because a large modern rolling mill represents a large capital investment, it is necessary that the equipment be used as efficiently as possible. Automated controls have been developed to pace and schedule the rate at which steel ingots are fed to the rolling mill and the resulting product is withdrawn. Since, in a large operation, several different steel ingots for various customers with various specifications may be in process at any given time, the production control task of scheduling these various orders through the mill at maximum efficiency and keeping track of the different products requires rapid, massive information gathering and analysis while the process is in operation. In the most modern of mills, this production control task has been effectively integrated with the pacing and rolling control operations.

Machining. The shaping of metal by cutting tools was one of the first processes to be mechanized and then automated. The earliest numerically controlled machine tools were direct adaptations of conventional machines. Starting with conventional milling machines, motors were attached to the handles and shafts that moved the work piece. These motors replaced the skilled hands of the machinists and eventually were actuated by a digital controller. Control instructions were prepared in the form of magnetic or punched paper tape. The program information was read from the tape and converted to electrical signals that controlled the actuating motors. Thus the workpiece was programmed through a series of passes under the cutting tool until shaping was completed.

In the earliest machines there was no feedback; it was assumed that the control instructions would move the workpiece to the appropriate position. In later models, automatic correction devices in the form of servomechanisms were used to assure proper positioning. In both cases it was necessary to determine by experimentation in advance the optimum speeds of the cutting head, the feed rate, lubrication feeds, and all other variables to insure acceptable tool wear and finish of the cut surface.

In more recent variations, instrumentation has been developed to permit "adaptive" control of the metal cutting process. In this further step of automation, torque, deflection, and vibration of the cutting head are continuously measured and the cutting speed is varied to optimize the process. In this "adaptive mode" the machine can respond to changes in the properties of the metal and can compensate for tool wear, substantially improving the quality of the finished part and the productivity of the

machine. Cutting machines can now automatically reset the cutting tool.

Numerically controlled machine tools have been developed to produce extremely large pieces, such as wing sections of airliners and turbine rotors of large power generating stations. Once the control tapes for a given part are available, any desired number of identical parts are readily produced, and precision and reproducibility are usually far greater than can be achieved by a human operator. Modern tools can be controlled in five axes of motion: three linear directions and rotation about two perpendicular axes.

Preparing control tapes

Preparing control tapes is still a major problem, even though special computer languages have been developed to simplify the procedure. Even with these aids, however, special skills are required to prepare efficient program control tapes from the designer's sketches of the part to be produced. As automated design aids (see below) are developed, the procedure may be greatly simplified and the effectiveness of numerically controlled, material-shaping machinery greatly increased.

Assembly. In the assembly industries, such as automobile production, automation's initial impact was on isolated areas of production. Thus, there exists a pattern of integrated manufacturing steps carried out by automated equipment, followed by manual operations in which human dexterity and flexibility maintain a substantial advantage over automatic equipment.

Automated assembly lines are still in a relatively primitive stage of development. The design of mechanisms that can manipulate piece parts of various sizes and shapes and align them with adequate precision for assembly has been a challenging and difficult task. The most common solution has been to design specialized machines and standardized piece part shapes especially adapted for automatic handling and assembly. Another approach is to develop robot mechanisms that simulate the articulation and movement of the human arm and hand. The movements of these mechanical arms can be programmed by physically moving the arm through the desired path, and automatically recording the various positions in an electronic memory that can then direct the mechanism to repeat the motion. Such mechanisms are quite flexible and are finding increasing application (see also ROBOT DEVICES).

Robot mechanisms

When mechanized assembly machines in the automotive and electronics industry are associated with transfer mechanisms that carry subassemblies between assembly machines, these later generally require human supervision and operate on an open-loop basis; *i.e.*, without feedback. When feedback is supplied it is often of a primitive type, such as a simple mechanical feeler arm or photoelectric cell to determine whether or not a piece part is in position.

In a few, more sophisticated assembly machines, automatic gauging insures that the assembly or machining operation has been performed with proper precision. In most cases the control action resulting from feedback simply accepts or rejects the assembled part. Again, in some more sophisticated cases, feedback automatically adjusts the assembly or machining mechanisms to improve operations.

An automated assembly operation

An example of a relatively highly automated assembly operation is the manufacture of a clutch assembly for an automatic automobile transmission. The line consists of 20 stations. Subassemblies are mounted on pallets and moved between assembly stations by a chain conveyor. A small computer monitors approximately 650 points in the operation and provides signals to some 200 control elements that adjust the operations as required. The basic control cycle of monitoring performance and providing adjustment is repeated three times a second. When the computer is unable to adjust the station as required, action is stopped and an alarm message is automatically printed out at a control centre. The computer also monitors pneumatic and hydraulic power and the lubrication of the entire machine. Malfunctions are signalled by printing an alarm message, and the computer keeps a complete record of all operating difficulties and also counts the assemblies produced and the rejects.

Electronic equipment manufacture. In the production of electronic equipment, automated systems have been developed for the manufacture of components, the assembly of circuits, and the wiring and testing of complete systems.

In the manufacture of components such as transistors, integrated circuits, resistors, and capacitors, automated subsystems that perform certain process operations are often combined with manual operations, especially for positioning and transfer. Thus a modern component factory is a mixture of automated subsystems and human operators. In some instances fully automated systems have been designed. In one installation, for example, raw materials such as ceramic and metal piece parts and hydrocarbon gases are fed into the line, and fully tested and labelled precision carbon resistors are produced without human intervention.

Automated systems are widely used in the manufacture and test of printed circuits. Program-controlled machines are used for generating circuit patterns for both conventional circuits and integrated circuits. Automated systems are used to assemble components on printed wiring boards and to check that proper components have been used. Testing of completed complex circuits is also performed by automatic equipment often controlled by digital computers that not only test circuit performance but also diagnose and report the source of any malfunction.

Testing completed electronic equipment

In the manufacture of complex electronic systems, such as a telephone switching system or a general purpose computer, there are often tens of thousands of wired connections that must be made and tested to assure against mistakes. After final assembly it is necessary to test the completed equipment and systems by operating them under conditions similar to those that can be expected in the field. Thus, in a telephone switching system, final tests are designed to exercise the equipment under the conditions it will experience when thousands of telephone customers are attempting to call each other simultaneously.

Automated wiring machines have been built to operate under numerical control using a control tape to position a wiring head over the desired terminal, make the connection, lead the unconnected end of the wire through a prescribed path to the second desired terminal, make a final connection, and cut the wire. The machine then makes an electrical test to assure that the proper terminals have been connected before proceeding to the next programmed operation. If a connection is faulty, it is identified on a printout.

Automatic testing machinery has been developed to check equipment that has been manually wired. In these cases the equipment will test each connection against every other connection, assuring that only the proper terminals have been wired together. Again, errors are printed out for manual correction. A piece of equipment containing 1,000 terminals requires 500,000 measurements to assure correct wiring. Such an automatic testing machine can do the task in a few minutes.

Communications. One of the earliest practical applications of mechanization and automation was in telephone switching. The earliest switching machines, invented near the end of the 19th century, were simple mechanical switches that the telephone user controlled remotely by pushing buttons or turning a dial on his telephone.

Automatic switching equipment

The next generations of automatic switching equipment were the electromechanical common control systems that first appeared in the 1920s and 1930s. Constructed of relays and other electromechanical switches, they presaged today's electronic computers. They can monitor thousands of telephone lines, determine which are demanding service, provide dial tone, remember the digits of each number as it is dialed, translate the number to identify the central office and line of the called party, transmit appropriate billing and other data between offices, send ringing current to the called party, set up the required connections, monitor the call during its progress, and disconnect the phones when the call is completed. They thus

perform most of the functions of a human operator. They also can time and bill toll calls, make repeated attempts to find alternate routings of calls if the most direct path is busy or not functioning, run tests of their own operations, and automatically print out trouble reports.

Modern electronic telephone switching machines contain highly sophisticated electronic computing elements that provide major improvements in economy and function over their electromechanical predecessors. In addition to the functions mentioned above, the newest electronic systems automatically transfer calls to alternate numbers, call the user back when a busy line becomes free, and perform other customer services in response to simple dialled codes. These systems also automatically check their function, diagnose difficulties when they arise, and print out detailed instructions for repairmen.

Today's huge-volume telephone service would hardly be possible without automation. Even the operators who handle the services that machines cannot provide, such as directory assistance and person-to-person calls, rely on electronic machines for assistance.

"Store and forward" systems

Automation is employed in other areas of communications such as in automatic "store and forward" systems that accept, store, and deliver a message to one or more destinations. The ability to store and introduce a prescribed delay allows handling of multiple-address messages and improves circuit utilization. An example of such a system is the automatic dispatch of teletypewriter messages. Other communication systems automatically perform functions such as the monitoring, adjusting, and alarm reporting on telephone, radio, and television networks. Satellite communications would not be possible without the automated guidance systems that place and retain satellites in predetermined orbits.

Many national postal systems have mechanized functions such as the grading of letters and parcels by size, orientation of letters for automatic stamp cancellation and, most important, sorting of mail. Completely automated systems to aid in the collection and redistribution of postal communications have yet to be designed.

Power and utilities. The principles of automation have been applied to both the generation and distribution of electric power. These installations have become particularly complex and critical as power systems from several geographic areas have been interconnected to pool the generating capacity of many different utilities. Special attention has been given to automated control systems designed to insure continuous power service and to prevent wide-area "blackouts."

In the early 1970s there was a strong trend toward automated control of power stations, occasionally for direct, on-line control under regular operation and sometimes for control during starting up or shutting down generating capacity. Most such systems supervise and monitor operating functions; they compile data, compare it to normal operating limits, and provide alarm signals to supervisory personnel when normal limits are exceeded.

The start-up of a multimegawatt power generator is a period of major and rapid change both for the generator itself and for the power grid to which it is connected. It is important that the power output of the generator be phased smoothly into the network without overloading or otherwise stressing the generator or the network.

Similar conditions apply when a generator is shut down. In both cases a large number of variables must be monitored and controlled. An automated system can often control the transition, reacting to unexpected events much more effectively than a human operator. In a typical application involving a 300-megawatt coal-fired steam turbogenerator, the automated control system monitors 600 inputs such as temperatures and pressures and switch positions. Each input is scanned repeatedly during the start-up period and controls may be changed as frequently as every second. Critical control signals are displayed for an operator who can override the automatic system if malfunction occurs.

During the normal operation of a power generating station, when conditions are changing much more slowly, automation has been limited to monitoring and alarm functions. In these cases advantage is taken of the ability of a computer controlled system to scan a large number of measurement points, compare the results with predetermined limits, and call to the operator's attention only those parameters which require adjustment. In one example, a monitoring system for a 750-megawatt generating station scans over 1,000 analog and 1,500 digital inputs. Deviations from the desired operating ranges are brought to the operator's attention immediately through cathode-ray tube display devices similar to television screens. They are also printed out on electric typewriters for a permanent record. The operator can call for data on any operating part of the system through a command console. In addition, overall boiler efficiency is constantly monitored, computed, and recorded. Any unwarranted change in temperature is noted and an alarm message is then printed out.

A near-automated electric power distribution system

In the distribution of electric power, automation has been limited primarily to the collection and analysis of data and the presentation of signals to operators who in turn actuate the switching gear or other devices to control and balance network loads. On-line direct digital control of a large power distribution grid has not yet been entrusted to an automated system. However, great advantages are obtained in the monitoring and alarm functions. An excellent example is an automated distribution system designed to include eight hydroelectric generating plants with a total annual output of 6,200,000,000 kilowatt-hours, 21 pumping stations, several storage reservoirs, and a 450-mile aqueduct that connects the water supply for the entire project. Along the 450-mile length the system contains 170 separate electromechanical units such as generators and pumps, 213 separate water control gates, and 49 major water turnouts where the water flow can be diverted for other purposes. The entire system is divided into four control areas, each with its own control centre and independent monitoring computer. In addition, a single master control centre was planned, with a computer to receive data from the four separate control areas and aid the central control operators in integrating the entire system. Again decisions and actions are taken by human control operators on the basis of data and analyses supplied by the automated monitoring systems and electronic computers, but the system has been designed so that completely automated control can be instituted in the future.

These distribution network concepts have application to other utility functions such as gas and oil pipe lines. Here also computer-controlled monitoring and alarm systems have been most effective in aiding the control of many hundreds of miles of large integrated networks.

Transportation. *Lunar missions.* The applications of automation in transportation range from the control of automobile traffic to the guidance and control of lunar landing missions. The most sophisticated applications of automation have been made in aviation and astronautics. Missions such as the Apollo Project would not have been possible without automated control and guidance systems. The launching of a complex and large rocket and the subsequent guidance and staging of the various modules require a coordination of measurement and control signalling well beyond human response time. Beginning with the countdown, the checking of all systems is programmed and monitored by automatic controls, the results compared automatically with predetermined limits, and deviations reported back to the launch team. From the moment of ignition all critical factors are automatically checked and controls adjusted as required. In case of major malfunctions emergency routines are automatically called upon to abort the mission. During the powered phases of the launch the vehicle is continuously tracked by radar and optical means and any deviations from the predetermined trajectory produce automatic control signals that correct the vehicle's flight. During the flight itself there is continuous telemetry of all critical data that again are automatically compared with preselected values with resultant alarms as required. Simi-

larly, the details of lunar landings, return, and re-entry are automatically guided and controlled. Manual overrides are provided for critical malfunctions, but the general system plan represents one of the most sophisticated examples of closed-loop automation.

Aviation. Automation plays several roles in aviation. Major airlines use automated reservation systems to keep continuous track of the status of all flights, compare requests for space with the status of each flight, grant space when available, and automatically update the reservation status files. This permits ticket agents at widely dispersed locations to confirm the availability of space on any flight in a matter of seconds.

Control of air traffic and piloting of aircraft still depend on human control. Although the air traffic controller is greatly aided by radar and other instruments he himself determines traffic patterns and airplane spacings and gives verbal commands by radio to the pilots.

Automatic pilots and navigation systems

Most commercial aircraft are equipped with automatic pilots that under normal flying circumstances can guide the craft over a predetermined route by detecting changes in the aircraft's orientation and heading from gyroscopes or similar instruments, and provide control signals to the aircraft's steering mechanisms to keep it on the desired course. Automatic navigation systems and instrument landing systems use radio signals from ground beacons and provide signals to the pilot to aid him in following a proper course. In general, though, when the aircraft is within the traffic pattern for ground control, the pilot assumes total control.

Flight simulators, used for aircrew training, are now used widely. A cockpit with controls and instruments is provided and the pilot's movements are transmitted to a computer that displays the results on instruments. Necessary adjustments are made by the pilot. Different problems can be fed into the computer and then presented to the crew for a solution. The simulator is a money-saving device for the airlines, having reduced actual training time by 25 percent.

Although automatic landing systems have been developed and experimentally tested, they are not generally applied in commercial aviation. As the reliability of these systems improves, the time can be foreseen when much of the traffic and aircraft control is automated with human supervision and human intervention only in case of system malfunction.

Railroads. Automation is only now beginning to have a significant impact on ground transportation. A series of successful tests of new automatic controls applied to a moving train have been reported from the U.S.S.R. The equipment is in essence an electronic analog computer designed to give optimal control of an electric locomotive running on a given track and following a given timetable. It limits the action of the driver to simple on and off operations of the automatic equipment during stops at the stations; he takes full control of the engine only in an emergency. The control commands are continuously fed to a series of servomechanisms in response to data gathered by sensing elements placed in strategic positions on the engine and track. The system operates the train with greater precision and economy than even the most experienced driver; results show a 7 percent saving in power consumption and a 15 percent increase in traffic. Similar systems have met with success in other countries, especially Japan, where the high-speed New Tokaido Line is highly automated.

A major problem for the railroads is keeping track of freight shipments and empty rolling stock over such vast and international areas as western Europe and North America. In North America an information network and a system of automatic car identification by colour code was under development in the early 1970s. The system will maintain an inventory of the over 2,000,000 Mexican, United States, and Canadian rail cars and notify each of the 65 participating railroads of the interchange of their rolling stock.

The management of rail yards is being substantially aided by automated systems that integrate the signalling and communication functions in the classification yard. Beginning with records of the car complement of an incoming train, the computer-based system automatically controls switches in the yard to sort the cars onto tracks with other cars with a common destination. As each outgoing string of cars is assembled, a record of the cars and their weight is automatically prepared for use at the next destination.

Rail yard management

BART system. One of the most ambitious automated rail transportation systems was planned for the Bay Area Rapid Transit (BART) system in the San Francisco, California, area. The system consists of over 75 miles of track with 105 trains operating at peak hours between 33 stations. The trains can attain speeds of 80 miles per hour with spacings as short as 90 seconds. Each train carries one operator whose role is that of an observer and communicator able to override the automatic system in an emergency. The design includes one subsystem, to protect trains by assuring a safe distance between them and by controlling their speed, and a second to control routing and adjust the performance of each train to keep the entire system operating on schedule.

Limited service began in 1972. Under full operation, as a train enters the station it will automatically transmit to a receiver its identification, destination, and length, thus lighting up a display board for passenger information and transmitting signals to the control centres. Signals will be automatically returned to the train to control its time in the station and its running time to the next station. At the beginning of each day an ideal schedule will be determined, and as the day progresses the performance of each train will be compared with the schedule, and adjustments to individual train performance made as required. If unexpected situations arise, such as train breakdown, the system can automatically adjust itself to minimize the effect. The entire system is controlled by two identical computers so that if one malfunctions the other assumes control while the first is under repair. Should the automated control fail, manual controls can keep traffic flowing.

Traffic control. Although there is much experimentation with automated highways for automobiles, economically practical systems have not yet been designed. The principles of automation have, however, been applied to traffic signal control for several years. These range from signal controllers that respond to an automobile crossing a treadle to lines of traffic lights whose interrelated timing is controlled from a remote traffic-control centre. Essentially every large city has automatic systems for remotely monitoring and controlling traffic flows. In most cases, changes in signal patterns are under the direct control of human traffic directors at the control centres.

Systems are being developed, however, in which decision and control will be determined by computer. In one of these, 77 intersections with traffic signals are under automated control; there are 26 primary sensor points at which traffic flow is continuously and automatically measured, and a central computer analyzes the data and selects a control program from a repertoire of signal control patterns stored in the computer. Commands are sent automatically to the traffic light controllers at each of the intersections to produce whatever changes are required.

Experience with the system has brought a 16 percent reduction in the number of stops that a vehicle must make, a 30 percent reduction in the average delay experienced by each car, and an 8 percent decrease in accidents. The average peak hour speed on the approaches to and from the downtown area has increased from about 20 miles an hour to greater than 30 miles an hour. It is estimated that the system is saving motorists approximately 19,500 hours of travel time annually. Advanced traffic-control systems also exist in Munich and Toronto.

Ships. Automation has also found application in ships, principally in the form of monitoring systems that measure the performance of boilers, turbines, diesels, etc. and provide operating data and alarms to engine officers and pilots. These systems permit more accurate and efficient control of the vessels as well as reduction in crews.

Management, services, research, and education. *Management information.* The principal uses of electronic computers in industry today are in information systems, which range from the simple mechanization of clerical operations, such as the computation and printing of payroll checks, to highly integrated systems that collect, analyze, and report management information required to operate a major plant or industrial complex. In the latter cases the systems may provide long-range business forecasts, receive customer orders over data communication links, schedule the orders through the plant, automatically generate orders for raw materials, keep complete track of orders as they progress through various stages of manufacture, maintain inventories, prepare priority lists for rush orders, prepare bills and invoices, and issue detailed shipping instructions. Some advanced systems include on-line data collection stations dispersed throughout the manufacturing plant so that the completion of manufacturing operations, both by people and by production machines, may be recorded as they happen. The memory banks of the computers contain information that describes the complete status of the manufacturing plant. It can compare the actual status with detailed schedules and plans and inform management when there are deviations from plan. There are very few modern manufacturing plants that do not use some kind of computer-based mechanized or automated information system for inventory and production control, although only a few have highly developed management-information systems.

Types of automated management operations

Automated warehouses. Automated warehouses are becoming increasingly common. As merchandise arrives at the warehouse, a vacant location is assigned by a computer that generates a punched card or similar record associated with the merchandise. The punched card can be used with mechanized handling equipment that automatically transports the material to the assigned location. When the item is needed the computer determines its location and presents a card that is used by the automatic materials-handling equipment to retrieve the item. The computer also updates its files to maintain an accurate inventory list and prints out notices or orders if inventories fall below prescribed limits. The computer maintains records of stock activity and changes stock levels to provide desired service levels.

Driverless trucks are a popular addition to the new automated warehouses. Trucks are battery-electric-powered and controlled to follow a continuous wire embedded in the floor. The trucks are then able to make scheduled stops around the warehouse and the routes can be easily changed. Safety devices, designed to stop the trucks automatically if they contact another object, have been devised.

Banks and financial institutions. Automated systems are used in many other areas of service that require the manipulation and analysis of data. The sorting of checks and verification of bank balances have been mechanized in most financial institutions. The stock exchanges use automatic systems to report stock transactions by ticker tape or closed circuit television. Stock certificates may be issued in the form of punched paper cards to facilitate record keeping in sales and exchanges. Restaurants, retailers, and credit-card organizations are using systems that automatically check the validity of a credit card and the credit standing of its holder in a matter of seconds while the customer waits.

Retail trade. Automation in retail trade is mainly concerned with inventory control or stock balances in stores. Valuable information can be obtained, such as analyses of sales and stock positions. The system can work at night, so that all information can be obtained the next morning. Information can also be got using special tapes that are printed or punched.

Medicine. Hospitals throughout the world are increasing their use of automation; these systems replace labour and sometimes are capable of surpassing human effort. For example, in hospitals in Sweden, several measurements such as heartbeat, pulse rate, blood sugar level, etc. are taken simultaneously every hour from seriously ill patients in intensive care wards. At any time, the doctor can request a complete printed record of his patient's hourly progress. The system also provides continuous monitoring, sounding an alarm should any critical measurement pass beyond the limits of safety. Even a large team of medical personnel using conventional methods could not provide such precise monitoring.

Automated laboratories

Automated hospital laboratories can perform ten different tests on a blood sample in the time it takes to do one manually. In one pilot operation, doctors who requested a specific test received the results of ten at no additional charge. These data caused changes in half the original diagnoses and led to widespread adoption of the system in Europe and America.

Libraries. Library information-retrieval systems are available that automatically search titles and abstracts contained in computer memory and present a bibliography of sources that appears to match a requested subject or combination of subject areas. The lack of low-cost, large capacity data-storage files with random access at electronic speeds has been a primary limitation. Substantial improvements in the cost–performance characteristics of data storage and retrieval devices and systems would increase the attractiveness of automation in this area.

Research. In the research laboratory, automated instruments are used to analyze chemical composition and determine crystal structure. Other systems automatically collect data from instruments, analyze the data, instruct the instrument to take the next series of readings, and present calculated results to the researcher.

Teaching machines. Automated teaching aids are expected to improve the teachers' ability to give large numbers of students more individualized attention. In one experimental system a student is presented a series of questions and he indicates the answers through a typewriter keyboard. Each succeeding question is automatically chosen depending upon his answers to the preceding questions. In this way areas in which his knowledge is weak can be explored and appropriate instructional material presented. Thus, each student, with the help of the machine, can study at his own pace, calling on a teacher if he reaches difficulties that the machine cannot help him resolve.

Automated design. Automated systems based on the data-handling ability of electronic computers are being used as aids to the designers of products, machinery, and even other automated systems. In their simplest and currently most widely used forms, these serve as mechanized aids to drafting. Thus a designer may use a "light pen" and sketch the details of a part on the face of a cathode-ray tube display. The automatic drafting system can convert the designer's rough lines and circles into precise lines and curves of specified mathematical shape. As the designer completes his sketch, information describing the details of his drawing is recorded in digital form in the drafting system's electronic memory. This information can later be used to drive an automatic drafting machine that will make a paper-and-ink copy of the design. The information can also be processed to produce a tape to drive a numerically controlled instrument to machine a part or produce a printed circuit pattern. In some sophisticated systems still under active development, especially in the automotive industry, the designer can sketch several plan views of his conception. The automated design system combines these views into a three-dimensional image that can be displayed in any desired aspect or rotated in perspective on the cathode-ray display. Similarly, the information can be used to drive a numerically controlled machine that will produce a three-dimensional model to any desired scale.

In more sophisticated automated design systems a more substantial part of the designer's task can be performed for him; rules (*e.g.*, the means for partitioning a circuit into modules, the permitted placement of electronic components, and the wire routing paths in a complex circuit) can be fed into the system. Then the designer need only state the basic electronic function that he desires and the

system will automatically determine the required modules and select the required components and their interconnection paths. The system may also generate graphic plots of the response of the described circuit to various inputs, thus enabling the designer to determine if the circuit design is satisfactory. The system may then provide manufacturing information such as parts lists, printed circuitry patterns, and component insertion programs as well as information to control automatic test sets to measure the finished product.

Though such automated design systems are still in their earliest state of development, they suggest the potential of a technology in which machines aid in the design as well as manufacture of other machines. The image of a self-generating and perpetuating machine technology has been a source of stimulation for science-fiction writers as well as pioneering students of cybernetics. Although still remote in time, the basic elements and concepts for their development are rapidly becoming available.

AUTOMATION AND SOCIETY

Stages of industrial development. Automation is the third stage in the development of industrial society.

In the first stage, the age of craftsmanship, each individual craftsman was responsible for a total production process from raw material to finished product. Each manufactured item differed from similar products, because each reflected the skill of its maker. The volume and cost of production tended to be directly proportional to the number of craftsmen.

The second stage, the period of mechanization, produced a complete reorganization of the methods of production. The period is characterized by the analytical study of the production process and the development of scientific management in the early years of the 20th century. Analysis led to the subdivision of the production process into a number of well-defined steps and the design of specialized machines to aid in the performance of each step. Workers were carefully but narrowly trained to operate their own tools and machines with great efficiency. This led to the economies of scale associated with mass production, by which the volume of production increased far more rapidly than the number of employees, and the unit costs of production decreased.

The third stage, automation, carries many of the characteristics of mechanization a large step further and introduces qualitatively new characteristics as well. A much broader and deeper base of engineering planning and design is required. The nature and character of the product are completely determined during the course of this planning. Capital investment is substantially larger. The volume of product is only loosely related to the number of productive employees. The cost of production is strongly dependent upon the cost of raw materials, capital investment, and equipment maintenance, rather than on the cost of labour. Productive employees no longer enter directly into the production process. Instead, the individual monitors and maintains machines, and takes control only in case of severe malfunction.

Effects on the individual. Each of these stages of industrial development has had profound effects on the role of the individual, his skill level, and his productivity. During the craft period each craftsman's skill was relatively unique. The value of his product was closely related to his skill level and the volume of individual productivity was relatively constant with time. The capital investment required for production was relatively small and business was organized on an individual or family level.

With the beginning of the Industrial Revolution and mechanization, production per employee began to rise rapidly. At any given time production was generally proportional to employment, but the ratio changed rapidly as a function of time, with labour content decreasing as productivity increased. Mechanization also created a demand for a broader range of skills including the highly skilled machine-design engineer, the machine builders and toolmakers, skilled maintenance technicians, and a wide range of skilled machine operators. The capital investment required for efficient production increased rapidly. This factor and the production economies that mechanization made possible led to the growth of business enterprises. In turn, complex organizational structures were developed to coordinate the highly subdivided and specialized parts of the production process. The management skills required to run large organizations became increasingly important.

As automation becomes more important the relationships between production and employment become very complex. The man-hours of labour per unit of production diminish while the level of skills and the responsibilities of employees increase. There is a large increase in the level of engineering required and in capital investment. This increases the growth tendencies of businesses and accentuates the requirements for close coordination and planning.

Effects on society. Mechanization and the Industrial Revolution produced profound changes in society. The growth of large industries stimulated migration from country to city. The ability to use narrowly trained and relatively low skill levels and the concentration of employees in large centres led first to cases of labour abuses and then to labour laws and unionization.

Massive increases in productivity stimulated population growth and also created opportunities for expansion of education, of the arts and sciences, and of leisure time. Increased productivity must be coupled with increased consumption and this, in turn, accentuated man's effect on the environment, depleting some resources and abusing others.

Automation and unemployment. During the early periods of the Industrial Revolution there was great concern about massive unemployment as machines took over the jobs formerly performed by human muscle. The feared unemployment, as a continuing factor, however, never materialized. Population growth, human tastes, and human consumption have kept up with increased productivity, and except for brief dislocations, the increased productivity that has resulted from mechanization has created many more jobs than it has destroyed.

In the 1960s automation prompted similar questions. Most studies were controversial and generally inconclusive, but as with the earlier mechanization, widescale unemployment did not materialize. Productivity continued to increase but consumer desires equalled or surpassed the increases. Some effects of automation on employment have, however, become more evident. There have been significant dislocations. In mining regions automation has produced a net decrease in local employment. More widely, increases in consumer demand have outpaced increases in productivity resulting from automation, leading to an increase in employment and a change in its nature. Routine repetitive tasks have been replaced by more broadly skilled jobs of designing, installing, maintaining, and monitoring the automated systems. The need for a higher ratio of the newer skills has obvious implications for education of the young and re-education of the displaced.

Finally, the increased interest in the quality of living in the early 1970s has significance in respect to automation. Automation will be needed to monitor and control new productive systems that will relieve the burden on the physical environment as well as in systems designed to cleanse and restore the environment.

BIBLIOGRAPHY. Much has been written in the popular and lay press about the principles of automation and its specific applications. General discussions in this vein include: G.H. and P. AMBER, *Anatomy of Automation* (1962); A.D. BOOTH, *Automation and Computing*, 2nd ed. (1966); D.B. FOSTER, *Automation in Practice* (1968); and J. ROSE, *Automation: Its Uses and Consequences* (1964), *Automation: Its Anatomy and Physiology* (1967).

Detailed articles on automated systems are found throughout the technical literature. Review articles are appearing in increasing numbers. References relating to two specific industries are: T. ISOBE, "Automatic Control in the Iron and Steel Industry," *Automatica*, 6:111–121 (1970); and G. QUAZZA,

"Automatic Control in Electric Power Systems," *ibid.*, pp. 123–150.

Broad discussions of the impact of automation on industrial management, workers, and the social structure are found in: J.R. BRIGHT, *Automation and Management* (1958); W.S. BUCKINGHAM, *Automation: Its Impact on Business and People* (1961); J. DIEBOLD, *Automation: Its Impact on Business and Labor* (1959), *Man and the Computer* (1969); W.A. FAUNCE, *Problems of an Industrial Society* (1968); G.W. TERBORGH, *The Automation Hysteria* (1966); and the NATIONAL COMMISSION ON TECHNOLOGY, *Automation and Economic Progress* (1966).

(M.T.)

Automobile

An automobile is a self-propelled passenger vehicle designed to be operated on ordinary roads. In general, motor vehicles may be classified into several types, according to the service they render. Most numerous are the private passenger cars described in this article, intended usually to transport up to nine individuals; a subdivision of this type is formed by the so-called sports cars, which are designed primarily for speed and good road-handling characteristics, with passenger comfort a secondary consideration. Commercial passenger vehicles for carrying larger numbers of persons—buses and motor coaches—and commercial freight vehicles for carrying goods and materials are described in the article TRUCKS AND BUSES. The development of the relevant manufacturing technology is covered in AUTOMOTIVE INDUSTRY.

This article is divided into the following sections:

I. History
  Early developments
    Early experiments
    The first automobile
    The age of steam
    Benz and the gasoline car
    Development of the Daimler
    Early efforts in the United States
  Recent developments
    Electric power
    Ford and the automotive revolution
    Age of the classic cars
    Developments after World War II
II. Modern automobiles
  Automotive systems
    Bodies
    Chassis
    Engines
    Fuel and lubrication
    Cooling system
    Electrical system
    Transmission
    Other mechanical subsystems
    Tires
  Problems of automotive design
    Safety considerations
    Pollution problems
    Future systems

I. History

EARLY DEVELOPMENTS

Unlike many other major inventions, the original idea of the automobile cannot be attributed to an individual, and many individuals worked simultaneously on self-powered road vehicles. Certainly, the idea occurred long before it was first recorded in the *Iliad*, in which Homer (in Alexander Pope's translation) states that Vulcan in a single day made 20 tricycles, which

> Wondrous to tell instinct with spirit roll'ed
> From place to place, around the blest abodes,
> Self-moved, obedient to the beck of gods.

Early experiments. Leonardo da Vinci considered the idea of a self-propelled vehicle. In 1760 a Swiss clergyman, J.H. Genevois, suggested mounting small windmills on a cartlike vehicle, their power to be used to wind springs that would move the road wheel. Genevois's idea probably derived from a windmill cart of about 1714. Two-masted wind carriages were running in the Netherlands in 1600, and a speed of 20 miles (30 kilometres) per hour with a load of 28 passengers was claimed for at least one of them. The first recorded suggestion of wind use was probably Robert Valturio's unrealized plan (1472) for a cart powered by windmills geared to the wheels.

Other inventors considered the possibilities of clockwork. Probably in 1748, a carriage propelled by a large clockwork engine was demonstrated in Paris by the versatile inventor Jacques de Vaucanson.

The air engine is thought to have originated with a 17th-century German physicist, Otto Von Guericke. Guericke invented an air pump and was probably the first to make metal pistons, cylinders, and connecting rods: the basic components of the reciprocating engine. In the 17th century, a Dutch inventor, Christiaan Huygens, produced an engine that worked by air pressure developed by explosion of a powder charge. Denis Papin of France built a model engine on the vacuum principle, using the condensation of steam to produce the vacuum. An air engine was patented in England in 1799, and a grid of compressor stations was proposed to service vehicles. An air-powered vehicle is said to have been produced in 1832.

As early as the 16th century, steam propulsion was proposed, and in 1678, Ferdinand Verbiest, a Belgian Jesuit missionary to China, made a model steam carriage based on a principle suggestive of the modern turbine. Another early proponent of steam power encountered such skepticism, however, that he was certified insane.

In the 18th century a French scientist, Philippe Lebon, patented a coal-gas engine and made the first suggestion of electrical ignition. In Paris, one Isaac de Rivas made a gas-powered vehicle which he patented in 1807; his engine used hydrogen gas as fuel, the valves and ignition were operated by hand, and the timing problem appears to have been difficult.

The first automobile. The Royal Automobile Club of Great Britain and the Automobile Club de France now agree that Nicolas-Joseph Cugnot of Lorraine was the constructor of the first true automobile. Cugnot's vehicle

By courtesy of the Montagu Motor Museum, Beaulieu, Hampshire

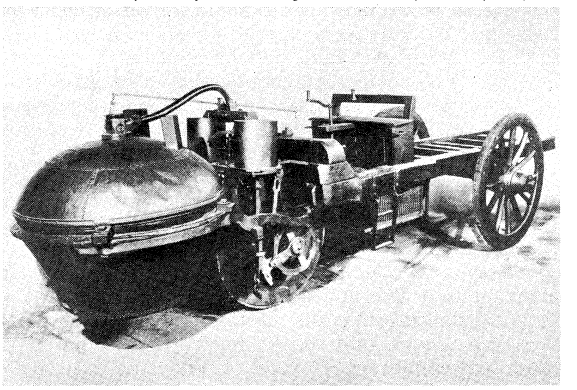


Figure 1: The 1769 Cugnot steamer, a three-wheeled artillery tractor with front-wheel drive, carried four passengers at a speed of 2¼ miles per hour.

Cugnot's steam tricycle

was a huge, heavy, steam-powered tricycle, and his model of 1769 was said to have run for 20 minutes at 2.25 miles (3.6 kilometres) per hour while carrying four people and to have recuperated sufficient steam power to move again after standing for 20 minutes. Cugnot was an artillery officer, and the more or less steam-tight pistons of his engine were made possible by the invention of a drill that accurately machined cannon bores. A replica of the Cugnot vehicle, partially original, is preserved in the Conservatoire des Arts et Métiers in Paris.

Most authorities are inclined to honour Carl Benz and Gottlieb Daimler of Germany as the most important pioneer contributors to the gasoline-engine automobile. Benz ran his first car in 1885, Daimler his in 1886. Although there is no reason to believe that Benz had ever seen a motor vehicle before he made his own, he and Daimler had been preceded by a Frenchman, J.-J.-E. Lenoir, and an Austrian, Siegfried Marcus, in 1862 and 1864–65, respectively, but neither Lenoir nor Marcus had persisted. Benz and Daimler did persist—indeed, to such purpose that the firm of Daimler–Benz could, in the

1960s, trace its origins back to 1895 and claim, with the Panhard firm of France, to be one of the oldest automobile-manufacturing firms in the world. Oddly, Benz and Daimler never met.

The four-stroke principle upon which most modern automobile engines work was discovered by a French engineer, Alphonse Beau de Rochas, in 1862, the year in which Lenoir ran his car from Paris to Joinville-le-Pont. The four-stroke cycle is often called the Otto cycle, after the German Nikolaus August Otto, who designed an engine on that principle in 1876. De Rochas held prior patents, however, and litigation in the French courts upheld him. Lenoir's engine omitted the compression stroke of the Otto cycle; fuel was drawn into the cylinder on the intake stroke and fired by a spark halfway on the next reciprocal stroke.

The idea for Siegfried Marcus' 1864–65 car apparently came to him by chance while he was considering the production of illumination by igniting a mixture of gasoline and air with a stream of sparks. The reaction was so violent that it occurred to him to use it as a power source. His first vehicle was a cart carrying a two-cycle engine geared to the rear wheels without any intervening clutch. It was started by having a strong man lift the rear end while the wheels were spun. It did run, however, a distance of about 200 yards. In 1898 the Austrian Automobile Club arranged an exhibition of motor cars, and Marcus was a guest of honour. He denied interest in the entire idea of the automobile, however, calling it "a senseless waste of time and effort." Marcus' second model, the 1874–75 car, was sturdy and sufficiently well-preserved to make a demonstration run in the streets of Vienna on April 16, 1950, at a rate of three miles per hour.

Other inventors have been put forward as originators of the internal-combustion automobile. In 1823 or 1826 an English mechanical engineer named Samuel Brown drove a self-powered road vehicle up Shooter's Hill in London. Brown's patent specifications and an eyewitness account of his drive both exist. His engine had separate combustion and working cylinders and apparently used carbureted hydrogen as fuel. Léon-Paul-Charles Malandin of France produced a four-wheel vehicle with a four-stroke, two-cylinder gasoline engine in 1883, and he, too, is sometimes cited as the first to run an automobile with a true internal-combustion engine. His claim, like the others, falters before the continuity of the Benz–Daimler efforts.

The age of steam. Before any internal-combustion engine had run, followers of Nicholas Cugnot were on the road, notably in England, although the first post-Cugnot steam carriage appears to have been that built in Amiens, France, in 1790. Steam buses were running in Paris about 1800. Oliver Evans of Philadelphia ran an amphibious steam dredge through the streets of that city in 1805. Less well known were Nathan Read of Salem, Massachusetts, and Apollo Kinsley of Hartford, Connecticut, both of whom ran steam vehicles during the period 1790–1800.

English inventors appear to have been active, and by the 1830s the manufacture and use of steam road carriages approached the status of a minor industry in the British Isles. James Watt's foreman, William Murdock, ran a model steam carriage on the roads of Cornwall in 1784, and Robert Fourness showed a working three-cylinder tractor in 1788. Watt was opposed to the use of steam engines for such purposes; his feelings were so strong that he inserted into the lease of a house he owned a clause stating that "no steam carriage should on any pretense be allowed to approach the house." Watt's low-pressure steam engine would have been too bulky for road use in any case, and all of the British efforts in steam derived from the earlier researches of Thomas Savery and Thomas Newcomen.

Richard Trevithick developed Murdock's ideas, and at least one of his carriages, with driving wheels ten feet (three metres) in diameter, ran in London. Sir Goldsworthy Gurney, the first commercially successful steam-carriage builder, based his design upon an unusually efficient boiler. He could not, however, be convinced that smooth wheels could grip a roadway, and so he arranged propulsion on his first vehicle by iron legs digging into the road surface. His second vehicle was said to weigh only 3,000 pounds (1,400 kilograms) and to be capable of carrying six persons. He made trips as long as 84 miles (135 kilometres) in a running time of nine hours 30 minutes and once recorded a speed of 17 miles (27 kilometres) per hour. A rate of 32 miles (51 kilometres) per hour was later claimed by two other builders. To allay the public's fear of boiler explosions, Gurney designed a "drag," which was in effect a tractor, or locomotive, pulling one or more unpowered coaches.

Steam coaches on British roads

Gurney equipment was used on a regularly scheduled Gloucester–Cheltenham service of four round trips daily that at times did the nine miles (14 kilometres) in 45 minutes. Between February 27 and June 22, 1831, steam coaches ran 4,000 miles (6,400 kilometres) on this route, carrying some 3,000 passengers. The equipment was noisy, smoky, destructive of roadways, and admittedly dangerous; hostility arose, and it was common for drivers to find the way blocked with heaps of stones or felled trees. A minor accident occurred and finally aroused so much public feeling that the passenger service was ended. Nevertheless, many passengers had been carried by steam carriage before the railways had accepted their first paying passenger.

The most successful era of the steam coaches lasted from 1831 to 1838. Ambitious routes were run, including one from London to Cambridge. But by 1840 it was clear that the steam carriages had little future. They had had much to contend with, including the anti-machinery attitude of the public and the enmity of the horse-coach interests, which resulted in such penalties as a charge of £5 for passing a tollgate that cost a horse coach only 3*d*. The government was of small help: a select committee of Parliament recommended a research grant of £16,000 for Goldsworthy Gurney, but the treasury refused to pay it. The crushing blow was the Locomotives on Highways Act of 1865, which reduced permissible speeds on public roads to two miles (three kilometres) per hour within cities and four miles (six kilometres) per hour in rural areas. This legislation was known as the Red Flag Act because of its requirement that every steam carriage mount a crew of three, one to precede it carrying a red flag of warning. The act was amended in 1878, but it was not repealed until 1896, by which time its provisions had effectively stifled the development of road transport in the British Isles.

Inventors in other countries also experimented with steam carriages, and a few ran on roads in Europe and the United States.

The decline of the steam carriage did not prevent continued effort in the field, and much attention was given to the steam tractor for use as a prime mover. Beginning in about 1868, Britain was the scene of a vogue for light steam-powered personal carriages; if the popularity of these vehicles had not been legally hindered, it would certainly have resulted in the appearance of widespread enthusiasm for motoring in the 1860s rather than in the 1890s. Some of the steamers were designed to carry as few as two people and were capable of speeds on the order of 20 miles (30 kilometres) per hour. The public climate remained unfriendly, however.

Light steam cars were being built in the United States, France, Germany, and Denmark during the same period, and it is possible to argue that the line from Cugnot's lumbering vehicle runs unbroken to the 20th-century steam automobiles made as late as 1926. The grip of the steam automobile on the American imagination has been strong ever since the era of the Stanley brothers (one of whose "steamers" took the world speed record at 127.66 miles [205.4 kilometres] per hour in 1906), and in the 1960s it was estimated that there were still 7,000 steam cars in the United States, about 1,000 of them in running order. Though the pollution crisis of the 1960s inspired a resurgence in steam-car research, the major Detroit companies were unwilling to finance a crash program, and of the individuals who attempted construction only

The Stanley steamer

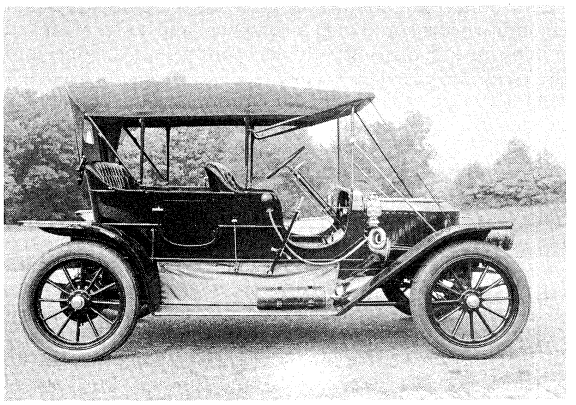


Figure 2: Stanley, 1911, a four-passenger touring car.
By courtesy of the Magic Age of Steam, Yorklyn, Delaware

the industrialist William Lear maintained a long-term program. In 1972 Lear demonstrated a steam-powered bus.

Benz and the gasoline car. Carl Benz was completely dedicated to the proposition that the internal-combustion engine would supersede the horse and revolutionize the world's transportation. He persisted in his efforts to build a gasoline-fuelled vehicle in the face of many obstacles, including lack of money to the point of poverty and the bitter objections of his associates, who considered him unbalanced on the subject.

Benz's day of triumph

Benz ran his first car, a three-wheeler powered by a two-cycle, one-cylinder engine, early in 1885, on a happy and triumphant day. He circled a cinder track beside his

By courtesy of the Smithsonian Institution, Washington, D.C.

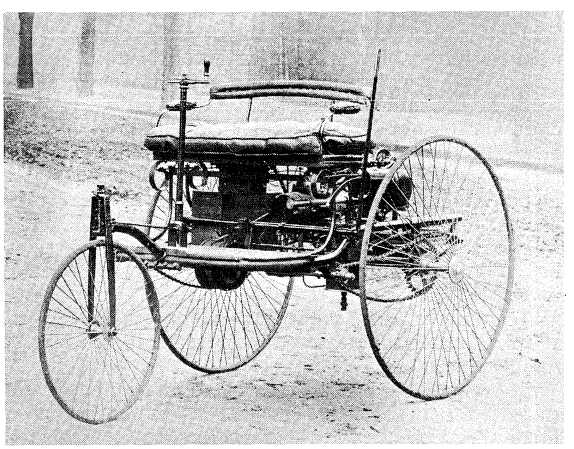


Figure 3: The first Benz, 1885, a three-wheeled vehicle with a steel frame in the shape of a horseshoe. This car was first driven in public in Mannheim on July 3, 1886, where a speed of 15 kilometres per hour was reached.

small factory, his workmen running beside the car, his wife running, too, clapping her hands; the little machine made four circuits of the track, stalling only twice, before a broken chain stopped it. Even Max Rose, Benz's skeptical partner, whose money had made the car possible, conceded that he was mildly impressed; but, like Siegfried Marcus, he remained convinced to the end of his association with Benz that there was no future in the horseless carriage.

In the autumn of 1885, when Benz attempted a public showing of a slightly improved model, he forgot to steer and smashed it against the brick wall surrounding the yard of his own house.

Benz made his first sale to a Parisian named Émile Rogers in 1887. Gradually, however, the soundness of his design and the quality and care that went into the material and the construction of his cars bore weight, and they sold well. By 1888 he was employing 50 workmen to build the tricycle car; in 1890 he began to make a four-wheeler.

In his way, Benz was almost as dogmatic and reactionary as Marcus had been; he objected to redesign of his original cars, and some authorities believe that he was never really convinced that his original concepts had been improved upon.

Development of the Daimler. Gunsmithing was Gottlieb Daimler's first vocation, and he showed marked talent, but he abandoned the trade to go to engineering school, studying in Germany, England, Belgium, and France. In Germany he worked for various engineering and machining concerns, including the Karlsruhe Maschinenbaugesellschaft, a firm that much earlier had employed Benz.

In 1872 Nikolaus A. Otto offered Daimler the position of technical director of his firm, then building stationary gasoline engines. Daimler was of notable utility to Otto during the next decade, when important work was done on the four-stroke engine. Daimler brought in several brilliant researchers, among them Wilhelm Maybach, but in 1882 both Daimler and Maybach resigned because of Daimler's conviction that Otto did not understand the potential of the internal-combustion engine. They set up a shop in Bad Cannstatt and built an air-cooled, one-cylinder engine. The first high-speed internal-combustion engine, it was designed to run at 900 revolutions per minute (rpm), while Benz's first tricycle engine had operated at only 250 rpm. Daimler and Maybach built a second engine and mounted it on a wooden bicycle, which first ran on November 10, 1885. The next year the first Daimler four-wheeled road vehicle was made: a carriage modified to be driven by a one-cylinder engine. Daimler appears to have believed that the first phase of the automobile era would be a mass conversion of carriages to engine drive; Benz apparently thought of the motorcar as a separate device. Daimler's licensees in France were René Panhard and Émile Levassor. In 1889 they entered the field independently, and the Panhard–Levassor designs of 1891–94 are of primary importance. They were true automobiles, not carriages modified for self-propulsion.

Daimler's 1889 car was a departure from previous practice. It was based on a framework of light tubing, it had the engine in the rear, its wheels were driven by belt, and it was steered by tiller. Remarkably, it had four speeds. This car had obvious commercial value, and in the following year the Daimler Motoren-Gesellschaft was founded. The British Daimler automobile was started as a manufactory licensed by the German company but later became quite independent of it. (To distinguish machines made by the two firms in the early years, the German cars are usually referred to as Cannstatt–Daimlers.) Products of the Daimler–Benz company (the two firms were merged in 1926) are sold under the name Mercedes-Benz.

In France the giants were De Dion-Bouton, Peugeot, Renault (the last two are still in existence). The Italians were later in the field: the Steffanini-Martina of 1896 is

By courtesy of the Veteran Car Club of Great Britain



Figure 4: Panhard–Levassor, 1891, the first vehicle with an internal-combustion engine mounted at the front of the chassis.



Figure 5: Hammel, 1886, possibly the oldest automobile still in running order.
By courtesy of the Montagu Motor Museum, Beaulieu, Hampshire

thought of as the foundation of the industry in Italy, and Isotta-Fraschini was founded about 1898. Giovanni Agnelli founded Fiat (Fabbrica Italiana Automobili Torino) in 1899, saw it grow into one of the weightiest industrial complexes in the world, and maintained personal control until his death in 1945. Fabricators of lesser puissance but great repute were Lancia, Alfa Romeo, Maserati, and Ferrari, for years the standard against which other Grand Prix and Gran Turismo motorcars were judged.

The Grand Prix and Gran Turismo cars

The smaller European nations produced makes that were to remain less well known: the Belgian Minerva, Métallurgique, and Excelsior; the Swiss Martini; the Austrian Austro-Daimler, Steyr, Gräf, and Stift; and the Czechoslovakian Skoda and Tatra, the latter technically interesting for its big (V-8) rear-mounted engine. Spain had the Elizalde, and the classic Hispano-Suiza by the great Swiss designer Marc Birkigt was Spanish financed. Curiously, what some authorities believe to be the oldest automobile still in running order early in the 1970s was an 1886 Hammel, made in Denmark.

Early efforts in the United States. The Daimler–Benz claim to the invention of the automobile was attacked in 1895 when U.S. patent 549,160 was granted to George B. Selden as inventor of the automobile. Selden had filed on May 8, 1879, although he had not at that time built an automobile. He was successful in an effort to keep the patent pending for 16 years.

Some authorities credit Charles E. and J. Frank Duryea with creating the first American gasoline-powered automobile, in 1892–93. The idea for the car apparently originated with Charles, and the machine was built by Frank. The Duryea consisted of a one-cylinder gasoline engine,

By courtesy of the Montagu Motor Museum, Beaulieu, Hampshire

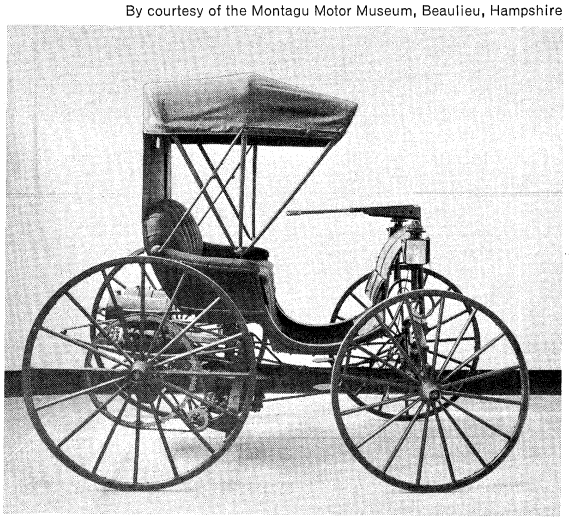


Figure 6: Duryea, 1893, a one-cylinder gasoline engine with electrical ignition.

with electrical ignition, installed in a secondhand carriage. It ran first on September 21, 1893. Driving a later model, J. Frank Duryea won the first automobile race in America in which more than two cars competed, the *Chicago Times-Herald* Race from Chicago to Evanston, Illinois, and return, in November 1895; the distance was 54.36 miles (87.47 kilometres). Duryea cars remained on the market until 1917. Despite this, many historians are convinced that the Duryea was not the first U.S. internal-combustion automobile and that this distinction should be assigned to a car built in 1890 and run in 1891 by John William Lambert of Ohio City, Ohio.

The Duryea was certainly not the first U.S.-built road vehicle. A number of steam carriages had been built after Oliver Evans' first example. In March 1863 the magazine *Scientific American* described tests of a vehicle that weighed only 650 pounds (about 300 kilograms) and achieved a speed of 20 miles (30 kilometres) per hour. Another American, Frank Curtis of Newburyport, Massachusetts, is remembered for building a personal steam carriage to the order of a Boston man who failed to meet the payment schedule, whereupon Curtis made the first recorded repossession of a motor vehicle.

By courtesy of the Lambert Corp.

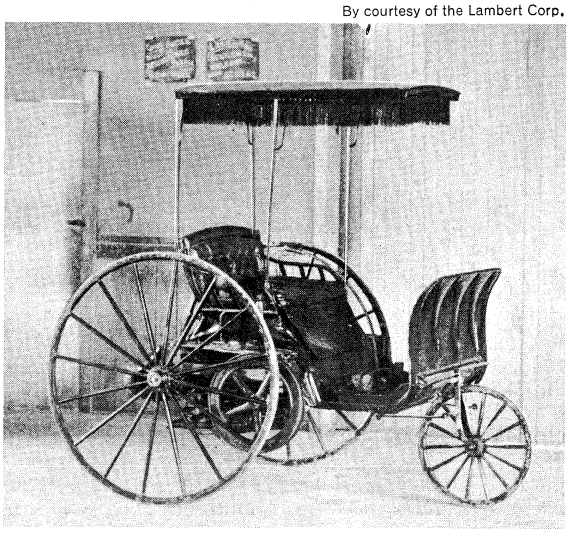


Figure 7: The Lambert, 1891, a gasoline-driven, single-cylinder automobile with a maximum speed of 15 miles per hour.

The U.S. Patent Office issued patents in September 1889 and April 1893 for three-wheeled gasoline-engine carriages, and Gottfried Schloemer of Milwaukee, Wisconsin, in 1890 built a successful car that still exists. A machinist of Allentown, Pennsylvania, in 1899 produced an automobile that ran quite well despite its dependence on the most primitive possible carburetor: a wooden wick that delivered fuel to the cylinder by capillary attraction. Charles Black in 1891, aided by Elwood Haynes, constructed a car in Indianapolis, Indiana, and Haynes followed the Duryea brothers with a gasoline car demonstrated in Kokomo, Indiana, on July 4, 1894. Charles Brady King built a car in Detroit, the first of the millions to issue from the city, that first ran on March 6, 1896.

Ransom Eli Olds, whose name survives in the Oldsmobile, was active in gasoline-engine research in the 1890s, after initially being interested in steam; so were Alexander Winton and James Ward Packard. By 1898 there were more than 50 automobile companies in existence—although the name automobile had not been settled upon (also considered were motor fly, diamote, automotive, autometon, mocole, oleo locomotive, motorig, bolvite, locomotive car, autobaine, autokenetic, and electrobat).

The first commercially successful U.S. car

The three-horsepower, curved-dash Oldsmobile was the first commercially successful American-made automobile: 425 of them were sold in 1901 and 5,000 in 1904 (the model is still prized by collectors), and the firm prospered. Its prosperity was noted by others, and, from 1904 to 1908, 241 automobile-manufacturing firms went

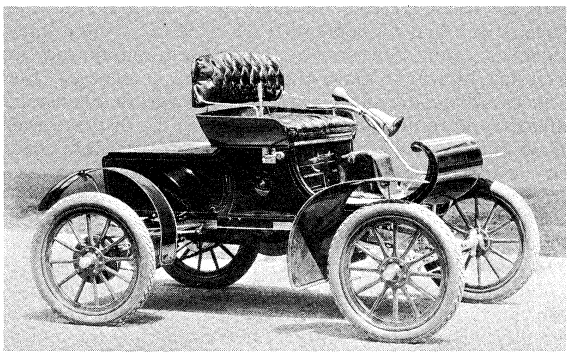


Figure 8: Oldsmobile, 1903, curved dash.
By courtesy of Oldsmobile Division

into business in the U.S. One of these was the Ford Motor Company, which was organized as a corporation in June 1903 and sold its first car on the following July 23; the company produced 1,700 cars during its first full year of business.

RECENT DEVELOPMENTS

Electric power. At the turn of the century, 40 percent of U.S. automobiles were powered by steam, 38 percent by electricity, and 22 percent by gasoline. In the face of the gasoline car's unreliability, noise, and vibration and the steamer's complications and thirst, the electric offered attractive selling points: notably, instant self-starting, silence, and minimal maintenance. The first automobile to exceed 100 kilometres (60 miles) an hour was an electric (Camille Jenatzy's La Jamais Contente, 1899). An electric, also Jenatzy's, had been the easy winner in 1898 of a French hill-climb contest to assay the three forms of power.

Invention of the storage battery, by Gaston Planté of France in 1859–60, and its improvement by Camille Faure in 1881 made the electric vehicle possible, and what was probably the first, a tricycle, ran in Paris in 1881. It was followed by other three-wheelers in London, 1882, and Boston, 1888. The first American battery-powered automobile, built in Des Moines, Iowa, in 1890, could maintain a speed of 14 miles (23 kilometres) per hour.

The peak year of the electric's acceptance was 1912, when 20 companies were in the trade and 33,842 cars were registered in the United States, the country in which the electric car had maximum acceptance. It was another application of battery power, the electric self-starter, that did as much as anything to doom the electric car by eliminating the dreaded hand crank and making the internal-combustion engine car amenable to operation by women. Further, the electric had never been really suited to any but limited urban use because of low speed (15–20 miles per hour), short range (30–40 miles), and dependence on recharging equipment.

The electric came in for its share of attention during the general re-examination of the automobile following World War II; many variations on the theme were proposed, and a number of vehicles were constructed and marketed. The problems of battery weight, low speed, and short range were not notably diminished, however, and new forms of battery—tubular lead–acid, silver peroxide–zinc, cobalt, ceramic—were expensive, exotic, or dangerous. Manufacturers of gasoline-powered cars resisted proposals on electric propulsion on the grounds that, like the steam car, it posed technical problems of such complexity that solutions would require large expenditure and many years.

Universal changeover to the electric automobile, seriously proposed in some quarters as an antipollution imperative, would have vast effects on the world's economy. The electric car is extraordinarily durable, is mechanically almost trouble-free, cannot support a host of accessories, and uses almost no oil, but would impose a serious demand on the world's electric-power system, much of which is fossil fuelled.

Ford and the automotive revolution. Henry Ford produced eight versions of cars before the Model T of 1908, with which his name is synonymous; these were the models A, B, C, F, K, N, R, and S. They were not remarkable automobiles, but public response to the less expensive ones (the firm made some fairly costly cars at first) indicated the soundness of Ford's idea—to turn the automobile from a luxury and a plaything into a necessity by making it cheap, versatile, and easy to maintain.

Within two decades, the American automobile had won the revolution Henry Ford had begun. The country was on wheels, and the manufacture and sale of the automobile were a main prop under the U.S. economy. The closed car was no longer exclusively a rich's man possession. In 1920 most cars had been open models, the occupants protected from the weather by canvas-and-isinglass side curtains; ten years later Detroit was producing closed models almost exclusively.

The 1920s saw also the emergence of the great European producers, Austin, Morris, Singer, Fiat, Citroën, companies founded between 1906 and 1919, as the Ford doctrine was carried forward on the Continent and elsewhere. Universal motor transportation was a long way off, but the concept of the small car that arose in the Austin Seven and the Fiat Topolino, to name two of the descendants of Ettore Bugatti's tiny Bébé Peugeot of 1911, was to have a profound effect.

By courtesy of the Ford Motor Company

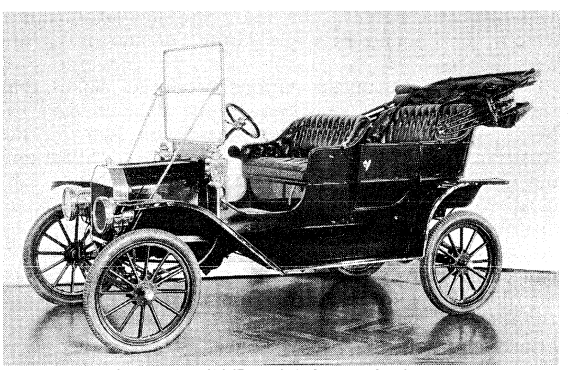


Figure 9: Ford, 1909 Model T, which featured a brass radiator, four-cylinder engine cast in one piece, and a detachable cylinder head.

Age of the classic cars. The decade 1925–35 was notable not only for the appearance of many new small automobiles; it also saw the building of many ultralarge ones. The years from 1925 to 1942 are cited by collectors of automobiles as the classic years, a period that saw the rise of the luxurious fast motorcar to a peak it seemed unlikely to reach again.

Rolls-Royce and Bentley

The first name in this field was Rolls-Royce, founded in 1904. Most Rolls-Royce chassis are designed for limousine and large sedan bodies, but the firm once made a comparatively light car (called the Twenty), and it has throughout its history produced fast models in addition to its regular line; *e.g.*, after World War II, the Continental, built under the Bentley Motors Ltd. label.

Other motorcars of this type included the Hispano-Suiza of Spain and France; Bugatti, Delage, Delahaye, Hotchkiss, Talbot (Darracq), and Voisin of France; the Duesenberg, Cadillac, Packard, and Pierce-Arrow of the United States; the Horch, Maybach, and Mercedes-Benz of Germany; the Belgian Minerva; and the Italian Isotta-Fraschini. These were costly machines, priced roughly from $7,500 to $40,000; fast, at 90 to 130 miles (140 to 210 kilometres) per hour and as comfortable as the state of the art would allow; and limited in luxury only by the purse of the purchaser. The great custom coach builders of England who furnished bodies for Rolls-Royce machines, unruffled by the whims of their clients, were prepared to satisfy any request, whether for upholstery in matched ostrich hide with ivory buttons or for a dashboard in rosewood.

The most expensive standard automobile of which there exists convincing record was the Type 41 Bugatti, pro-

duced in the 1920s by Ettore Bugatti, an Italian of extraordinary gifts who built cars in France, most of them racing and sports types, from 1909 to 1939. The Type 41 Bugatti, also called La Royale, was cataloged at a chassis price of 500,000 francs, about $20,000. Only six of the cars were built, and one of them, carrying a four-passenger convertible sedan body by a German coach maker, cost $43,000.

The market collapse of 1929 ended the era of the really luxurious motorcar. After World War II even Rolls-Royce abandoned its long-held policy of producing a standard chassis with custom-made bodies and offered a standard sedan that could be bought straight off the showroom floor. Custom bodies were still available to Rolls-Royce purchasers through the old-line suppliers, of course. In 1971 an overextension of Rolls-Royce's aero-engine division put the company into receivership and resulted in nationalization and a separate corporate entity for the profit-making automobile division, which continued in production without interruption.

The antique car

A few hundred persons in the 1930s were seriously interested in the collection of antique automobiles, purchasing pre-World War I specimens for $10, $25, $50—or the trouble of towing them away from a junkyard. Distaste for the mechanically dull and stylistically unpleasant 1946–50 U.S. cars attracted many Americans to the hobby, prices rose, and membership in concerned organizations such as the Antique Automobile Club of America increased. In 1951 the Museum of Modern Art in New York held an exhibition of the automobile as art object that notably influenced opinion molders. Publications devoted to the automobile proliferated; the old European idea of the *concours d'élégance* (competition in design elegance) began to be reflected in formal showings of fine pre-1939 motorcars. Shops specializing in the restoration of old automobiles were set up, and the best of them were soon booked months ahead. Collectors worked over primary sources so heavily that the discovery of a desirable car tucked away in a forgotten corner of a barn became rare indeed, and interest in the automobile as a hedge against inflation drove prices steeply upward. For example, a 1913 Mercer Raceabout, which had cost $1,250 in 1951, appreciated, in 20 years, to $45,000, and a 1936 Bugatti, $5,000 in 1954, fetched $59,000 at an auction in Los Angeles in June 1971. An offer of $50,000 for a 1933 Duesenberg was thought risible, and market observers confidently predicted six-figures for desirable items in the 1970s.

But rare motorcars were increasingly being taken off the market by museums; the first of which to do so may have been the Dangerfield Museum set up in England in 1912. The Harrah Collection in Reno, Nevada, became by far the largest, with about 1,350 vehicles, and, by many standards, incomparably the best of the world's 170-odd motorcar museums. The Montagu Motor Museum in England, the Museo dell'Automobile Carlo Biscaretti di Ruffia in Italy, and the Daimler–Benz Museum in Germany house superb collections, and there are many splendid smaller exhibits throughout Europe.

Developments after World War II. The effect of Italian ideas on the world's automobile-body designers was profound when manufacture began to be resumed in 1946. Pininfarina of Turin was the best known of the coach makers who established the characteristic Italian approach: grace, lightness in line and substance, and minimal use of decoration. Designs clearly derivative of those of Italian origin appeared everywhere, and manufacturers in France, England, and the United States contracted for the services of Italian *carrozzerie* (body factories).

The trend toward the small automobile in the United States, clear if not obtrusive after 1932, was strongly accelerated by World War II. A leading factor was the return from duty in Europe of servicemen who had previously not known of the existence of the sheer variety of automobiles the world afforded. The sports car, designed for pleasure, not transport, was particularly new to young Americans. The characteristics of automobiles such as the British sports two-seater M.G., plus their availability at a time of short domestic supply, made them attractive, and the importation of European-made models into the U.S. increased rapidly.

Auto racing as a major sport

At this time, automobile racing, which in the Vanderbilt Cup days around 1910 had drawn the biggest crowds in U.S. sports history, began to regain popularity. By 1954 motor racing had become a high-ranking American spectator sport, and by 1969 estimated attendance was 41,300,000, higher than that for baseball or football. Only horse racing showed a total higher than auto racing. In the 1950s and 1960s, U.S. manufacturers returned to racing as exploitation (a standard practice 1900–30); Ford was most successful, winning the Le Mans 24-hour Grand Prix d'Endurance race—the first U.S.-built car to do so—in 1966 and 1967 and producing, in a remarkably short time, a racing engine that dominated major U.S. tracks.

While the size of the standard U.S. motorcar increased steadily from the late 1940s to the early 1960s, and its design seemed to many to become even more bizarre, a small segment of the population was demonstrating a preference for smaller cars and for comparatively uncluttered styling. The success of the German Volkswagen and other small cars eventually led the major U.S. producers simultaneously to undertake the production of automobiles generically termed compact. At around 110-inch (280-centimetre) wheelbase, they were smaller than the less big U.S. cars, bigger than the average European models. By the mid-1960s a demand for more highly individualized luxury models of compact size had brought in the "fastback," a two-door coupé with roof line extending in a continuous curve to a rear bumper, a design reflecting the European tradition of simpler, cleaner lines. Not only in design but also in more basic engineering problems, much U.S. experimentation followed research begun by the European industry: development of gas turbine engines; experimentation with fuel-injection systems; disk brakes; return to body-and-frame assembly; introduction of rear-engine and later, conversely, of front-wheel-drive models.

In Europe and Japan, which became a major producer in the 1960s, the small car continued to dominate, though the number of larger automobiles increased. In the United States the market remained significantly influenced by European small-car design, and a new generation of "really small" cars appeared.

A possibly significant sign in popular taste discernible in the early 1970s was a new fad among young people, especially in California. Young Californians, whose fads had often presaged trends, began to show disinterest in the high acceleration and top-speed characteristics that had preoccupied them since the 1930s and were shopping the used-cars lots for vans, which they transformed into mobile recreation rooms and ran sedately at legal speeds.

(K.W.P.)

II. Modern automobiles

AUTOMOTIVE SYSTEMS

The modern automobile is a complex technical system composed of more than 14,000 parts. The subsystems that comprise the complete vehicle have evolved over many decades and include the body, or passenger compartment; the engine, or power source; the chassis, or framework supporting the engine and body; a system of shafts and gears (transmission) for transmitting power from the engine to the wheels; braking and steering mechanisms; cooling and electrical systems; fuel and lubrication; and an increasingly sophisticated group of accessories. The principal mechanical subsystems, such as gears, drives, couplings, brakes, bearings, and shafts, are treated in detail in the article MACHINES AND MACHINE COMPONENTS.

Between 25,000,000 and 30,000,000 automobiles a year were being produced throughout the world in the 1970s. While the number of manufacturers was declining, the number of designs and models was proliferating. In an effort to increase effectiveness, cars were being engineered for longer production runs to utilize tooling more fully. Specific designs for individual segments of the

market—subcompact, compact, sporty, intermediate, regular, and luxury—have been more successful. Firms with less capital have developed designs to meet particular needs. By late 1970 there were 243,000,000 cars, trucks, and buses in the world.

Motorists in the United States in the early 1970s numbered about half the world's total and drove about 1,000,000,000,000 miles each year, one-fifth of that total at speeds of 60 to 70 miles (100 to 110 kilometres) per hour on limited-access express highways linking major cities of the nation. Similar road networks and driving patterns were emerging in other industrialized nations. Vehicle stability and properly selected, manufactured, and assembled vehicle components are important in highway driving. Vehicle stability depends principally upon the distribution of weight between the front and rear wheels, the height of the centre of gravity and its position relative to the aerodynamic centre of pressure of the vehicle, suspension characteristics, and whether front or rear wheels are used for propulsion. Weight distribution depends principally upon the location and size of the engine. The common practice of front-mounted engines exploits the stability that is more readily achieved with this layout. The development of aluminum and magnesium engines and new manufacturing processes have, however, made it possible to locate the engine at the rear without necessarily compromising stability.

Stability and weight distribution

Bodies. Automotive body styles are frequently categorized according to the number of doors in the body and the manner in which the roof is supported. Automobile roofs are conventionally supported by three pillars on each side of the body. Convertible models with retractable fabric tops rely upon the pillar at the side of the windshield for upper body strength, as convertible mechanisms and glass areas are essentially nonstructural. Beginning in 1956, automotive stylists sought to increase the openness of conventional bodies by eliminating the side pillar behind the front door. This became known as hardtop styling, and the two-door hardtop was the most popular body style in 1970.

The high cost of new tools makes it impractical for manufacturers to produce totally new designs every year. New designs are usually programmed on three- to six-year cycles with generally minor refinements appearing during the cycle. As much as four years of planning and new tool purchasing is needed for a completely new design.

Automotive bodies are generally formed out of sheet steel. Elements are added to the alloy to improve its ability to be formed into deeper depressions without wrinkling or tearing in manufacturing presses. Steel is used because of its general availability, low cost, and good workability. For special applications, however, other materials, such as aluminum and fibre-glass-reinforced plastic, are used because of their special properties. Plastics have low ductility (are more brittle) in comparison with the same weight of steel or aluminum. Therefore, they are used in generally nonstructural areas.

To protect bodies from corrosive elements and to maintain their strength and appearance, special priming and painting processes are used. Bodies are first dipped in cleaning baths to remove oil and other foreign matter. They then go through a succession of dip and spray cycles. Enamel and acrylic lacquer are both in common use. Electrodeposition of the sprayed paint, a process in which the paint spray is given an electrostatic charge and then attracted to the surface by a high voltage, helps assure that an even coat is applied and that hard-to-reach areas are covered. Ovens with conveyor lines are used to speed the drying process in the factory.

Chassis. The chassis of the modern automobile is the main structure of the vehicle. In most designs a pressed-steel frame forms a skeleton on which the engine, wheels, axle assemblies, transmission, steering mechanism, brakes, and suspension members are mounted. The body is flexibly bolted to the chassis during the manufacturing process. The combination of body and frame absorbs the reactions from the movements of the engine and axle, receives the reaction forces of the wheels in acceleration and braking, absorbs aerodynamic wind forces and road shocks through the suspension, and absorbs the energy of impact in the event of an accident.

Since the 1930s there has been a trend toward combining the chassis frame and the body into a single structural element. In this arrangement the steel body shell is reinforced with braces that make it rigid enough to resist the forces which are applied to it. While this arrangement has been almost universally adopted for small cars, there has been a return to separate frames for other cars in order to achieve better noise-isolation characteristics and simplified tooling for body changes. The presence of a heavy-gauge frame member at the perimeter of the body also tends to limit intrusion in accidents.

Engines. A wide range of energy-conversion systems has been used experimentally and in automotive production. These include electric, steam, solar, turbine, rotary, and a variety of piston-type internal-combustion engines. The most successful for automobiles has been the reciprocating-piston internal-combustion engine, operating on a four-stroke cycle (see also GASOLINE ENGINE), while diesel engines are widely used for trucks and buses (see also DIESEL ENGINE). The gasoline engine was originally selected for automobiles because it could operate more flexibly over a wide range of speeds and the power developed for a given weight engine was reasonable, it could be produced by economical mass-production methods, and it used a readily available, moderately priced fuel—gasoline. Reliability, compact size, and range of operation later became important factors.

Predominance of gasoline engines

Beginning in the 1960s there has been a reassessment of these priorities with new emphasis on the pollution-producing characteristics of automotive power systems. This has created new interest in alternate power sources, but by 1972 there were no commercially successful replacements for the gasoline engine, which had developed new emission-control devices to improve its emission performance.

The pollutants requiring control include hydrocarbons that are unburned fuel compounds, carbon monoxide, and oxides of nitrogen. It was estimated in 1970 that automotive emissions produced 39 percent of the national tonnage of pollution in the United States, a proportion that was being reduced as old cars were replaced by new ones with better control systems.

In the late 1940s a trend began to increase engine horsepower, particularly in United States models. Design changes incorporated all known methods of increasing engine capacity, including increasing the pressure in the cylinders to improve efficiency, increasing the size of the engine, and increasing the speed at which power is generated.

The higher dynamic forces and pressures created by these changes created engine vibration and size problems that led to stiffer, more compact engines with V and opposed cylinder layouts replacing longer straight-line arrangements. In passenger cars, V-8 layouts were adopted for all piston displacements greater than 250 cubic inches.

The advent of smaller cars brought a return to smaller engines, four- and six-cylinder designs rated as low as 80 horsepower, compared with the standard-size V-8 of large cylinder bore and relatively short piston stroke with horsepower ratings in the range from 250 to 350.

Multiplicity of engine types

European automobile engines were of a much wider variety, ranging from one to 12 cylinders, with corresponding differences in overall size, weight, piston displacement, and cylinder bores. A majority of the models had four cylinders and horsepower ratings from 19 to 120. Several three-cylinder, two-stroke-cycle models were built. Most engines had straight or in-line cylinders. There were, however, several V-type models and horizontally opposed two- and four-cylinder makes. Overhead camshafts were frequently employed. The smaller engines were commonly air cooled and located at the rear of the vehicle; compression ratios were relatively low.

Fuel and lubrication. Gasoline is essentially the only fuel used for automobile operation, although diesel fuels are used for many trucks and buses and a few automo-

biles. The most important requirements of a fuel for automobile use are proper volatility, sufficient antiknock quality, and freedom from polluting by-products of combustion. The volatility is adjusted seasonally by refiners so that sufficient gasoline vaporizes, even in extreme cold weather, to permit the engine to be started but without being so volatile that vapour lock (a bubble of vapour that blocks gasoline flow) will prevent proper functioning of the fuel system at the highest outdoor temperatures encountered.

Antiknock qualities are controlled by refining processes that produce compact, stable molecules instead of long, straight-chain molecules. Antiknock compounds, principally tetraethyl lead, are added to some gasolines as a more economic way of attaining the same result. Small deposits of lead on such places as engine-valve seats serve to improve valve life. These same deposits, however, can contaminate certain emission-pollution-control systems such as catalytic converters. Antiknock quality is rated by the octane number of the gasoline. The octane-number requirement of an automobile engine depends primarily upon the compression ratio of the engine but is also affected by combustion-chamber design and deposits formed on the chamber walls. Various additives intended to minimize the effects of deposits are used in most automotive gasolines. No advantage is gained by using a fuel of octane number higher than that necessary to eliminate objectionable knock. In the early 1970s regular gasoline carried an octane rating of around 93 and high-test in the neighbourhood of 100. All moving parts of an automobile require lubrication. Without it, friction would increase power consumption and damage the parts. The lubricant also serves as a coolant, a noise-reducing cushion, and a sealant between engine piston rings and cylinder walls.

The engine lubrication system consists of an oil sump and a gear-type pump that delivers the oil under pressure to a system of drilled passages leading to various bearings. Drilled passages in the crankshaft conduct oil to lubricate the connecting-rod bearings, and leakage from these bearings sprays oil up into the cylinder bores to lubricate the pistons, piston pins, and piston rings. The spray also lubricates the cams and valve lifters. Usually the oil passes through a paper or cotton-waste filter before it enters the distribution passages; bypass filters sometimes are used to receive a small stream of oil from the system continuously and return it to the sump after removing any solids.

Engine oils

Engine oils may be identified by letters designating the severity of service for which the oil is suitable and by numbers indicating the viscosity range. Selection of the proper oil depends upon the conditions under which it will be used—for example, continuous high speeds, heavy loads, frequent starts and stops—and engine components such as hydraulic valve lifters that are sensitive to oil condition.

Wheel bearings and universal joints require a fairly stiff grease; other chassis joints require a soft grease that can be injected by pressure guns. Hydraulic transmissions require a special grade of light hydraulic fluid, and manually shifted transmissions use a heavier gear oil similar to that for rear axles. Hypoid gears used in most rear axles require oils that are compounded to resist heavy loads on the gear teeth. Gears and bearings in lightly loaded components, such as generators and window regulators, are fabricated from self-lubricating plastic materials (see also LUBRICATION).

Cooling system. The vast majority of automobiles employ liquid cooling systems for their engines. A typical automotive cooling system comprises (1) a series of channels cast into the engine block and cylinder head, surrounding the combustion chambers with circulating water or other coolant to carry away excessive heat; (2) a radiator, consisting of many small tubes equipped with a honeycomb of fins to radiate heat rapidly, that receives and cools hot liquid from the engine; (3) a water pump, usually of the centrifugal type, to circulate coolant through the system; (4) a thermostat, which maintains constant temperature by automatically varying the amount of coolant passing into the radiator; and (5) a fan, which draws fresh air through the radiator.

For operation at outdoor temperatures below freezing, it is necessary to prevent the coolant from freezing. This is usually done by adding some compound to depress the freezing point of the coolant. Alcohol formerly was commonly used, but it has a relatively low boiling point and evaporates quite easily, making it less desirable than organic compounds with a high boiling point, such as ethylene glycol. By varying the amount of additive, it is possible to protect against freezing of the coolant down to any minimum temperature normally encountered.

Air cooling offers the important advantage of eliminating not only freezing and boiling of the coolant at temperature extremes but also corrosion damage to the cooling system. Control of engine temperature is more difficult, however, and the engine temperature is more likely to vary with the outside temperature. Air-cooled cylinders operate at higher, more efficient temperatures, but compression ratios (with the same fuels) must be lower, because combustion control is more difficult. In addition, it is more difficult to utilize engine heat to warm the car interior during winter. In spite of these disadvantages, in the early 1970s a sizable number of cars with air-cooled motors were being manufactured.

With the aim of eliminating seasonal draining of liquid cooling systems, manufacturers in the later 1950s and early 1960s introduced liquid coolants intended to replace water–antifreeze mixtures. These contained corrosion inhibitors that were designed to make it necessary to drain and refill the cooling system only once a year.

Permanent coolants

Other compounds under development in the early 1960s, some of which were organic silicates, contained no water and were permanent in characteristics. Such liquids permit the cooling system to be sealed and require no attention. The ultimate practicability of these permanent coolants depends upon their ability to carry heat away from the engine. All previously known liquids with appropriate boiling and freezing points were poorer than water in this respect and so were not as effective as engine coolants.

Pressurized cooling systems with operating pressures up to 14 pounds per square inch have been used to increase effective operating temperatures. Partially sealed systems using coolant reservoirs for coolant expansion if the engine overheats were introduced in 1970.

Electrical system. Originally, the electrical system of the automobile was limited to the ignition equipment. With the advent of the electrical starter on a 1912 model, electric lights and horns began to replace the kerosene and acetylene lights and the bulb horns. Electrification was rapid and complete, and by 1930 six-volt systems were standard everywhere. The electrical system comprises a storage battery, generator, or dynamo, starting (cranking) motor, lighting system, ignition system, and various accessories and controls.

Increased engine speeds and higher cylinder pressures of the post-World War II cars made it increasingly difficult to meet high ignition voltage requirements. The larger engines required higher cranking torque. Additional electrically operated features, such as radios, window regulators, and multispeed windshield wipers, also added to system requirements. To meet these needs, 12-volt systems generally replaced the six-volt systems in 1956 production.

The ignition system provides the spark to ignite the air–fuel mixture in the cylinders of the engine. In order to jump the gap between the electrodes of the spark plugs, the 12-volt potential of the electrical system must be stepped up to about 20,000 volts. This is done by a circuit that starts with the battery, one side of which is grounded on the chassis and leads through the ignition switch to the primary winding of the ignition coil and back to the ground through an interrupter switch, or circuit breaker. Interrupting the primary circuit induces a high voltage across the secondary of the coil. The high-voltage secondary terminal of the coil leads to a distributor that acts as a rotary switch, alternately connecting the coil to each of the wires leading to the spark plugs. The system

consists of the spark plugs, contact breaker, coil, distributor, and battery.

The generator

The source of electrical energy for the various electrical devices of the automobile is a generator, or dynamo, that is belt driven from the engine crankshaft. The generator is usually a two-pole, direct-current type with a field controlled by a voltage regulator the function of which is to match the generator output to the electrical load and also to the charging requirements of the battery, regardless of engine speed. High-wattage electrical loads, resulting from the addition of many electrical accessories, has made it increasingly difficult to design direct-current generators with sufficiently high capacity to maintain the battery in a fully charged condition. Alternating-current generators, or alternators, with built-in rectifiers were introduced in 1958 to provide for these loads.

A lead–acid battery serves as a reservoir to store excess output of the generator by chemical changes in the sulfuric acid electrolyte and in the composition of the lead plates. Energy for the starting motor is thus made available along with power for operating other electrical devices when the engine is not running or when the generator speed is not sufficiently high to carry the load (see also BATTERIES AND FUEL CELLS).

The starting motor drives a small spur gear so arranged that it automatically moves into mesh with gear teeth on the rim of the flywheel as the starting-motor armature begins to turn. When the engine starts, the small gear is disengaged, thus preventing damage to the starting motor from overspeeding. The starting motor is designed for high current consumption and delivers considerable power for its size for a limited time.

Headlights that satisfactorily illuminate the highway ahead of the automobile for night driving without temporarily blinding approaching drivers have long been sought. Early dimming devices of the resistance type, which decreased the brightness of the headlights when meeting another car, gave way to mechanical tilting reflectors and later to double-filament bulbs with a high and a low beam, called sealed-beam units.

The double-filament headlight unit of necessity has only one of its filaments at the focal point of the reflector. Because of the greater illumination required for high-speed driving with the high beam, the lower beam filament was placed off centre, with a resulting decrease in lighting effectiveness. In the 1950s some manufacturers began equipping their models with four headlights to improve illumination. In most instances the outer lamps were double-filament types with a high-wattage low beam and a "soft" high beam for city driving. The inner lamps have a single-filament high beam of high wattage that produces a spotlight effect. The low beam of the two outer lamps is focussed slightly to the right and is used when overtaking or meeting other vehicles. The high beams of all four lamps are used on the open highway when there is no approaching traffic. Dimming is automatically achieved on some cars by means of an automatic photocell-controlled switch in the lamp circuit that is triggered by the lights of an oncoming car. Larger diameter double-filament lamps with improved photometrics permitted a return to two-headlamp systems on some cars. Total intensity of forward lighting systems in many places is limited by law to 75,000 candlepower.

Signal lamps and other special-purpose lights increased in usage in the 1960s. Amber-coloured front and red rear signal lights are flashed as a turn indication; all these lights are flashed simultaneously in the "flasher" system for use when a car is parked along a roadway or is travelling at a low speed on a high-speed highway. Marker lights that are visible from the front, side, and rear are also widely required by law. Red-coloured rear signals are used to denote braking, and, on some models, cornering lamps to provide extra illumination in the direction of an intended turn are available. These are actuated in conjunction with the turn signals. Backup lights provide illumination to the rear when backing up.

Transmission. The gasoline engine must be disconnected from the load (the rear wheels) when it is started and when idling. This characteristic necessitates some type of unloading and engaging device to permit gradual application of load to the engine after it has been started. In starting the engine, the crankshaft is rotated (cranked) by an electric motor; this in turn starts action in the cylinders. The torque, or turning effort, that the engine is capable of producing is low at low crankshaft speeds, increasing to a maximum at some fairly high speed representing the maximum, or rated, horsepower.

Engine efficiency

The efficiency of an automobile engine is highest when the load on the engine is high and the throttle is nearly wide open. At moderate speeds on level pavement the power required to propel an automobile is only a portion of the power the engine is capable of developing in its upper range of speeds. Thus, under normal driving conditions at constant moderate speed, the engine may operate at an uneconomically light load unless some means is provided to reduce its speed and power output.

The transmission is such a speed-changing device. It is installed in the drive train that connects the crankshaft of the engine to the driving wheels. This permits the engine to operate at a higher speed when its full power is needed and to slow down to a more economical speed when less power is needed. Under some conditions, as in starting a stationary vehicle or in ascending steep grades, the torque of the engine is insufficient, and amplification is desirable. Most devices employed to change the ratio of the speed of the engine to the speed of the driving wheels multiply the engine torque by the same factor by which the engine speed is increased.

The simplest automobile transmission is the sliding-spur-gear type with three or four forward speeds and reverse. The desired gear ratio is selected by manipulating a shift lever which slides a spur gear into the proper position to engage the various gears. Early devices of this type required considerable skill on the part of the operator to shift the gears smoothly and without clashing the teeth. The shift from low to second gear was the most troublesome, because this shift required that two gears, moving at different speeds, be slid sidewise into tooth engagement. This problem was eliminated by a constant-mesh second gear in which the driven gear of the train is not keyed to the driven shaft of the transmission. Second gear is engaged by means of a toothed or jaw clutch, similar to that used to engage direct drive, arranged so that it can keep the driven gear from turning on the driven shaft.

Ease of shifting into second and high was further improved by the use of synchronizing clutches that engaged ahead of the toothed clutches and caused the two portions of the positive clutch to turn in unison before their engaging teeth touched each other. Another device, a blocker ring, was added later; its function was to block the two members of the clutch from contacting each other before their speeds were synchronized. The only difficulty remaining in the operation of the sliding-gear transmission was the need for simultaneously operating the accelerator pedal, the clutch pedal, and the gearshift lever. The automatic transmission was developed to eliminate this manipulation. Most automatic transmissions employ either a fluid coupling or a hydraulic torque converter, a device for transmitting and amplifying the torque produced by the engine. One manufacturer, DAF of The Netherlands, uses a mechanical linkage. Each type provides for manual selection of reverse and a low range that either prevents automatic upshifts or employs a lower gear ratio than is used in normal driving. Grade-retard provisions are also sometimes included to provide engine braking on hills.

Hydraulic transmissions

By 1970 the hydraulic torque-converter-type automatic transmission dominated its field. In hydraulic transmissions, shifting is done by a speed-sensitive governing device that changes the position of valves that control the flow of hydraulic fluid. The vehicle speeds at which shifts occur depend upon the position of the accelerator pedal, and the driver can delay upshifts until higher speed is attained by depressing the accelerator pedal further. Control is by hydraulically engaged bands and multiple-disk clutches running in oil, either by the driver's operation of the selector lever or by a speed-sensitive governor. Compound planetary gear

trains with multiple sun gears and planet pinions have been designed to provide a low forward speed, an intermediate speed, a reverse, and a means of locking into direct drive. This unit is used with various modifications in almost all hydraulic-torque-converter transmissions.

In hydraulic-torque-converter transmissions, torque is multiplied by means of gear trains and a hydraulic member with three or more elements.

As shown schematically in Figure 10, oil in the housing is accelerated outward by rotating vanes in the pump impeller and, reacting against vanes in the turbine impeller, forces them to rotate. The oil then passes into the stator vanes, which redirect it to the pump. The stator serves as a reaction member providing more torque to turn the turbine than was originally applied to the pump impeller by the engine. Thus, it acts to multiply engine torque by a factor of up to 2½ to 1.

Adapted from *Automotive Industry* (October 1970)

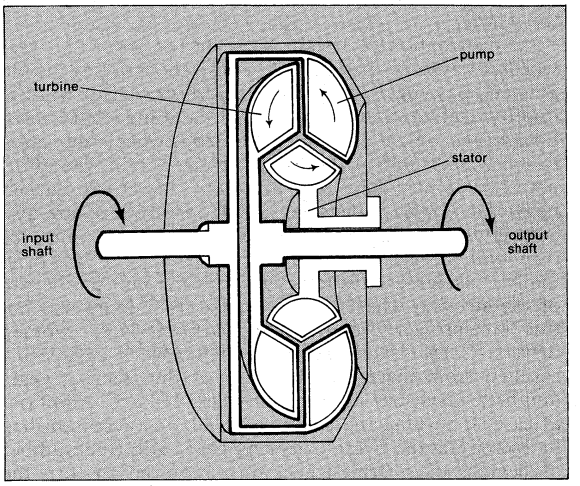


Figure 10: Torque converter.

Blades in all three elements are specially contoured for their specific function and to achieve particular multiplication characteristics. Through a clutch linkage, the stator is allowed gradually to accelerate until it reaches the speed of the pump impeller. During this period torque multiplication gradually drops to approach 1 to 1.

The hydraulic elements are combined with two or three planetary gear sets, which provide further torque multiplication between the turbine and the output shaft.

Automatic transmissions not only require no skill to operate but make possible better performance than is obtainable with transmissions that necessitate the release of a clutch during gear shifts. Power is applied to the driving wheels continuously whenever the accelerator pedal is depressed, and gears are shifted without closing the engine throttle or interrupting the drive.

Small, low-powered European cars usually have manually shifted, four-speed transmissions. Power-operated clutches are sometimes used in combination with pre-selective or semi-automatic transmissions. Fully automatic transmissions are used for the most part in larger cars with ratings above 60 horsepower.

Other mechanical subsystems. *Axles.* Power is conveyed from the transmission to the rear axle by a propeller shaft and universal joints. As body lines were progressively lowered, the floor level came closer to the propeller shaft, necessitating floor humps or tunnels to provide clearance. The adoption of hypoid or offset spiral bevel gears in the rear axle provided an increase in this clearance by lowering the drive pinion below the center of the axle shafts.

The differential

The ring gear of the rear axle surrounds the housing of a differential gear train that serves as an equalizer in dividing the torque between the two driving wheels while permitting one to turn faster than the other when rounding corners. The axle shafts terminate in bevel gears that are connected by several smaller bevel gears mounted on radial axles attached to the differential housing and carried around with it by the ring gear. In its simplest form this differential has the defect that one driving wheel may spin when it loses traction, and the torque applied to the wheel, being equal to that of the slipping wheel, will not be sufficient to drive the car. Several differentials have been developed to overcome this difficulty.

European manufacturers quite generally adopted a wide variety of articulated rear axles that provide individual wheel suspension at the rear as well as the front. Individual rear suspension not only eliminates the heavy rear axle housing but also permits lowered bodies with no floor humps, because the transmission and differential gears can be combined in a housing mounted on a rear cross member moving with the body under suspension-spring action.

In some instances, European manufacturers use articulated or swing axles similar to American rear axles except that the tubular housings surrounding the axle shafts terminate in spherical head segments that fit into matching sockets formed in the sides of the central gear housing. Universal joints within the spherical elements permit the axle shafts to move with the actions of the suspension springs. The gear housing is supported by a rear cross member of the chassis and moves with the sprung portion of the vehicle, as does the drive shaft. Other types eliminate the axle shaft housings and drive the wheels through two open axle shafts, each fitted with two universal joints. The wheels are then individually supported by radius rods or other suitable linkage.

The driving trains for rear engine, rear-wheel-drive cars and front engine, front-wheel-drive cars are simplified by individually supported wheels in that a combined transmission and differential assembly can form a unit with the engine. Two short transverse drive shafts, each having universal joints at both ends, transmit power to the wheels. In this case there is no axle as such, and the wheels may be individually suspended. Front-wheel-drive vehicles built in the United States since 1966 have, however, employed solid rear axles.

Brakes. Most automobile brakes are of the internally expanding shoe type, in which two nearly semicircular brake shoes are pressed against the inner surface of a drum attached to the wheel. Until the 1930s, most brake systems were of the mechanical type; *i.e.*, foot pressure exerted on the brake pedal was carried directly to the brake shoes by a system of flexible cables. Mechanical brakes, however, are difficult to keep adjusted so that equal braking force is applied at each wheel; and, as vehicle weights and speeds increased, more and more effort on the brake pedal was demanded of the driver.

Hydraulic brakes

Mechanical brakes have been almost universally supplanted by hydraulic braking systems, in which the brake pedal is connected to a piston in a master cylinder and thence by steel tubing with flexible sections to individual cylinders at the wheels. Front and rear hydraulic circuits are separated. The wheel cylinders are located between the movable ends of the brake shoes, and each is fitted with two pistons that are forced outward toward the ends of the cylinder by the pressure of the fluid between them. As these pistons move outward, they push the brake shoes against the inner surface of the brake drum attached to the wheel. The larger diameter of the piston in the master cylinder provides a hydraulic force multiplication at the wheel cylinder that reduces the effort required of the driver.

Further increases in vehicle weights and speeds in the 1950s made even hydraulic brakes difficult for drivers to operate effectively, and many of the larger automobiles consequently were equipped with power brake systems. These are virtually the same as the hydraulic system except that the piston of the master cylinder is operated by a vacuum piston and cylinder instead of by the pressure exerted on the brake pedal (see Figure 11). The master cylinder and vacuum power cylinder are formed in one unit. When the driver starts to depress the brake pedal, the control valve closes the atmospheric ports. Further movement of the valve opens the vacuum port connecting the vacuum inlet to the cylinder chamber at the left of the piston. The vacuum inlet pipe leads to the intake

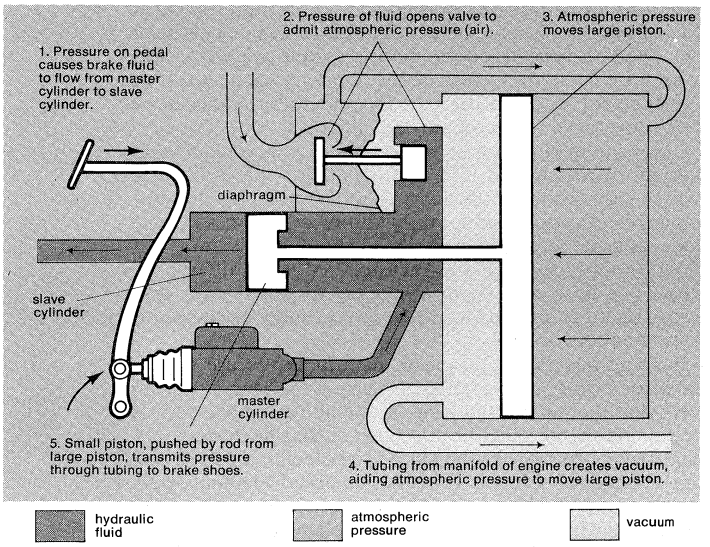


Figure 11: Schematic cross section of vacuum-actuated power brake.
By courtesy of F.E. Compton Co.

manifold, and, when the vacuum valve opens the port, a vacuum is formed on the left side of the power piston. Atmospheric pressure acting on the right side of the piston causes it to move to the left. Plungers in the master cylinder cover the compensating ports and place the hydraulic brake fluid under pressure in the wheel cylinders. As the brake-fluid pressure builds up, the control valve is pushed back. If the pedal is not depressed further, the vacuum port will close, and the brakes will be held in application until the pedal is released. The amount of brake-fluid pressure depends upon how far the pedal is depressed; the driver feels a resistance to pedal movement proportional to the fluid pressure actuating the brakes. The force he must exert is much smaller, however, than is necessary with ordinary brakes, and no appreciable physical effort is necessary to apply the brakes fully.

Overheating of the brake drums and shoes causes the brakes to fade and lose their effectiveness when held in engagement for a considerable length of time. This problem has been attacked by the use of aluminum cooling fins bonded to the outside of the brake drums to increase the rate of heat transfer to the air. Vents in the wheels are provided to increase the air circulation for cooling.

Disk brakes, originally developed for aircraft, have been fitted to some automobiles. Although there are some four-wheel systems, usually disks are mounted on the front wheels, and conventional drum-type units are retained at the rear. Disk designs are somewhat more costly but have become increasingly popular. They were fitted on 56 percent of all new cars in 1970. Each wheel has a hub-mounted disk and a brake unit or caliper rigidly attached to the suspension. The caliper employs two friction-pad assemblies, one on each side of the disk. When the brake is applied, hydraulic pressure forces the friction pads against the disk. This arrangement is self-adjusting, and the ability of the disks to dissipate heat rapidly in the open airstream makes them practically immune to fading.

Parking brakes

Parking brakes usually are of the mechanical type, applying force only to the rear brake shoes by means of a flexible cable connected to a hand lever or pedal. On cars with automatic transmissions, an additional lock is usually provided in the form of a pawl that can be engaged, by placing the shift lever in the "park" position, to prevent the drive shaft and rear wheels from turning.

Steering. Automobiles are steered by a system of gears and linkages that transmit the motion of the steering wheel to the pivoted front wheel hubs. The gear mechanism, located at the lower end of the shaft carrying the steering wheel, is usually a worm-and-nut or cam-and-lever combination that rotates a shaft with an attached crank arm through a small angle as the steering wheel is turned. Tie rods attached to the arm convey its motion to the wheels. In cornering, the inner wheel must turn through a slightly greater angle than the outer wheel, because the inner wheel negotiates a sharper turn. The geometry of the linkage is designed to provide for this.

When the front wheels are independently suspended, the steering must be designed so the wheels are not turned as the tie rods lengthen and shorten as a result of spring action. The point of linkage attachment to the steering gear must be placed so that it can move vertically with respect to the wheel mountings without turning the wheels.

The distribution of weight between the front and rear wheels of automobiles shifted toward the front as the engine and passenger compartment were moved forward to improve riding comfort and road-handling characteristics. As the weight carried on the front wheels increased to more than half of the total vehicle weight, the effort necessary to turn the wheels in steering increased. Larger, heavier cars with wider tires and lower tire pressure also contribute to drag between tires and road that must be overcome in steering, particularly in parking. It was originally considered satisfactory to limit the pull on the rim of the steering wheel to 30 pounds (14 kilograms), but this limit proved to be too high, particularly for women drivers. Considerable reduction in the work of steering resulted from increased efficiency of the steering gears and better bearings in the front wheel linkage. Additional ease of turning the steering wheel was accomplished by increasing the overall steering gear ratio (the number of degrees of steering-wheel turn required to turn the front wheels one degree). Large steering gear ratios make high-speed manoeuvrability more difficult, however, because the steering wheel must be turned through greater angles. On the other hand, steering mechanisms of higher efficiency are also more reversible; that is, road shocks are transmitted more completely from the wheels and must be overcome to a greater extent by the driver. This causes a dangerous situation on rough roads or when a front tire blows out, because the wheel may be jerked from the driver's hands.

Power steering gear was developed to solve the steadily increasing steering problems. Power steering was first applied to heavy trucks and military vehicles early in the 1930s, and hundreds of patents were granted for devices to help the driver turn the steering wheel. Most of the proposed devices were hydraulic; some were electrical and some mechanical. A pump driven by the engine maintains the hydraulic fluid under pressure. A valve with a sensing device allows the fluid to enter and leave the power cylinder as necessary.

Suspension systems. The suspended portion of the automobile is attached to the wheels by elastic members designed to cushion the impact of road irregularities. The nature of the attaching linkages and spring elements varies widely among United States, European, and Japanese automobiles. Riding comfort and the handling qualities of the vehicle are greatly affected by the functioning of the suspension system. Mechanical simplification is gained by connecting the front wheels with a rigid front axle just as the rear wheels are connected by the conventional rear axle housing. Several important advantages, however, are gained by so-called independent-suspension systems that permit the wheels to move independently of each other. The unsprung weight of the vehicle is decreased, softer springs are permissible, and front-wheel vibration problems are minimized. Independent front suspensions entirely replaced the rigid-axle type after World War II, and numerous independent rear suspensions came into use, first on European cars. Spring elements used for automobile suspension members, in increasing order of their ability to store elastic energy per unit of weight, are leaf springs, coil springs, torsion bars, rubber-in-shear devices, and air springs.

Independent suspension systems

The leaf spring, although comparatively inelastic, has the important advantage of accurately positioning the wheel with respect to the other chassis components, both laterally and fore and aft, without the aid of auxiliary linkages.

Some cars equipped with rear leaf springs employ a drive in which the axle housing is rigidly attached to the spring seats so that the driving torque reaction, the brake torque, and the driving force are all imparted to the spring. Those using rear coil springs may have a torque tube enclosing the drive shaft that takes these forces to its point of attachment at the front of the axle housing. A diagonal transverse rod connecting one outer end of the axle housing to the chassis frame at the opposite side is provided to position the axle laterally. Other types with open drive shafts provide for these forces by radius rods and other linkages.

An important factor in spring selection is the relationship between load and deflection known as spring rate, defined as the load in pounds divided by the deflection of the spring in inches. A soft spring has a low rate and deflects a greater distance under a given load. A coil or a leaf spring retains a substantially constant rate within its operating range of load and will deflect ten times as much if a force ten times as great is applied to it. The torsion bar, a long spring-steel element with one end held rigidly to the frame and the other twisted by a crank connected to the axle, can be designed to provide an increasing spring rate.

A soft-spring suspension provides a comfortable ride on a relatively smooth road, but the occupants move up and down excessively on a rough road. The springs must be stiff enough to prevent a large deflection at any time because of the difficulty in providing enough clearance between the sprung portion of the vehicle and the unsprung portion below the springs. Lower roof heights make it increasingly difficult to provide the clearance needed for soft springs. Road-handling characteristics also suffer because of what is known as sway, or roll, the sidewise tilting of the car body that results from centrifugal force acting outward on turns. The softer the suspension, the more the outer springs are compressed and the inner springs expanded. Front-end "dive" under brake action is more noticeable with soft front springs. Air springs offer several advantages over metal springs, one of the most important of which is the possibility of controlling the spring rate. Inherently, the force required to deflect the air unit increases with greater deflection, because the air is compressed into a smaller space, and greater pressure is built up, thus progressively resisting further deflection.

A combination hydraulic-fluid-and-air suspension system has been developed in which the elastic medium is a sealed-in, fixed mass of air, and no air compressor is required. The hydraulic portion of each spring is a cylinder mounted on the body sill and fitted with a plunger that is pivotally attached to the wheel linkage to form a hydraulic strut. Each spring cylinder has a spherical air chamber attached to its outer end. The sphere is divided into two chambers by a flexible diaphragm, the upper occupied by air and the lower by hydraulic fluid that is in communication with the hydraulic cylinder through a two-way restrictor valve. This valve limits the rate of movement of the plunger in the cylinder, since fluid must be pushed into the sphere when the body descends and returned when it rises. This damping action thus controls the motion of the wheel with respect to the sprung portion of the vehicle supported by the spring.

The amount of hydraulic fluid in each spring unit is varied by front and rear levelling valves that transfer fluid to and from a central hydraulic system in which pressure is maintained by an engine-driven pump. These leveling valves serve to maintain constant clearance between the wheels and the sprung portion of the vehicle under varying passenger and luggage load.

Tires. The pneumatic rubber tire is the point of contact between the automobile and the road surface. It functions to provide traction for acceleration and braking and limits the transmission of road vibrations to the automobile body. Inner tubes within tires were standard until the 1950s, when seals between the tire and the wheel were developed, leading to tubeless tires, now used almost universally.

Tire-tread design

Tire-tread designs are tailored for the characteristics of the surface on which the vehicle is intended to operate. Deep designs provide gripping action in loose soil and snow, while smooth surfaces provide maximum contact area for applications such as racing. Current passenger-car treads are a compromise between these extremes.

A typical tire casing is fabricated from layers, or plies, of varying proportions of rubber compounds reinforced with synthetic fibres or steel wire. The composition of the reinforcement and the angle of its application to the axis of the tread affect the ability of the tire to respond to sidewise forces created during cornering. They also affect harshness or vibration-transmission characteristics.

By 1970, longitudinal-, bias-, and radial-ply constructions were in use, with layers of two, four, or more plies, depending on the load capacity of the design. An additional factor relating to the load capacity of a particular construction is the pressure to which the tire is inflated. New designs also have lower height-to-width ratios to increase the road-contact area while maintaining a low standing height for the tire and consequently the car.

PROBLEMS OF AUTOMOTIVE DESIGN

Safety considerations. From its beginnings, the automobile exhibited a serious defect from the point of view of safety. Its speed and weight gave it an impact capacity both for its occupants and for pedestrians or other automobile passengers that produced great numbers of fatalities and serious injuries. Over the 20th century, the rates of death and injury declined in terms of passenger miles, yet, in both Europe and the United States and more recently in Japan and other countries, the total numbers rose because of the increased number of vehicles on the road. The worldwide pattern in the early 1970s varied little. Most fatal accidents occurred on either city streets or secondary rural roads; national expressway systems were relatively safer. Driver training, vehicle maintenance, highway improvement, and law enforcement were identified as key areas with potential for improving safety, but the basic design of the vehicle itself and the addition of special safety features were receiving increasing attention. Safety features of automobiles come under two distinct headings: accident avoidance and occupant protection.

Accident avoidance. Accident-avoidance systems are designed to help the driver maintain better control of his car. The dual-master-cylinder brake system is a good example. This protects the driver against sudden loss of brake-line pressure. Front and rear brake lines are separated so that if one fails the other continues to function. This system is sometimes coupled with an indicator lamp on the instrument panel to warn the driver if one section of the brake circuit has a failure.

Occupant protection. Systems to protect occupants in the event of an accident fall into four major classes: maintenance of passenger-compartment integrity, occupant restraints, interior-impact energy-absorber systems, and exterior-impact energy absorbers. Statistics indicate a far higher chance for survival among accident victims remaining inside the passenger compartment. Passenger-compartment integrity depends significantly on the proper action of the doors, which must remain closed in the event of an accident and must be sufficiently strong to limit deflection during intrusions. Door-latch mechanisms have been designed to resist forward, rearward, and sideward forces and incorporate two-stage catches, so that the latch may hold if the primary stage fails. Reinforcement beams in doors are designed to deflect impact forces downward to the more rigid frame structure below the door. Forces are directed through reinforced door pillars and hinges.

Occupant restraints are used to help couple the passenger to the car. They permit decelerating with the car rather than free flight into the car structure or into the air. Combination lap- and shoulder-belt systems are the most common restraint system. These consist of web fabrics that are able to withstand 6,000-pound (2,700 kilogram) test loading and are bolted to the car underbody and roof rail. Button-type latch release mechanisms for buckles, together with belt-operation reminder sys-

tems that detect unlatched belts, became legal requirements for all cars in the United States in 1972. This is intended to help offset the major drawback of belt systems or any other arrangements that require active participation of the occupant to insure proper operation. A 1970 survey indicated that 60 percent of car occupants do not use lap belts, and shoulder-belt usage is less than 3 percent. Another line of research aimed at overcoming this problem has centred on passive restraints that do not require any action on the part of the occupant. Experimental air-cushion restraint systems have been developed; an inflatable fabric bag pops out of the instrument panel and is filled with high-pressure gas within 40 milliseconds after an impact. The system is triggered by a sensor that is automatically actuated at impact. The occupant flies forward into the inflated cushion, which then controls his further motion and absorbs some of his energy. Energy is absorbed by forcing gas out of the cushion through a series of ports or orifices in the fabric. This system had not been accepted in the early 1970s because of developmental problems, including the possibility of inadvertent actuation, failure to actuate, the effect on occupants who are not sitting in the proper positions, the possible harmful effect on occupant hearing created by the high noise level of actuation, and the necessary safety precautions that must be taken when working with pyrotechnic devices and high-pressure gas cylinders.

Air-cushion restraint systems

Interior-impact energy-absorbing devices are designed to augment restraint systems by absorbing energy from the occupant while minimizing his injuries. The energy-absorbing steering column, introduced in 1967, is a good example of such a device (see Figure 12). In front-end collisions, the column resists being pushed rearward. The lower end compresses while the upper end is retained by U-shaped support brackets and bolts. A few milliseconds after impact the driver hits the steering wheel, and the support brackets break away, allowing the upper end of the column to move downward, absorbing energy by deforming a metal jacket surrounding the column. Instrument panels, windshield glass, and other surfaces that may be struck by an unrestrained occupant may be designed to absorb energy in a controlled manner.

By courtesy of General Motors Corporation

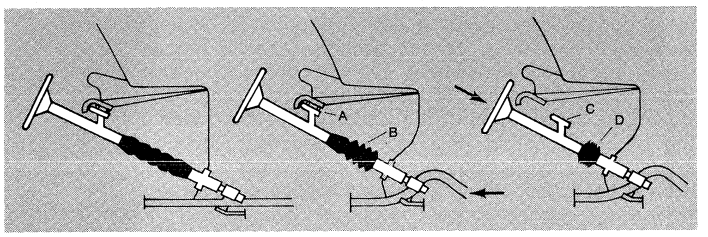


Figure 12: *Energy-absorbing steering column action* (see text). (Left) Assembly before impact. (Centre) An instant after impact. (A) Attachment bracket still holds; (B) lower steering column has started to collapse. (Right) An instant later. Impact of driver striking steering wheel breaks loose (C) attachment bracket and energy of impact is absorbed by further collapse of (D) lower steering column.

Exterior-impact-energy devices include the structural elements of the chassis and body, which may be tailored to deform in a controlled manner to decelerate the automobile more gradually and, as a result, leave less force to be experienced by the occupants. Stress risers in the form of section irregularities have been built into front frame members of some cars. These are designed to buckle under severe loads and absorb energy in the process. Retractable bumpers that have been developed absorb some of the impact energy of a collision as the bars move inward against the restraint of springs.

Pollution problems. Governments have also found a role in the automotive-air-pollution problem, to establish quality standards and to perform sample inspections to insure that standards are met. Standards have become progressively more stringent, and the equipment necessary to meet them has become more complex.

Positive crankcase ventilation (see Figure 13) was introduced in the United States in 1963, for the purpose of eliminating engine blow-by gas that formerly had been vented to the atmosphere as an emission source. This was achieved by routing the gas, along with crankcase-ventilation air, into the engine intake, where it could be recycled through the cylinders and properly burned. A small valve in the circuit was necessary to maintain the direction of flow. Beginning in 1966, cars increasingly were provided with air pumps to supply excess air to the exhaust stream near the exhaust valves. High gas temperatures in this zone promote thermal oxidation of hydrocarbons and carbon monoxide to form carbon dioxide and water.

By courtesy of General Motors Corporation

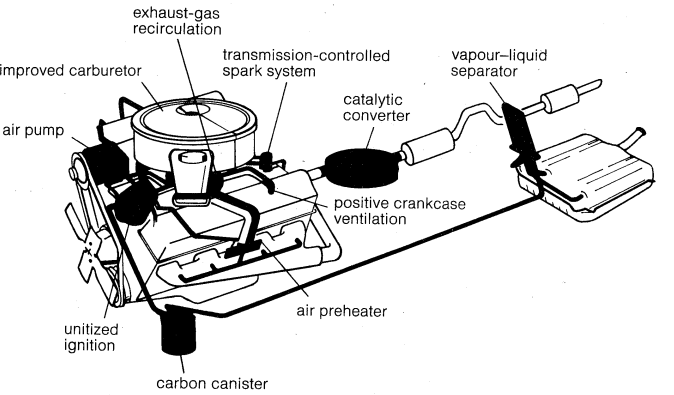


Figure 13: Emission-control systems for automobile engines (see text).

Various engine modifications that alter emission characteristics have been successfully introduced. These include adjusted carburetor air–fuel ratios, lowered compression ratios, retarded spark timing, reduced combustion chamber surface-to-volume ratios, and closer production tolerances. To improve drivability of some arrangements, preheated air from a heat exchanger on the exhaust manifold is ducted to the air cleaner.

Engine modifications to lower pollution

The undesired evaporation of gasoline hydrocarbons into the air has been controlled by sealing the gas tank and venting the tank through a liquid–vapour separator into a canister containing activated charcoal. During engine operation these vapours are desorbed and burned in the engine.

The effect of these controls was well established by the early 1970s. The average hydrocarbon emissions from equipped vehicles was reduced by about 87 percent. A 65 percent reduction in carbon monoxide was also obtained. Controls to reduce the formation of oxides of nitrogen brought a 30 percent reduction result. The chemical conditions that tend to reduce hydrocarbon and carbon monoxide sometimes tend to increase the formation of undesirable nitrous oxide. This presents a formidable challenge to research scientists and engineers.

Among emission-control devices under development in the early 1970s were catalytic converters (devices to promote combustion of unburned hydrocarbons in the exhaust), exhaust-gas-recirculation systems, manifold reactors, fuel injection, unitized ignition elements, and stratified charge combustion engines (see Figure 13).

A catalytic converter consists of an insulated chamber containing a porous bed of catalytic material through which hot exhaust gas must pass before being discharged into the air. The catalyst is in the form of pellets of a variety of metal oxides, which are heated by exhaust gas to about 900° F (500° C). At this temperature unburned hydrocarbons and carbon monoxide are further oxidized. Problems with catalysts involve their intolerance for leaded fuels and the need to prevent overheating.

Exhaust-gas recirculation is a technique to control oxides of nitrogen, which are formed by the chemical reaction of nitrogen and oxygen at high temperatures during combustion. Either reducing the concentrations of these elements or lowering peak cycle temperatures will reduce the amount of nitrous oxides produced. To achieve this, exhaust gas is piped from the exhaust manifold to the in-

take manifold. This dilutes the incoming fuel–air mixture and effectively lowers combustion temperature. The amount of recirculation is a function of throttle position but averages about 2 percent.

Manifold reactors are enlarged, insulated exhaust manifolds into which air is injected and in which exhaust gas continues to burn. The effectiveness of such units in experimental projects depends upon the amount of heat generated and the length of time the gas is within the manifold. Stainless steel and ceramic materials are used to provide durability at high operating temperatures (approaching 2,300° F [about 1,300° C]).

Fuel injection

Fuel injection, as a replacement for carburetion, has potential for reducing exhaust emissions. The precise metering of fuel for each cylinder provides a means of ensuring that the chemically correct air–fuel ratio is being burned in the engine. This eliminates cylinder-to-cylinder variations and the tendency of cylinders that are most remote from the carburetor to receive less fuel than is desired. A variety of metering and control systems are commercially available. A continuous-flow system was in use as early as 1957. Timed injection, in which a small quantity of gasoline is squirted into each cylinder or intake-valve port during the intake stroke of the piston, is employed on a number of European cars.

In one timed-injection system, individual pumps at each intake valve are regulated (timed) by an electronic controller. This unit monitors intake vacuum, engine temperature, ambient-air temperature, and throttle position and adjusts the time and duration of injection accordingly.

Another direction of emission control involved ignition performance. Unitized or solid-state ignition circuitry was introduced on some cars in the late 1960s to improve starting and ignition reliability and durability.

Another approach is the stratified charge engine, a variation from conventional cylinder combustion. Fuel is injected into a combustion-chamber pocket, and the nonhomogeneous, stratified charge is spark ignited. Operation is possible at very lean fuel–air ratios, thus permitting high thermal efficiency at light engine loads. This provides excellent reductions in exhaust hydrocarbons, carbon monoxide, and oxides of nitrogen. The primary problem with the system is to make it function over a wide range of speeds and loads with good transient response.

Future systems. Expansion of the total potential automotive market in the future and concern for the environment may be expected to change cars of the future. Special-purpose vehicles designed for specific urban or rural functions, with appropriate power systems for each type of use, may be needed. Possibilities include electric, steam, gas-turbine, nuclear, and other power sources.

In the early 1970s electric propulsion was possible for relatively short-range vehicles using power from batteries or fuel cells. Such systems powered vehicles used on the moon in 1971.

Conventional storage-battery systems do not possess high power-to-weight ratios for acceleration or energy-to-weight ratios for driving range to match gasoline-powered general purpose vehicles. Special-purpose applications, however, such as urban delivery vehicles, may be practical because of the excellent low-emission characteristics of the system. It may also be coupled with gasoline or other power sources to form a hybrid power plant with electric operation in cities and gasoline operation in less congested areas. Storage batteries having high power-to-weight ratios are under active development and may stimulate the production of electric cars in the future.

Steam power plants have been re-examined in the light of modern technology and new materials. The continuous-combustion process used to heat the steam generator offers potentially improved emission characteristics. Closed systems using vapours other than steam are also under active development, but further work is necessary to improve power-plant size, weight, and working fluid characteristics.

Gas turbines have been tested extensively and are expected to power high-speed, long-haul trucks and buses in the 1970s. Turbines have good torque characteristics, operate on a wide variety of fuels, have high power-to-weight ratios, and offer quiet operation. Studies have shown that the advantages of the system are best realized in heavy-duty vehicles operating on long, nearly constant speed runs. Successful designs require regenerative systems to recover energy from hot exhaust gas and transfer it to incoming air. This improves fuel economy, reduces exhaust temperatures to safer levels, and eliminates the need for a muffler in some designs.

Gas turbines

Nuclear energy offers the advantage of extremely low fuel weight. The obstacle for automotive use, however, is the great weight and volume of shielding required to protect the occupants from excessive nuclear radiation. This source can probably best be applied indirectly to produce stable chemical compounds or electricity, which could be converted to automotive uses.

A number of other energy-conversion systems have been studied for future automotive applications. Many of these are variations of engine combustion cycles such as turbocharged four-stroke, two-stroke, and diesel (two- and four-stroke). Free-piston and other geometric configurations, including rotary engines, offer other possibilities. Great interest has been shown in the Wankel, a type of rotary engine invented by Felix Wankel of Germany, and several auto manufacturers are conducting tests on the motor. In 1971, at least one company (Japanese) was manufacturing a Wankel-powered automobile.

Because the automobile promises to remain a part of the overall mobility system of modern man for some time to come, it requires continuing improvement in safety and emission control as well as in its older values of performance characteristics and economic contribution.

(O.C.C./G.C.C.)

BIBLIOGRAPHY

Historical development: G.N. GEORGANO (ed.), *The Complete Encyclopedia of Motorcars, 1885–1968* (1968), the definitive work on the subject, copiously illustrated, covering some 4,000 makes of automobile; THE AUTOMOBILE QUARTERLY, *The American Car Since 1775* (1971), an exhaustive treatment of some 5,000 makes of car produced or projected in North America; LAWRENCE J. WHITE, *The Automobile Industry Since 1945* (1971), based on critical, scholarly research; NEW ENGLISH LIBRARY, *History of the Motorcar* (1971), a lavishly illustrated, brightly-written treatment of the self-propelled vehicle from Leonardo da Vinci to the present: T.R. NICHOLSON, *The World's Motor Museums* (1970), brief descriptions of some 170 museums wholly concerned with the motorcar.

Modern automobiles: WILLIAM H. CROUSE, *Automotive Mechanics*, 5th ed. (1965) and *Automotive Engine Design* (1970), provide a complete introductory course on each subject and cover the theory of operation and construction, maintenance, repair, and adjustment of automotive components, as well as their design. Important textbooks on the subject include: LESTER C. LICHTY, *Combustion Engine Processes* (1967), a technical book dealing particularly with fuels and combustion, carburetion and fuel injection, fuel properties, manifolding, and ignition, including the new transistorized systems; E.F. OBERT, *Internal Combustion Engines*, 3rd ed. (1968), which presents the fundamental and practical development of the application of science in the design of combustion engines, including material on pollution; M.W. STOCKEL, *Auto Mechanics Fundamentals* (1969), which contains basic information about various devices; C.F. and E.S. TAYLOR, *The Internal-Combustion Engine*, 2nd ed. (1961), which provides analytical methods that lead to a basic understanding of engine behaviour; and COLIN CAMPBELL, *The Sports Car: Its Design and Performance*, 2nd ed. rev. (1959), a discussion of the design and performance primarily of high-powered vehicles, sometimes called muscle cars, but with much broader fundamental information also included. Periodicals that provide much current information include: *Auto World*, a widely recognized semi-monthly Detroit publication largely dealing with developments in the automotive industry; *Automobile Engineer*, a monthly British publication that presents much international information; *Automotive Engineering*, the monthly journal of the Society of Automotive Engineers, a standard reference magazine in the industry; and *Automotive Industries*, a comprehensive semi-monthly magazine.

(O.C.C./G.C.C./K.W.P.)

Automotive Industry

The term automotive industry is generally applied to all those companies and activities involved in the manufacture of motor vehicles, including most components, such as engines and bodies, but excluding tires, batteries, and fuel. The industry's principal product is passenger automobiles; commercial vehicles, though important, are secondary, and are discussed in the article TRUCKS AND BUSES.

The history of the automobile industry, while brief compared with many others, has exceptional interest because of its implications for 20th-century history. Though the automobile originated in Europe, the United States completely dominated the world industry for the first half of the century through the invention of mass-production techniques. In the second half of the century the situation altered sharply as the western European countries and Japan became major producers and exporters.

HISTORY OF THE AUTOMOTIVE INDUSTRY

Though steam-powered road vehicles were produced earlier, the origins of the automotive industry are rooted in the development of the gasoline engine (*q.v.*) in the 1860s and 1870s, principally in France and Germany. (For an account of the invention of the automobile, see AUTOMOBILE.) By the turn of the 20th century, German and French manufacturers had been joined by British, Italian, and United States makers.

Developments before World War I. The typical early automobile companies were small shops, hundreds of which produced a few hand-manufactured cars, and nearly all of which abandoned the business soon after going into it. The handful that survived into the era of large-scale production had certain characteristics in common.

Types of early shops

First, they fell into one of three well-defined categories: they were makers of bicycles, such as Opel in Germany and Morris in Great Britain; builders of horse-drawn vehicles, such as Durant in the United States; or, most frequent of all, they were machinery manufacturers—these included makers of stationary gas engines (Daimler of Germany, Lanchester of Britain, Olds of the United States), makers of marine engines (Vauxhall of Britain), makers of machine tools (Leland of the United States), makers of sheep-shearing machinery (Wolseley of Britain), makers of washing machines (Peerless of the United States), makers of sewing machines (White of the United States), and makers of woodworking and milling machinery (Panhard and Levassor of France). One company (Pierce of the United States) had made birdcages. Two notable exceptions to the general pattern were Rolls-Royce in Britain and Ford in the United States, both of which were founded by partners combining engineering talent and business skill.

In the United States practically all of the producers were assemblers who put together components and parts manufactured by separate firms. The assembly technique was also a method of financing. It was possible to begin building motor vehicles with a minimal investment of capital by buying parts on credit and selling the finished cars for cash; the cash sale from manufacturer to dealer has been integral in the marketing of motor vehicles in the United States ever since. European automotive firms of this period tended to be more self-contained.

The pioneer automobile manufacturer not only had to solve the technical and financial problems of getting into production, but also had to make a basic decision about what to produce. After the first success of the gasoline engine, there was widespread experimenting with steam and electricity; for a brief period the electric automobile actually enjoyed the greatest acceptance because it was quiet and easy to operate, but the limitations imposed by battery capacity proved competitively fatal.

Steam power, a more serious rival, was aided by the general adoption of the so-called flash boiler after 1900, a modification that provided for a rapid raising of steam. The steam car was easy to operate because it did not require an elaborate transmission. On the other hand, high steam pressures were needed to make the engine light enough for use in a road vehicle, requiring expensive construction and difficult maintenance. By 1910 most manufacturers of steam vehicles had turned to gasoline power. The Stanley brothers in the United States, however, continued to manufacture steam automobiles until the early 1920s.

Patent controversies

As often happens with a new technology, the automotive industry experienced patent controversies in its early years. Most notable were two long-drawn-out court cases in Britain and the United States, in each of which a promoter sought to gain control of the new industry by filing comprehensive patents. In Britain the claim was rejected by the courts in 1901, five years after the patent application. In the United States a longer, more complex legal battle ultimately resulted in an agreement among manufacturers for cross-licensing patents (1915).

Mass production. The outstanding contribution of the automotive industry to technological advance was the introduction of full-scale mass production, a process combining precision, standardization, interchangeability, synchronization, and continuity. Mass production was a United States innovation. The United States, with its large population, high standard of living, and long distances, was the natural birthplace of the technique, which had been partly explored in the 19th century. Though Europe had shared in the experimentation, the United States role was underlined in the popular description of standardization and interchangeability as "the American system of manufacture." The fundamental techniques were known, but they had not previously been applied to the manufacture of a mechanism so complex as a motor vehicle (see MASS PRODUCTION).

The kind of interchangeability achieved by the "American system" was dramatically demonstrated at the British Royal Automobile Club in London in 1908, when three Cadillac cars were disassembled, the parts mixed together, 89 parts removed at random and replaced from dealer's stock, the cars assembled, and driven 500 miles without trouble. Henry M. Leland, founder of Cadillac Motors and the man responsible for this feat of showmanship, later enlisted the aid of a noted electrical engineer, Charles F. Kettering, in developing the electric starter, a significant innovation in promoting the acceptability of the gasoline-powered automobile.

Ford and the assembly line. The mass-produced automobile is generally and correctly attributed to Henry Ford, but he was not alone in seeing the possibilities in a mass market. Ransom E. Olds made the first major bid for the mass market with a famous curved-dash Oldsmobile buggy in 1901. While the first Oldsmobile was a popular car, it was too lightly built to withstand rough usage. The same defect applied to Olds's imitators. Ford, more successful in realizing his dream of "a car for the great multitude," designed his car first and then considered the problem of producing it cheaply. The car was the so-called Model T, the best known motor vehicle in history. It was built to be durable for service on the rough American country roads of the period, economical to operate, and easy to maintain and repair. It was first put on the market in 1908, and over 15,000,000 were built before it was discontinued in 1927.

The Model T

When the design of the Model T proved successful, Ford and his associates turned to the problem of producing the car in large volume and at low unit cost. The solution was found in the moving assembly line, a method first tested in the assembly of magnetos. After more experimentation, in 1913 the Ford Motor Company displayed to the world the complete assembly-line mass production of motor vehicles. The technique consisted of two basic elements: a conveyor system and the limitation of each worker to a single repetitive task. Despite its deceptive simplicity, the technique required elaborate planning and synchronization.

The first Ford assembly line permitted only very minor variations in the basic model, a limitation which was compensated for by the low cost. The price of the Model T touring car dropped from $950 in 1909 to $360 in 1916, and an incredible $290 in 1926. By that time Ford was producing half of all the motor vehicles in the world.

Spread of mass production. This success inspired imitation and competition, but Ford's primacy remained unchallenged until he himself lost it in the middle 1920s by refusing to recognize that the Model T had become outmoded. More luxurious and better styled cars appeared at prices not greatly above that of the Model T, and these were increasingly available to low-income purchasers through a growing used-car market. Abroad, William R. Morris (later Lord Nuffield) in Britain undertook to emulate Ford as early as 1912, but found British engineering firms reluctant to commit themselves to large-scale manufacture of automotive parts. Morris in fact turned to the United States for his parts, but these early efforts were cut short by World War I. In the 1920s Morris resumed the production of low-priced cars, along with his British competitor Herbert Austin and André-Gustave Citroën and Louis Renault in France. British manufacturers had to face the problem of a tax on horsepower, calculated on a formula based on bore and the number of cylinders. The effect was to encourage the design of small engines that had cylinders with narrow bore and long stroke, in contrast to the wide-bore, short-stroke engines favoured elsewhere. This handicapped the sale of British cars abroad and kept the level of production from growing. It was not until 1934 that Morris Motors finally felt justified in installing a moving assembly line; the Hillman Company had preceded Morris in this by a year or two.

The effect of the horsepower tax on design

Large-scale organization. Though the appearance of mass production in the automotive industry coincided with the emergence of large-scale business organization, the two had originated independently. They were, nevertheless, related, and reacted on each other as the industry expanded. Only a large firm could make the heavy investment in plant and tooling that the assembly line required, and Ford was already the largest single American producer when it introduced the technique. The mass producer in turn enjoyed a cost advantage that tended to make it increasingly difficult for smaller competitors to survive. Exceptions can be noted along the way, but the trend has been consistent.

General Motors. General Motors, which ultimately became the world's largest automotive firm, and in fact the largest privately owned manufacturing enterprise in the world, was founded in 1908 by William C. Durant, a carriage manufacturer of Flint, Michigan. In 1904 he assumed control of the ailing Buick Motor Company and made it one of the principal American producers. Durant developed the idea for a combination that would produce a variety of models and control its own parts producers. As initially formed, General Motors included four major vehicle manufacturers (Buick, Cadillac, Oldsmobile, Oakland) and an assortment of smaller firms. The combine ran into financial trouble in 1910 and was reorganized by a financial syndicate. A similar combination, the United States Motor Corporation, formed in 1910, collapsed in 1912 and was reorganized as the Maxwell Motor Company. General Motors survived. A new reorganization took place with the return of Durant, with du Pont Company backing, to control in 1916.

Rise of the Big Three. At the end of World War I, Ford was the colossus, dominating the automotive scene with the Model T not only in the United States but through branch plants throughout the world. British Ford was the largest single producer in the United Kingdom. General Motors was emerging as a potential major competitor in the United States. No other automotive firms of comparable size existed.

The next decade brought a striking transformation. The depression of 1921 had far-reaching effects on the American automotive industry; General Motors was plunged into another financial crisis. Alfred P. Sloan became president of the corporation in 1923 and raised it to its unchallenged first place in the industry. Among other steps he gave General Motors a staff-and-line organization, with autonomous manufacturing divisions, which facilitated management of a large corporate structure and became the model for other major automotive combinations. Ford also went through a crisis because the crash caught him involved in the construction of a large new plant (River Rouge) and in the process of buying out his stockholders. Ford weathered the storm (though many of his dealers, unable to sell cars and not permitted to return them, went out of business), but the Ford Motor Company had reached its crest.

The third member of the American "Big Three" was created at this same time. When the Maxwell Motor Company failed in the 1921 depression, Walter P. Chrysler, formerly of General Motors, was called in to reorganize it. It became the Chrysler Corporation in 1925 and grew to major proportions with the acquisition of the Dodge firm in 1928. When Ford went out of production in 1927 to switch from the Model T to the Model A (a process taking 18 months), Chrysler was able to break into the low-priced car market with the Plymouth.

Chrysler's entry

The independents. By 1929 the Big Three had three-quarters of the American market for motor vehicles with most of the remainder divided among the five largest independents (Hudson, Nash, Packard, Studebaker, Willys-Overland). In less than ten years the number of automobile manufacturers in the United States dropped from 108 to 44. Some of the minor firms have technological or personal interest, like Nordyke and Marmon, makers of luxury cars, or E.L. Cord, who marketed a front-wheel-drive car in 1929. The depression years of the 1930s eliminated all but the largest independents and increased still further the domination of the Big Three. Motor vehicle production declined from a peak of over 5,000,000 in 1929 to a low of just over 1,000,000 in 1932. It rose again steadily but had not returned to the 1929 figure when World War II broke out.

Yet, while these years were difficult economically, they saw some significant developments within the industry. Greater emphasis was placed on style in passenger car design, with the general trend in the direction of incorporating body, bumpers, and mudguards into a single pattern of smoothly flowing lines. A number of technical features came into general use: the V-8 engine, introduced by Ford in 1932; three-point engine suspension; free-wheeling (permitting the car to coast freely when the accelerator was released); overdrive (a fourth forward speed); and, on a limited scale, automatic transmission.

Growth in Europe. The period from 1919 to 1939 also witnessed significant growth in automobile manufacturing in Europe, although on a considerably smaller scale than in North America. The European industry was moving in the same directions as the American, toward a mass market for motor vehicles, but it made slower progress for a variety of reasons: lower living standards with less purchasing power, smaller national markets, and more restrictions in tax and tariff policies. Nevertheless, the same trend toward concentration is discernible. British automotive production rose from 73,000 in 1922 (both private and commercial vehicles) to 239,000 in 1929, while the number of producers declined from 90 to 41. Three firms (Austin, Morris, Singer) controlled 75 percent of the British market in 1929.

The apparent analogy to the American experience was temporary. British production was not yet at the level at which the economies of scale gave the larger firms as commanding a lead as in the United States, and there were other factors to create a somewhat different situation. During the 1930s British automotive production continued to increase steadily, in contrast to the American, so that the smaller concerns were not forced to compete for a shrinking market. With output reaching almost half a million in 1937, at the end of the decade there were six major British producers instead of three: Morris, Austin, Standard, Rootes, Ford, and Vauxhall. The last two represented entry by American firms. Vauxhall was bought by General Motors in 1925; Ford had been in Britain since 1911, had lost ground in the 1920s, and had recovered. The Rootes Group, based on Hillman and Humber, was a combine formed by a family that had built a large automobile sales concern and moved from sales to production. The replacement of Singer by Standard was simply the rise of one company and the decline of another, as evidence that open competition could still change the structure of the British automotive industry.

British production gains in the 1930s

In France, three major firms, Peugeot, Renault, and Citroën, emerged in the 1920s. Citroën accounted for 40 percent of French automotive production in 1925, but had reached that dominating position at the cost of financial stability. When André Citroën died before the decade ended, his company came into the hands of Michelin Tire. A new French firm, Simca, rose to prominence in the 1930s. The German automobile industry suffered from the dislocation of World War I and Germany's subsequent economic difficulties. The major developments of the 1920s were the merger of Daimler and Benz in 1926, after the founders of both firms had died (their bitter rivalry for the distinction of being the inventor of the gasoline automobile made any such union during their lifetimes unthinkable), and the entry of General Motors into the German scene through the acquisition of the Opel company in 1929. The Germans were ardent admirers of Henry Ford and his methods, which they termed *Fordismus*, but Ford never succeeded in becoming a power in the German automotive world. During the 1930s the Nazi regime sought to emulate Ford by undertaking mass production of a low-priced car, the Volkswagen, but the onset of war interrupted this project. Italian automobile manufacturers gained a reputation for highly engineered sports and racing cars, but Italy had no mass market and therefore only small-scale production at that time.

The automotive industry at war. During World War I the productive capacity represented by the automotive industry first demonstrated its military value. Motor vehicles were used extensively for transport and supply, and in addition, automotive plants could readily be converted into facilities for the manufacture of military equipment, including tanks and aircraft. For all of the belligerents the conversion of automotive facilities was an afterthought, improvised after the beginning of hostilities, and the United States industry, involved only for a short time, never fully utilized its capacity.

"Shadow factories"

More preparation was made to use the resources of the various automotive industries as World War II approached. The British government built "shadow factories" adjacent to their automotive plants, equipped to go into military production (principally aircraft) when war came, with managerial and technical personnel drawn from the automotive industry. France attempted conversion, but only belatedly and inefficiently. The German automotive industry, which built the military vehicles needed for blitzkrieg warfare, was not fully converted to military production until 1943. In the United States preparation for industrial mobilization was negligible until 1940; in fact, there was no serious effort even to restrict civilian automobile production until after the attack on Pearl Harbor. The American automotive industry, nevertheless, represented such a concentration of productive capacity and skill that once its resources had been harnessed to war production its contribution was stupendous. Between 1940 and 1945 automotive firms made almost $29,000,000,000 worth of military materials, a fifth of the entire United States output. The list included 2,600,000 military trucks and 660,000 jeeps, but production extended well beyond motor vehicles. Automotive firms provided one-half the machine guns and carbines made in the United States during the war, 60 percent of the tanks, all the armoured cars, and 85 percent of the army helmets and aerial bombs.

It had been assumed that automotive facilities could be readily converted to aircraft production, but this proved more difficult than anticipated. Automobile assembly plants did not readily accommodate airframes, nor could an automobile engine factory be converted without substantial modification. In time these problems were resolved and automobile companies contributed significantly to aircraft production.

Britain was better prepared to use the resources of its automotive industry, at that time the world's second largest. The shadow factories became operative, and Austin, Morris, Standard, Daimler, Ford, and Rootes participated in filling the wartime demand for aircraft and aircraft engines. Leyland and Vauxhall built tanks. Lord Nuffield made a notable contribution to the production effort by establishing a system for repairing aircraft, employing the sales and service organization of Morris Motors, and it was subsequently extended to a large number of small contractors.

The automotive industries of the other belligerents were smaller in scale and the facilities for armaments manufacture proportionately greater than in the United States or Great Britain. Consequently, the automotive firms in these countries were concerned chiefly with meeting the insatiable demand for vehicles. The various Ford properties that came under German control and Volkswagen, which turned out the German equivalent of the jeep, were employed in this manner. Renault, a tank manufacturer since World War I, built tanks for France and later for Germany.

The automotive industry since 1945. The years since World War II ended have seen two major developments in the automotive industry. First, there has been a spectacular expansion on a worldwide scale. In 20 years (1950–1970) world output of motor vehicles trebled, from 10,000,000 to 30,000,000 a year; it is significant that the bulk of that increase took place outside the United States, whose share of world automotive production fell from almost 76 percent in 1950 to 28 percent in 1970. The increase in output among the principal producers is shown in Table 1. This phenomenal growth of motor vehicle production was both an index and a cause of the remarkable economic recovery of western Europe and Japan from the devastation of the war.

Table 1: Comparative Totals, Motor Vehicle Production

	1935	1950	1970
U.S.	3,971,241	8,003,056	8,283,949
Japan	...	31,597	5,289,157
West Germany	242,934*	306,064	3,842,247
France	165,000†	357,700	2,750,086
U.K.	416,899	783,672	2,098,496
Italy	50,493	127,847	1,854,252
Canada	172,877	390,102	1,193,572
U.S.S.R.	...	403,700†	844,300†
Sweden	3,318	17,553	310,141

*All Germany. †Estimated.
Source: *Automobile Facts and Figures, 1971; The Motor Industry of Great Britain, 1965*, 1966.

The second development has been the continuation and intensification of the trend to concentration in the automotive industry, a trend common to all producing countries. The Soviet Union is a special case, of course, since the state controls all production and distribution of motor vehicles.

Concentration in the United States. At the end of World War II the American automobile industry found itself with its physical plant intact; it had, in fact, been increased because of facilities built by the government for military production. At the same time there was a great pent-up demand for motor vehicles. This situation brought about an ambitious attempt by a newcomer to enter the industry, the first such effort on a substantial scale since the creation of the Chrysler Corporation. The new venture, Kaiser-Frazer, took over the Ford wartime bomber assembly plant at Willow Run, Michigan, as its main assembly centre and produced a range of small and large cars. But it had no technical advance to offer and lacked the resources in capital, management, and sales outlets to compete with the established firms and after several years abandoned the field. The remaining United States independents sought survival in merger. One merged company, Studebaker-Packard, abandoned the American automobile market in 1964 and moved its operations to Hamilton, Ontario. A second, American Motors, managed to remain competitive by introducing the first American "compact," named the Rambler. In competition with the high-powered, ornately decorated cars that dominated the American scene in the 1950s, it succeeded well enough so that the other companies had to

Kaiser-Frazer

meet the competition by producing compacts of their own. The popularity of small foreign cars on the American market also stimulated the trend to compacts.

By the beginning of the 1970s the automotive industry in the United States was concentrated in four major firms (General Motors, Ford, Chrysler, and American Motors, in that order), and one important manufacturer of commercial vehicles, the White Motor Company, which during the 1950s had absorbed most of the separate truck manufacturing firms. A few producers of specialized vehicles remained, along with an assortment of companies making automotive parts and components. Technically, the 1950s and 1960s were marked by improvement and refinement rather than by any important innovation. Diesel engines were increasingly used on trucks and buses. Automatic transmission became virtually standard equipment for passenger cars, and power brakes and power steering found widespread acceptance. In the early 1960s Chrysler experimented with a gas turbine engine for passenger automobiles, but it had too many technical problems for general use.

Styling became increasingly important in automotive design as a marketing device. The general trend in styling became established late in the 1920s when cars began to lose their square, boxlike lines and develop flowing curves. The flow in time encompassed both body and chassis, integrating such formerly separated features as mudguards, running boards, and bumpers. A combination of pressures led American cars in the 1950s to extravagances in the use of chrome and exaggerated tail fins, which ended when the public found the simpler lines of imported cars more attractive.

Growth in Europe. In Europe, motor vehicles were recognized as an export item that could help restore war-shattered economies. Britain, for example, earmarked over one-half its automotive output for export and restricted domestic purchases for several years after the war. In addition, the horsepower tax was abandoned to enable British manufacturers to build profitably for the world market. The most popular British designs (excluding such a specialized luxury vehicle as the Rolls-Royce) continued to be light cars, including a number of models with an ingenious front-wheel drive. The trend to concentration led to the merger of Morris and Austin in 1952, which became the British Motor Corporation, Ltd., a combine that accounted for about two-fifths of Britain's motor vehicle production. Another British combine was formed around Leyland Motors, which had grown to be Britain's largest manufacturer of commercial vehicles and became a power in the passenger car field by acquiring Standard-Triumph and Sunbeam in the 1950s. Leyland and the British Motor Corporation united in 1968 as the British Leyland Motor Corporation, Ltd., a move sanctioned by the government to forestall possible American domination of the British automobile industry. Except for Rolls-Royce, whose automobile production was only a very small part of the company's business, British automobile output was controlled by four firms: British Leyland Motors, Ford, Vauxhall, and Rootes, which came under Chrysler control in 1967.

Introduction of the Volkswagen

The revival of the German automobile industry from almost total destruction was a spectacular feat, with most emphasis centring upon the Volkswagen. At the end of the war the Volkswagen factory and the city of Wolfsburg were in ruins. Restored to production, the plant, in a little more than a decade, was producing one-half of Germany's motor vehicles and had established a strong position in the world market. Breaking away from what had become standard design, the Volkswagen used a four-cylinder air-cooled engine at the rear of the car. The feature proved popular enough to induce other manufacturers to imitate it, but with qualified success. The rear-engine arrangement did not require a long drive shaft but involved problems of weight distribution and consequently of stability. The Volkswagen also dispensed with the annual model change that had become customary with other automobile manufacturers. The company was founded by the German government, but in the 1960s the government divested itself of 60 percent of its interest by selling stock to the public, an unusual case of denationalization in an era when nationalization of industry was far more common.

Fiat (Fabbrica Italiana Automobili Torino), a firm founded in 1899 but without a mass market until the 1950s, dominated Italian automotive production. The French industry was centred on Renault, Peugeot, Citroën, and Simca. Renault was nationalized in 1944 and merged with Peugeot in the 1960s. Simca became a Chrysler property in 1958, and Fiat acquired a 15 percent interest in Citroën. Although Sweden was a relatively small producer, Swedish-built Volvos and Saabs became a factor in the world market during the 1960s.

Japan. The most spectacular rise in automotive production was Japan's. From a negligible position in 1950 Japan climbed in 20 years to become the world's second largest producer of motor vehicles, passing both Britain and Germany. Although the rest of the world knew the Japanese industry through two passenger cars, the Datsun and the Toyota, half of Japan's motor vehicle output consisted of trucks and buses. Japan's post-World War II economic growth created a demand for commercial highway transportation before there was a mass market for private automobiles. Passenger cars now constitute 70 percent of Japanese production.

THE MODERN INDUSTRY

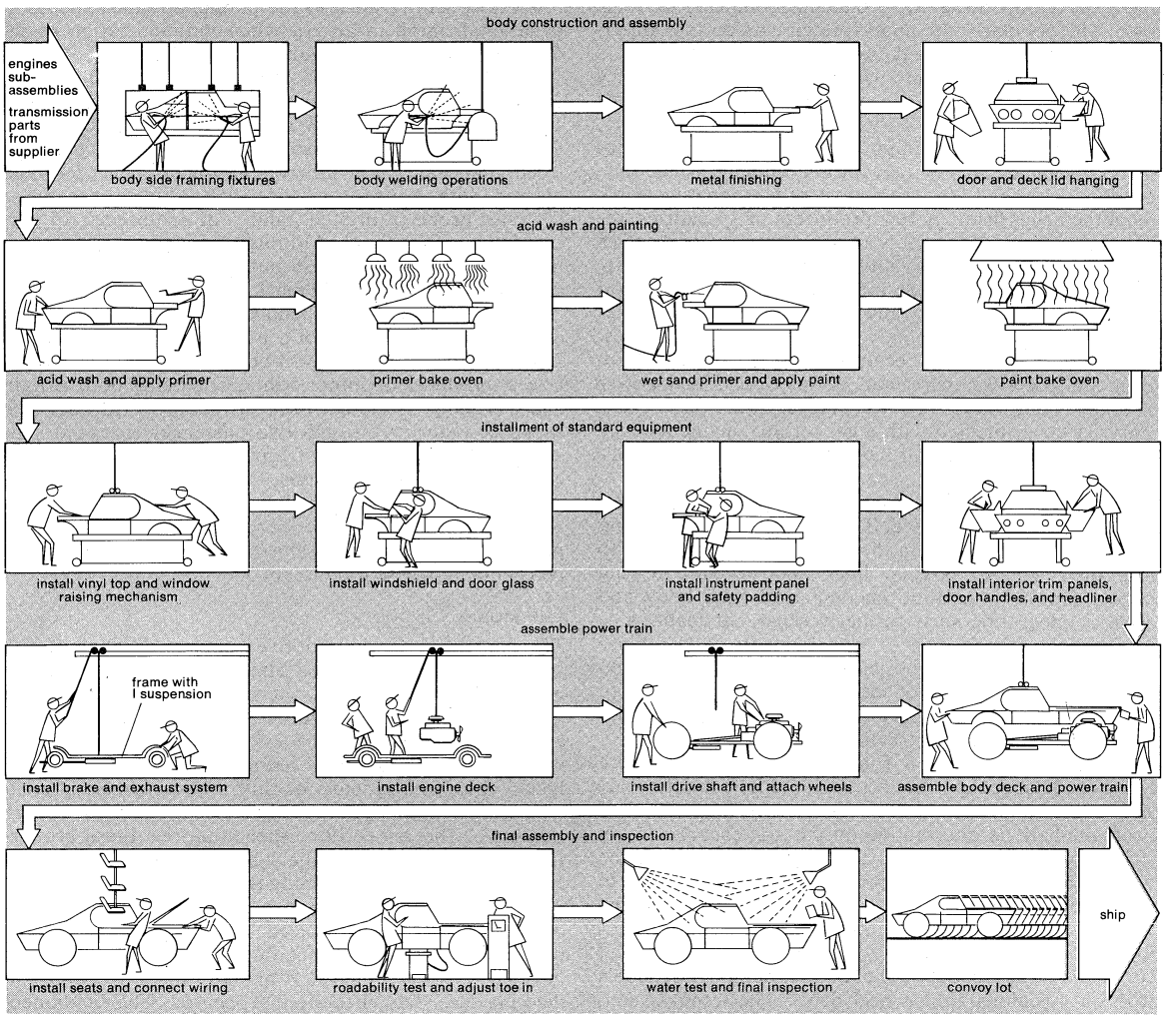
The contemporary automotive industry is a giant and still growing. In the United States, it is the largest single manufacturing enterprise in terms of total value of products, value added by manufacture, and number of wage earners employed. One business in six is dependent on the manufacture, distribution, servicing, or use of motor vehicles; sales and receipts of automotive firms represent 17 percent of the country's wholesale business and 24 percent of the retail. For other countries these proportions are somewhat smaller, but western Europe and Japan are rapidly approaching the United States level.

Concentration. The trend to concentration in the industry has already been traced. In each of the major producing countries the output of motor vehicles is in the hands of a few very large firms, and the small independent producer has virtually disappeared. The fundamental cause of this trend is the technique of mass production, which requires a heavy investment in equipment and tooling and is therefore feasible only for a large organization. Once the technique is instituted, the resulting economies of scale give the large firm a commanding advantage, provided of course that the market can absorb the number of units that must be built to justify the investment. Although the numbers required are difficult to determine exactly, the best calculations, considering both the assembly operation and the stamping of body panels, place the optimum output between 200,000 and 400,000 cars per year for a single plant.

Significance of size of production run

The structural organization of these giant enterprises shows, despite individual variation, a recognizable tendency toward the pattern first adopted by General Motors in the 1920s. There is a central organization with an executive committee responsible for overall policy and planning. The operating divisions are semi-autonomous, each reporting directly to the central authority but responsible for its own internal management. In some situations the operating divisions even compete with each other. The Ford Motor Company was consciously reorganized on the General Motors pattern after World War II; other American automotive firms have similar structures, as do British Leyland Motors and Rootes.

In addition, the largest producers decentralize their manufacturing operations by means of regional assembly plants. These permit the central factory to ship frames and components rather than complete automobiles to the areas served by the assembly plants, effecting substantial savings in transportation costs. This system was worked out for the Ford company as far back as 1911. It was particularly well suited to the extensive market areas of the United States and Canada. The local assembly plant has also been generally employed by United States and European firms to penetrate foreign markets.



Automobile assembly line.
Adapted from *Automotive Industry* (October 1970)

Diversity of product line. The immense resources of the automotive industries in production facilities and technical and managerial skills have been devoted predominantly to the building of motor vehicles, but there has been a consistent strong incentive to extend into related products and occasionally into operations whose relationship to automobiles is remote. The Ford Motor Company has long manufactured tractors.

Ford and General Motors have both been involved in aircraft manufacture. Ford built the famous Ford trimotor, an all-metal transport, from 1924 to 1932. For about 20 years General Motors had an interest in North American Aviation and Bendix Aviation and continues to be an important manufacturer of aircraft engines. It also makes refrigerators, diesel engines, diesel-powered railway locomotives, and earth-moving equipment. All these operations resulted from the acquisition of companies that were actively or experimentally engaged in these fields.

American Motors manufactures electric refrigerators and other household appliances as a result of merging with the Kelvinator Corporation. Ford moved into electronics and astronautics in 1961 by acquiring the Philco Corporation, a manufacturer of radios, televisions, and electronic equipment, and by organizing an Aeronutronic Division. In Europe the same trend is observable. Fiat is an industrial giant that not only makes 90 percent of all Italian cars but also builds aircraft, railway locomotives, marine engines, and electric refrigerators. In addition, it has a division engaged in major civil engineering and construction works. Renault builds tractors and machine tools. British Leyland Motors makes tractors, aircraft engines (through its Alvis division), refrigerators, and household appliances. The major Japanese automobile manufacturers, Toyota and Nissan (Datsun), have affiliates that make railway rolling stock, aircraft and aircraft engines, road rollers, and diesel engines.

Market surveys

New car development. The process of putting a new car on the market has become largely standardized. If a completely new model is contemplated, the first step is a market survey. Since there may be an interval of five years between this survey and the appearance of the new car in the dealers' showrooms, there is a distinct element of risk, as illustrated by the case of the Ford Motor Company's Edsel. (Market research had indicated a demand for a car in a relatively high price range, but by the time the Edsel appeared, both public taste and economic conditions had changed.) Conferences then follow among engineers, stylists, and executives to agree on the basic design. The next stage is a mock-up of the car, on which revisions and refinements can be worked out.

Manufacturing processes. The bulk of the world's new cars come from the moving assembly line introduced by Ford, but much more refined and elaborated (see illustration). The first requisite of this process is an accurately controlled flow of materials into the assembly plants. No company can afford either the money or the space to stockpile the parts and components needed for any extended period of production. Interruption or confusion in the flow of materials quickly stops production. Ford dreamed of an organization in which no item was ever at rest from the extraction of raw material to the vehicle's completion, a dream not yet realized.

The need for careful control over the flow of materials is an incentive for automobile firms to manufacture their own components, sometimes directly but more often through subsidiaries. Yet complete integration does not

exist and is not desired. As a general rule, tires, batteries, and dashboard instruments are procured externally. In addition, and for the same reasons, the largest companies support outside suppliers even for items of inhouse manufacture. First, it may be more economical to buy externally than to provide additional internal facilities for the purpose. Second, the supplier firm may have special equipment and capability. Third, the outside supplier provides a check on the costs of the inhouse operation. United States companies rely more than others on independent suppliers. Only General Motors in the United States produces over half of its own parts.

Production of a new model also calls for elaborate tooling, and the larger the output the more highly specialized the tools. For example, it is expensive to install a stamping press exclusively to make a single body panel for a single year model, but if the model run reaches several hundred thousand, the cost is amply justified.

The assembly process itself has a quite uniform pattern throughout the world. As a rule there are two main lines, body and chassis. On the first, the body panels are welded together, doors and windows installed, and the body painted and trimmed (upholstery, interior hardware, wiring). On the second, the frame has springs, wheels, steering gear, and power train (engine, transmission, drive shaft, and differential) installed, plus brakes and exhaust system. The two lines merge at the point where the car is finished except for minor items and necessary testing and inspection. A variation on this process is "unitized" construction, whereby body and frame are assembled as a unit. In this system the undercarriage still goes down the chassis line for power train, front suspension, and rear axle, to be supported on pedestals until they are joined to the unitized body structure. Engines are constructed in the same way, as a rule, in separate assembly plants.

Automated assembly lines

Assembly lines have been elaborately refined by automatic control systems and transfer machines, which have replaced many manual operations where volume is high. Austin Motors in Britain seems to have pioneered with automatic transfer machines in 1950. The first large-scale automated installation in the United States was a Ford Motor Company engine plant that went into production in 1951. A universally employed form of automatic control has been the use of punch cards and computers to schedule assembly operations so that a variety of styles can be programmed along the same assembly line. Customers can be offered wide choices in body styles, wheel patterns, and colour combinations.

Sales and service organization. Mass production implies mass consumption, which in turn requires an elaborate distributive organization not only to sell the cars but to develop customer confidence that adequate service will be available. In the early days cars were sold direct from the factory or through independent dealers who might handle several different makes. Many of the bicycle manufacturers simply used their existing sales outlets when they added horseless carriages to their line. When sales in large quantities became the objective, however, more elaborate and better organized techniques of distribution became essential.

In the United States the restricted franchise dealership became the uniform and almost exclusive method of selling new cars. In this system the dealer may sell only the particular make of new car specified in his franchise, must accept a quota of cars specified by the manufacturer, and must pay cash on delivery. In return he receives some guarantee of sales territory and may be assisted in various ways by the manufacturer: financing or aid in advertising, for instance. Contracts also specify that the dealer must maintain service facilities on standards approved by the manufacturer.

Seemingly weighted in favour of the manufacturer, the system has been subjected to periodic dealer complaint, producing state legislation and a federal statute in 1956 to protect dealers from arbitrary action on the part of the manufacturers. Yet dealers have never been united in these attitudes, and no effective substitute for the restricted franchise has as yet been found. On the contrary, it is becoming the general practice in other parts of the world where large-scale markets for motor vehicles have been developing.

The United States has about 34,000 franchised passenger-car dealers, all of whom sell both new and used cars, and about 28,000 nonfranchised dealers engaged predominantly in buying and selling used cars. In addition, there are over 100,000 enterprises engaged in general and specialized repairs to motor vehicles and well over 200,000 gasoline stations, most of which do some repair and service work. These are all at least ancillary to the automotive industry.

The market in used cars is an important part of the distribution system for motor vehicles in all countries with a substantial motor vehicle industry because it affects the sale and styling of new cars. The institution of the annual model was adopted in the United States during the 1920s to promote new-car sales in the face of used-car competition. The new model must have enough changes in styling or engineering to persuade prospective buyers that it is indeed an improvement. At the same time it must not be so radically different from its predecessors as to give the buyer doubts about its resale potential.

Like all machinery, motor vehicles wear out. In the United States alone over 7,000,000 are scrapped each year. Some become scrap metal to feed steel furnaces; some go to wrecking yards where usable parts are salvaged. Throughout the world, however, the disposal of discarded motor vehicles has become a problem without a completely satisfactory solution. In too many areas landscapes are disfigured by abandoned wrecks or unsightly automobile graveyards.

International operations. The automotive industry is supranational in its organization and operation and is matched in this respect only by the petroleum industry. Reference has been made to the international expansion of the principal American automobile companies. General Motors owns Vauxhall in Britain, Opel in Germany, and Holden in Australia. Ford has been established for many years in Great Britain, Germany, France, Spain, Denmark, and Holland, and in the early 1960s acquired Willys of Brazil. Chrysler controls Rootes in Britain and Simca in France. The major European producers follow a comparable pattern with worldwide networks of subsidiaries and affiliates, although none operates on as large a scale as General Motors or Ford. Most of the principal motor-vehicle-producing countries export a much higher proportion of their total output than does the United States (Table 2), but this picture is misleading because it does not allow for the very substantial volume of motor vehicle production by American-owned companies in other countries.

Table 2: Motor Vehicle Exports as Percentage of Production
(1969)

	percentage
Canada	83.4
Sweden	60.8
West Germany	57.0
France	47.8
U.K.	46.3
Italy	39.5
Japan	10.4
U.S.	4.3
Australia	2.6

Source: *Automobile Facts and Figures, 1971.*

The most promising export markets for motor vehicles were, and fundamentally still are, developed countries with the purchasing power to create a demand for automobiles, essentially North America, Europe, Japan, such Commonwealth states as Australia and New Zealand, and South Africa. Since 1950 there has been a significant shift in market prospects in favour of the developing countries, in that vehicle registrations in those countries have grown almost twice as fast as in the highly developed nations (Table 3). There has been, in consequence,

International vehicle registrations

Table 3: Motor Vehicle Registrations in 20 Leading Countries
(1969)

	cars	commercial	total	population per vehicle
U.S.	86,861,334	18,235,269	105,096,603	1.9
Japan	6,933,732	8,192,934	15,126,666	6.8
West Germany	12,198,180	2,099,472	14,297,652	4.1
France	11,670,000	2,100,000	13,770,000	3.7
U.K.	11,504,300	1,901,040	13,405,340	4.0
Italy	9,028,400	834,174	9,862,574	5.4
Canada	6,433,283	1,461,571	7,894,854	2.7
U.S.S.R.	1,560,000	4,075,000	5,635,000	43.0
Australia	3,676,241	949,587	4,625,828	2.7
Brazil	2,041,000	1,161,000	3,202,000	28.0
Spain	1,998,500	688,500	2,687,000	12.0
The Netherlands	2,225,000	308,000	2,533,000	5.1
Sweden	2,193,634	156,181	2,349,815	3.4
Belgium	1,920,638	281,785	2,202,423	4.4
Argentina	1,390,000	760,500	2,150,500	11.0
South Africa	1,415,000	402,000	1,817,000	11.0
Mexico	1,112,116	524,618	1,636,734	30.0
Switzerland	1,283,670	105,649	1,389,319	4.5
Denmark	1,023,790	262,032	1,285,822	3.8
Austria	1,116,246	119,695	1,235,941	6.0
World total	181,211,319	50,817,677	232,028,996	15.0

Source: *Automobile Facts and Figures, 1971.*

Table 4: Automotive Materials Consumption as a Percentage of Total U.S. Consumption
(1969)

materials	percentage of output used by automotive industry
Steel	
Alloy steel (excluding stainless)	21.4
Stainless steel	13.6
Steel (including carbon, alloy, stainless)	
Hot-rolled bars	29.8
Cold-finished bars	14.1
Hot-rolled strip	36.3
Cold-rolled strip	20.3
Hot-rolled sheet	37.1
Cold-rolled sheet	44.3
Galvanized (sheets and strip)	16.0
Aluminum	9.9
Copper and copper alloys	8.3
Cotton	2.2
Gray and ductile iron	17.1
Lead	54.7
Malleable iron	41.7
Nickel	11.2
Rubber	
Natural	71.7
Reclaimed	59.4
Synthetic	63.9
Zinc	32.5

Source: Compiled by Automobile Manufacturers Association from various trade sources.

an intensification of both assembly and distribution in parts of the world not previously important in the automotive industry. In 1965 about 1,000,000 motor vehicles (out of a world total of 25,000,000) were produced outside the major manufacturing nations, about 80 percent in Australia, Spain, Argentina, Brazil, Belgium, and East Germany.

The great bulk of this production is assembly, done in plants affiliated with and usually operated by American, European, or Japanese automotive firms. Because of the small volume of output, the assembly operations are high cost. In a completely free market it would be much cheaper to ship cars from the home plants, but the market is not free. In order to develop their own industries, these nations by and large have tariff policies that make imported cars prohibitively expensive and, in addition, requirements that a substantial portion of the components used in local assembly plants be of domestic origin. A stipulated percentage of local ownership, public or private, is also a normal requirement. The rest of the financing and most of the initial managerial and technical skill come from the parent company. Argentina, Brazil, and Spain have been able to reach 90 to 100 percent of domestic content in car manufacture; India and Mexico between 60 and 80 percent. While these proportions may portend well for the future, they are an additional cost factor at present. An Indian car, for example, with 97 percent domestic content, costs twice as much to produce as the same vehicle in Europe; with 28 percent domestic content, the cost is one and one-half times what it is in Europe.

Economic and social impact. The automotive industry has become a vital element in the economy of the industrialized nations to the extent that motor vehicle production and sales are one of the major indices of the state of the economy in those countries. For the United Kingdom, Japan, France, Italy, Sweden, and West Germany, motor vehicle exports are essential to the maintenance of healthy international trade balances.

The effect of motor vehicle manufacturing on other industries is shown in part in Table 4. Almost one-fourth of United States steel production and well over three-fifths of its rubber output go to the automotive industry, which is also the largest single consumer of machine tools. More than that, the special requirements of automotive mass production have had a profound influence on the design and development of highly specialized machine tools and have stimulated technological advances in petroleum refining, steelmaking, plate glass, paints, and other industrial areas.

The indirect effects are also considerable, through the many businesses such as motor-freight operators and highway construction firms which are automotive related. Finally, truck transportation has grown steadily throughout the world.

Highway development. Before the advent of the motor vehicle, roads in most parts of the world were generally poor. The available methods of road transport were so costly and inefficient that, unless there were special considerations such as military movements, it was not worthwhile to maintain roads for other than local movements. The general use of automobiles created a strong demand for better highways. The first response was to provide for the improvement of existing road networks. Experience subsequently demonstrated that roads for automobile traffic needed to be differentiated functionally, depending on whether they were intended for through or local traffic. Main arteries are best designed as freeways (motorways, *autostrade*, *Autobahnen*); *i.e.*, divided highways with complete control of access and no intersections at grade (see ROADS AND HIGHWAYS).

Highways functionally differentiated

Social effects. A historian said that Henry Ford freed the common man from the limitations of his geography. The statement cogently summarizes the social transformations still proceeding throughout the world as a result of the motor vehicle. It has created mobility on a scale never known before, and the total effect on living habits and social customs is still incalculable. The automobile has radically changed urban life by accelerating the outward expansion of population into the suburbs. As with other automobile-related phenomena, the trend is most conspicuous in the United States but is rapidly appearing elsewhere. The decentralizing trend is accentuated by the fact that highway transportation encourages business and industry to move outward to sites where land is cheaper, access by car and truck is easier than in crowded central cities, and space is available for the one-story structures that permit optimum use of modern materials-handling techniques. Yet the impact on rural life has been, if anything, more pronounced than the effect on the city. In the days of horse-drawn transport the economical limit of wagon transportation was from 10 to 15 miles, so that any community or individual farm farther than 15 miles from a railroad or navigable waterway was effectively isolated from the mainstream of economic and social life.

In urban and rural areas alike the automobile is credited with spawning drastic changes in the sexual mores of the young, who have found in it a privacy not formerly attainable.

Motor vehicles and paved roads have narrowed much of the gap between rural and urban life. Not only can the farmer ship easily and economically by truck and drive to

town when convenient but institutions such as regional schools and hospitals are now accessible by bus and car.

It would be impossible to list all the specific effects of motor vehicle production, but two are especially illustrative. First, the marketing of automobiles has stimulated a great expansion in the use of credit. Installment buying existed before the automobile, but in a limited scope. The technique was introduced into the United States automobile industry in 1916 by manufacturers of medium-priced cars to help meet the competition of the low-priced Model T. It became universal practice in all countries in the purchase of motor vehicles and accustomed people to buying other durable consumer goods in the same way. Second, there has been a striking development of drive-in business, again first in the United States but spreading rapidly everywhere.

Recreational travel. One of the conspicuous impacts of the automobile has been to permit nearly everyone in the automotive countries to travel for recreation. The motor vehicle allows for such auxiliary devices as trailers (caravans in Europe), campers, and boat trailers, to broaden the scope of recreational opportunities.

Adverse effects. The mass use of motor vehicles was bound to have some unforeseen and undesirable consequences, of which three can be singled out: traffic congestion, air pollution, and highway accidents. The approach to each of these problems has illustrated a common propensity to blame the technology, rather than the way in which the technology has been used.

Congestion, pollution, and accidents

City streets were congested long before the automobile existed, but the problem has been compounded enormously by the masses of motor vehicles that attempt to enter or leave cities at peak traffic hours. The constantly growing number of automobiles throughout the world adds to the difficulty of finding remedies for congestion. The heart of the problem is that few city street systems have been designed for automobile traffic. Reliable estimates are that some two-thirds of the vehicles in central business districts are passing through and should have been routed on circumferential highways. Remedying this situation is difficult and expensive. It calls for modern highways to provide ready access both into and around downtown areas. Programs for this purpose encounter vigorous opposition, frequently justified, on the ground that building freeways in cities disrupts neighbourhoods and destroys scenic or historic values.

The widespread use of automobiles for work trips has also led in many cities to a decline in public-transit systems, and there has been much discussion of the need for encouraging the development and use of up-to-date mass-transit media. Given the trend toward dispersal of people and business in urban areas, it seems doubtful that mass transit will appreciably diminish motor vehicle traffic. In any event, for most cities bus systems can provide the needed capacity for public transportation and are the most economical way of doing so.

Atmospheric pollution also antedates the automobile, but the concentration of many thousands of motor vehicles in large cities has given the problem a new dimension. Automobile exhausts contribute from half the atmospheric pollutants in most cities to a high of about 80 percent in Los Angeles, where atmospheric conditions are peculiarly conducive to smog formation. Progress in combatting the problem has been slow because the development of suitable devices has taken time. Studies of alternative power plants are also being made. In the early 1970s only the United States had enacted legislation in this area, but with use of motor vehicles expanding more rapidly elsewhere, action against automobile-caused air pollution seems likely to spread.

The most distressing feature of large-scale use of motor vehicles has been the universally heavy toll of highway accidents. The United States, with the largest number of cars and the heaviest volume of highway travel, has the highest aggregate figures in this category: over 50,000 fatalities and a million disabling injuries annually since 1965. The economic and social costs are virtually beyond measure. Rising concern over this situation led to the passage in 1966 of the National Traffic and Motor Vehicle Safety Act, requiring the incorporation of safety features in cars, and the Highway Safety Act, for developing comprehensive programs of traffic safety. As automobile usage has increased elsewhere, other countries have also had to wrestle with this problem. Studies throughout the world make it clear that traffic safety has three major components—the vehicle, the driver, and the road—and that safety programs must give due weight to all three.

FUTURE TRENDS

Prediction about the future of the automotive industry is necessarily risky, but some trends seem reasonably certain unless drastic and currently unforeseeable change occurs. The production and use of motor vehicles will continue to increase, with the increase proportionately greater in those parts of the world that do not now have large, well-established automotive industries or widespread automobile use, especially the U.S.S.R., India, and Latin America.

Next, given existing public pressures, greater emphasis will be placed on incorporating safety and antipollution features into the design and construction of motor vehicles. Partly because of these pressures, but more because technology changes, the automobile of the future will eventually have a different power plant, although what it will be and when it will come cannot be accurately forecast. Experiments with steam have been much discussed but have not yielded economically practical results. Electric automobiles remain subject to the limitations of the battery. Fuel-cell power plants have acute technical problems to be overcome before they are ready for everyday use in passenger cars. Any or all of these may experience a technological breakthrough, but the most likely replacement for the internal combustion engine is some form of gas turbine.

The automotive industry itself will probably continue to be dominated by a few very large firms. Even where new production develops, current experience is that it occurs in some form of relationship with the existing automotive giants. This scale of operation is inherent in the technology of automotive production. There is likely to be increased diversification as a measure of protection against fluctuation in the market for motor vehicles and possibly against anti-automobile legislation, but this diversification will hardly affect the character of the basic operation.

BIBLIOGRAPHY

General histories: Informed general accounts of the American automotive industry include: R.M. CLEVELAND and S.T. WILLIAMSON, *The Road is Yours* (1951); M. DENISON, *The Power to Go* (1956); and J.B. RAE, *American Automobile Manufacturers: The First Forty Years* (1959), and *The American Automobile* (1965). Comparable accounts for Great Britain are H.G. CASTLE, *Britain's Motor Industry* (1950); and L.T.C. ROLT, *Horseless Carriage: The Motor Car in England* (1950).

Economic and Social Studies: Two comprehensive works on the impact of the automobile are C.D. BUCHANAN, *Mixed Blessing: The Motor in Britain* (1958); and J.B. RAE, *The Road and the Car in American Life* (1971). J.J. FLINK, *America Adopts the Automobile, 1895–1910* (1970), is a unique description of the acceptance of the motor vehicle.

Biographical works: A brief biography of Henry Ford is R. BURLINGAME, *Henry Ford: A Great Life in Brief* (1954). A.P. SLOAN, JR. has a good autobiography, *My Years with General Motors* (1963). The best biography of a leading British automotive figure is P.W.S. ANDREWS and E. BRUNNER, *The Life of Lord Nuffield* (1955). There is an excellent study of the German automotive pioneers in E. DIESEL, G. GOLDBECK, and F. SCHILDBERGER, *Vom Motor zum Auto* (1957; Eng. trans., *From Engines to Autos*, 1960).

Modern industry: An excellent analysis of the American automobile industry in the 1950s and early 1960s is C.E. EDWARDS, *The Dynamics of the United States Automobile Industry* (1965). A comparable study for Great Britain is G. MAXCY and A. SILBERTSON, *The Motor Industry* (1959). The most authoritative work on the expansion of the automotive industry to the developing countries is J. BARANSON, *Automotive Industries in Developing Countries* (1968).

(J.B.Ra.)

Autopsy

Autopsy is the dissection and examination of a dead body and its organs and structures to determine the cause of death, to observe the effects of disease, and to establish the sequences of changes and thus establish evolution and mechanisms of disease processes. The procedure is also variously called necropsy, postmortem, and postmortem examination.

History. Although the Greek physician Galen (AD *c.* 130–*c.* 200) accepted the ancient humoral theory of disease as a faulty mixture of the four body fluids or humours, he was the first to correlate the patient's symptoms (complaints) and his signs (what can be seen and felt) with what was found upon examining the "affected part of the deceased." This was the great leap forward that eventually led to the autopsy and broke an ancient barrier to progress in medicine. The early Egyptians did not study the dead human body for an explanation of disease and death, though some organs were removed for preservation. The first real dissections for the study of disease were carried out by the Alexandrian physicians Herophilus (flourished 300 BC) and Erasistratus (flourished 300 BC). The Greeks and the Indians cremated their dead without examination; the Romans, the Chinese, and Muslims all had taboos about opening the body. Human dissections were not permitted during the Middle Ages; only animals were examined. Medieval prohibition was based on the doctrine that "the Church abhors the shedding of blood." (Opposition to the autopsy still persists among some sects.)

The first dissections

Mondino dei Liucci, of Bologna, did the first public anatomical demonstrations in 1315. It was the rebirth of anatomy during the Renaissance, as exemplified by the work of Andreas Vesalius (*De Humani Corporis Fabrica,* 1543) that made it possible to distinguish the abnormal, as such (*e.g.*, an aneurysm), from normal anatomy. Leonardo da Vinci (1452–1519) dissected 30 corpses and noted "abnormal anatomy"; his records of these and many of his anatomical drawings have come to light only recently. Michelangelo (1475–1564), too, did a number of dissections. Earlier, in the 13th century, Frederick II ordered that the bodies of two executed criminals be delivered every two years to the medical schools, one of which was at Salerno, for an "Anatomica Publica," which every physician was obliged to attend.

In 1348 Pope Clement VI ordered the bodies of plague victims opened; his physician, Guy de Chauliac, described the changes. The first forensic or legal autopsy, wherein the death was investigated to determine presence of "fault," is said to have been one requested by a magistrate in Bologna in 1302; no guilt was found. Antonio Benivieni, a Florentine physician (*c.* 1440–1502), did 15 autopsies explicitly to determine the "cause of death" and significantly correlated some of his findings with prior symptoms in the deceased. Théophile Bonet of Geneva (1620–89) collated from the literature the observations made in 3,000 autopsies. Many specific clinical and pathologic entities were then defined by various observers, thus opening the door to modern practice.

The autopsy came of age with Giovanni Morgagni, the father of modern pathology, who in 1761 described what could be seen in the body with the naked eye. In his voluminous work *On the Seats and Causes of Diseases as Investigated by Anatomy*, he compared the symptoms and observations in some 700 patients with the anatomical findings upon examining their bodies. He proved the dictum of his teacher Antonio Valsalva (1666–1723) that an apoplectic hemorrhage in the brain caused a paralysis on the opposite side of the body. He described syphilitic aneurysms, cirrhosis, and acute yellow atrophy of the liver. In addition, he observed that in pneumonia the lung became solid and took on the substance and appearance of liver, introducing the pathologic term "hepatization." He noted the relationship of the occupation of stone-cutting to tuberculosis of the lung (consumption) and to caseous tuberculous lymph nodes (scrofula). Many other entities, such as brain abscess from a middle ear infection and calcifying hardening of the arteries of the heart, are found among his cases. Most important, in Morgagni's method, study of the patient replaced that of books and commentaries.

With Karl von Rokitansky of Vienna (1804–78), the gross (naked eye) autopsy reached its apogee. Rokitansky utilized the microscope very little and was limited by his own humoral theory. The French anatomist and physiologist Marie F.X. Bichat (1771–1802) stressed the role of the different generalized systems and tissues in the study of disease. It was the German pathologist Rudolf Virchow (1821–1902), however, who introduced the cellular doctrine—that changes in the cells are the basis of the understanding of disease—in pathology and in autopsy. In *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre* (*Cellular Pathology Based on Physiological and Pathological Histology*, 1858), he established that the cell is the unit of life, that all cells come from pre-existing cells, that disease is altered cell function or relations or both; proclaimed that the anatomical lesion at the cellular level is the disease; and destroyed utterly the humoral theory. He also warned against the dominance of pathologic anatomy—the study of the structure of diseased tissue—alone as such and stressed that the future of pathology would be physiologic pathology—study of the functioning of the organism in the investigation of disease.

The modern autopsy has been expanded to include the application of all knowledge and all of the instruments of the specialized modern basic sciences. The examination has been extended to structures too small to be seen except with the electron microscope, and to molecular biology to include all that can be seen as well as what still remains unseen.

The procedure. The autopsy procedure itself has changed very little during the 20th century. There has always been a gross examination of the exterior for any abnormality or trauma and a careful description of the interior of the body and its organs. For some time this has been followed by further studies, including microscopic examination, for which purpose portions of tissues and organs (blocks) are embedded in paraffin or other material and thin slivers (sections) are stained by one of the many techniques evolved over the years.

Gross examination

The main incisions in the body remain the same. For the torso, a Y-shaped incision is made. Each upper limb of the "Y" extends from either the armpit or the outer shoulder and is carried beneath the breast to the bottom of the sternum, or breastbone, in the midline. From this point of juncture at the bottom of the sternum the incision is continued down to the lower abdomen where both groins meet in the genital area. The incision extends through all the layers of the abdominal wall, and, with some undermining cuts, the flaps are turned back so that the abdominal cavity is laid open. The chest content is exposed by removal of the sternum, and the intestinal tract is usually removed by stripping the bowel from its attachment.

There are two different schools as to procedure beyond this point. In the Rokitansky method, each organ is removed separately for incision and study. In the en masse method the chest organs are all removed in a single group and all of the abdominal organs in another for examination. The great vessels to the neck, head, and arms are ligated—tied off—and the organs removed as a unit for dissection. The neck organs are explored *in situ* only or removed from below. Dissection then proceeds usually from the back, except where findings dictate a variation in the procedure. Usually groups of organs are removed together so that disturbances in their functional relationships may be determined. In examination of the head, an incision is made across the top or vertex of the scalp and the resulting flaps turned out front and back. The top of the skull is removed. After study of the brain in position, it is freed from its attachments and removed *in toto*. The spinal cord can be removed by sawing through the bodies of the vertebrae from in front or by making another incision from the back and removing the vertebral arches.

The dissector proceeds to examine the external and cut surface of each organ, its vascular structures, including arteries, lymphatics, fascial or fibrous tissue, and nerves. Specimens are taken for culture, chemical analysis, and other studies. Immediately upon completion of the procedure, all of the organs are returned to the body and all incisions carefully sewn. After the body's proper restoration, no unseemly evidence of the autopsy need remain.

After the gross examination of the body the findings are balanced one against another and a list of pathological findings is compiled; this list comprises the tentative or "provisional anatomical diagnoses." Such diagnoses are grouped and arranged in the order of importance and of sequence. On occasion a quick microscopic study is done to confirm a diagnosis so as to assure its proper listing. Finally the examiner lists as the cause of death the one lesion without which death would not have occurred. Though obviously all-important in forensic cases, this aspect of the autopsy analysis is also required in cases not required by law. After all studies—histological, chemical, toxicological, bacteriological, and viral—are completed, any errors of the provisional anatomical diagnoses are corrected and the final anatomical diagnoses and the final cause of death are listed. A statement of analysis of the autopsy that correlates the findings with the clinical picture, the "clinical pathological correlation," concludes the record of the autopsy.

Final diagnosis and final cause of death

The medicolegal autopsy. The forensic autopsy is an imperative function in modern society. The truth must prevail in certified detail if justice is to prevail. The forensic pathologist goes beyond the mere cause of death; he must establish all the facts, both lethal and nonlethal, with any potential bearing whatsoever on the criminal or civil litigation. The cause of death is not automatically revealed when the body is opened; it is not an isolated tangible and delimited entity; it is a *concept*—an opinion—as to mechanism or happening and as such is subject occasionally to differences in interpretation. The legal autopsy requires meticulous detailed descriptions, measurements, and documentation. Experience in the investigation of the scene of a death in medicolegal cases is important, for the evaluation of circumstances of death may be critical in establishing the mode of death—*e.g.*, suicide. The autopsy may not be able, of itself, to determine intent, whereas the scene and the circumstances, such as an empty pill bottle or a note beside the body, may provide unmistakable evidence. Photographic documentation is important in the medicolegal autopsy. The medicolegal postmortem examination must always be complete to rule out any other potential contributory cause of death and therefore must never be limited to a partial study. The identification of the deceased and of all specimens taken from the body is critical; the time of death and the blood grouping must, if possible, be established. The medicolegal autopsy substitutes facts for speculation in the courtroom; it helps convict the guilty and free the innocent. The external examination is important and must include every minute detail of trauma to help determine its nature. The small chemical burn on the lip may indicate poisoning. In the case of wounds, the description must note the depth, direction and extent, bleeding, discoloration, height of the wound above heel, and structures penetrated. All important is the presence or absence of powder burns and their size and configuration as well as whether the wound edges are turned out or in. In all autopsies, but especially in forensic cases, findings must be dictated to a stenographer or recording instrument during the actual performance of the procedure. The record often becomes legal evidence and therefore must be complete and accurate.

External examination

In some jurisdictions administrative codes are specific and provide that the death of any person from criminal violence, or by casualty, or by suicide, or suddenly, while in apparent health, or when unattended by a physician, or in any suspicious or unusual manner, shall be reported forthwith to the office of the chief medical examiner. As many as one-third of the deaths in a community may be reported for investigation to the medicolegal authorities.

Reasons for autopsy. Death represents the final failure of medical care, and so the autopsy is the final evaluation of the care given. The autopsy deals with the particular illness as evidenced in one individual and is more than simply a statistical average. Every autopsy is important to expose mistakes, to delimit new diseases and new patterns of disease, and to guide future studies. With the autopsy one can evaluate the numerous new drugs and modes of therapy; many are toxic and dangerous and can do more harm than good. Bizarre ideas about the nature of human disease are still extant; the autopsy continues to expose such erroneous dogma. Morbidity and mortality statistics acquire accuracy and significance when based on careful autopsies; they also often give the first indication of contagion and epidemics. The autopsy promotes a proper balance between all elements of the health complex—physician, hospital, research and education, and the integration of the multifarious specialties—so that excessive emphasis on any one at the expense of others can be corrected.

The autopsy also provides opportunities for the salvage of human organs for transplantation and for hormone extraction. Growth hormone from cadavers can make a dwarfed child normal; a transplant of a clear cornea can restore sight; a kidney can save a life. Genetic, hereditary, and degenerative diseases are becoming more frequent than those due to infection; the autopsy can point the way to counselling and preventive measures for the family. There is also psychological benefit for the family in the acceptance of the demonstrable disease, the reality of death, and the ultimate decomposition of the body.

Aid to medical education

Probably the most important role of the autopsy today is its contribution to medical education. The autopsy is invaluable in undergraduate, graduate, and continuing education of physicians. It is the focal point at which the profession learns to assess and to apply medical knowledge.

There was a time when virtually all progress in medicine was made at the autopsy table. In the 20th century many refinements in basic science and technology, which are applicable to the living patient, are important for the understanding and treatment of disease. Molecular concepts, however, are no substitute for what can be seen by the naked eye or for the older staples and simple truths of the autopsy. Disease cannot be known only from a battery of reports or from waves on a polygraph or oscilloscope but rather by seeing and understanding its extent and effect as it occurs in the human body.

The autopsy does more than merely determine the cause of death. While the medicolegal autopsy in particular has this important primary objective, most autopsies have a larger purpose: to reveal the sequential changes of disease from start to finish, as altered by intrinsic and extrinsic factors. By piecing together in proper order the static pictures found at each of many different stages of disease in a series of autopsies, an appropriate sequential analysis may be obtained of how the disease proceeds.

Future of the autopsy. Impressed by outstanding advances in biochemistry and molecular biology, some doctors consider the autopsy no longer important. The autopsy must continue to carry out the many routine functions noted above. Like all research, the autopsy is a deficit expenditure for hospitals and will be more costly as it becomes more thorough. The support provided by grants for experimental scientific research in other fields is also needed for the autopsy so that the autopsy can become more productive in research and publication as it becomes more precise and probing. In the future more biochemical, toxicological, and bacteriological studies should be done; viral studies are now more frequently indicated. The use of radioactive isotopes as tracers for test-tube study of the metabolism of tissues still viable after death should be explored. The special autopsy should include withdrawal—at the moment of death to forestall postmortem changes—of tissue and fluids from areas suspected of being foci of disease for study by the electron microscope. Appropriate animal experimentation should follow as indicated.

Only thus can the autopsy match other techniques in

the achievements of modern medicine. There has been a shift of emphasis in all pathology from a concentration on morphology alone to include study of function and molecular biology with sophisticated instrumentation. In the hospital the autopsy will continue to be important for the evaluation of clinical diagnoses and of medical and surgical treatment as well as for continuing education and clinical research. Some researchers believe that it is not unreasonable to look forward to the day when data on every illness, injury, drug, surgical procedure, and other event of significance in disease, including autopsy diagnoses, will be entered, with some form of identification for each individual, in a computer. Then, instead of being simply cancelled out at death the identifying symbol—such as a social security number—could be kept with all the entries; and with the addition of the coded autopsy data, there would be available a permanent storehouse of knowledge for future study of human disease. The life of each person would thus become an invaluable contribution to medical knowledge. In this and many other ways the autopsy can continue and enhance its vital role in medicine.

BIBLIOGRAPHY. Additional information may be found in the following articles: J.B. HAZARD *et al.*, "Symposium on the Autopsy," *JAMA*, 193:805–814 (1965); J.M. PRUTTING *et al.*, "Symposium on Medical Progress and the Postmortem," *Bull. NY Acad. Med.*, 44:793–861 (1968); A. ANGRIST, "Fitting the Old-Fashioned Autopsy into the Modern Medical Scene," *Amer. J. Clin. Path.*, 45:202–207 (1966); "Progress and Paradox in Pathology and Medicine," *Pharos*, 32:48–53 (1969); "Experimental Research and the Autopsy," *Bull. NY Acad. Med.*, 45:3–9 (1969); and "A Plea for Support of Training in Pathologic Anatomy in Our Medical Schools," *Arch. Path.*, 73:1–5 (1962).

(A.A.A.)

Averroës

Abū al-Walīd Muḥammad ibn Aḥmad ibn Muḥammad ibn Rushd—called Averroës by the medieval Latin scholars—was one of the most important Islāmic thinkers; he integrated Islāmic traditions and Greek philosophy, especially that of Aristotle, into a system of thought of his own. Averroës' commentaries greatly contributed to the rise of Scholastic philosophy as every page of the Schoolmen shows (*e.g.*, Albertus Magnus and Thomas Aquinas).

Life and works. Averroës was born into a distinguished family of jurists at Córdoba in 1126 and died at Marrakesh, the North African capital of the Almohad (al-Muwaḥḥidūn) dynasty, in 1198. Thoroughly versed in the traditional Muslim sciences (especially exegesis of the Qur'ān—Islāmic scripture—and Ḥadīth, or Traditions, and *fiqh*, or Law), trained in medicine, and accomplished in philosophy, Averroës rose to be chief *qāḍī* (judge) of Córdoba (Qurṭubah), an office also held by his grandfather (of the same name) under the Almoravids (al-Murābiṭūn). After the death of the philosopher Ibn Ṭufayl, Averroës succeeded him as personal physician to the caliphs Abū Ya'qūb Yūsuf in 1182 and his son Abū Yūsuf Ya'qūb in 1184. In 1169 Ibn Ṭufayl introduced Averroës to Abū Ya'qūb, who, himself a keen student of philosophy, frightened Averroës with a question concerning whether the heavens were created or not. The caliph answered the question himself, put Averroës at ease, and sent him away with precious gifts after a long conversation that proved decisive for Averroës' career. Soon afterward Averroës received the ruler's request to provide a badly needed correct interpretation of the Greek philosopher Aristotle's philosophy, a task to which he devoted many years of his busy life as judge, beginning at Seville and continuing at Córdoba. The exact year of his appointment as chief *qāḍī* of Córdoba, one of the key posts in the government (and not confined to the administration of justice), is not known.

Commentaries on Aristotle's works

Between 1169 and 1195 Averroës wrote a series of commentaries on most of Aristotle's works (*e.g.*, the *Organon*, *De anima*, *Physica*, *Metaphysica*, *De partibus animalium*, *Parva naturalia*, *Meteorologica*, *Rhetorica*, *Poetica*, and the *Nicomachean Ethics*). He wrote summaries, and middle and long commentaries—often two or all three kinds on the same work. Aristotle's *Politica* was inaccessible to Averroës; therefore he wrote a commentary on Plato's *Republic* (which is both a paraphrase and a middle commentary in form). All of Averroës' commentaries are incorporated in the Latin version of Aristotle's complete works. They are extant in the Arabic original or Hebrew translations or both, and some of these translations serve in place of the presumably lost Arabic originals; *e.g.*, the important commentaries on Aristotle's *Nicomachean Ethics* and on Plato's *Republic*. Averroës' commentaries exerted considerable influence on Jews and Christians in the following centuries. His clear, penetrating mind enabled him to present competently Aristotle's thought and to add considerably to its understanding. He ably and critically used the classical commentators Themistius and Alexander of Aphrodisias and the *falāsifah* (Muslim philosophers) al-Fārābī, Avicenna (Ibn Sīnā), and his own countryman Avempace (Ibn Bājjah). In commenting on Aristotle's treatises on the natural sciences, Averroës showed considerable power of observation.

His own first work is on *General Medicine* (*Kulliyāt*, Latin *Colliget*), written between 1162 and 1169. Only a few of his legal writings and none of his theological writings are preserved. Undoubtedly his most important writings are three closely connected religious–philosophical polemical treatises, composed in the years 1179 and 1180: the *Decisive Treatise on the Agreement between Religious Law and Philosophy* (*Faṣl*) with its *Appendix: Examination of the Methods of Proof Concerning the Doctrines of Religion* (*Manāhij*); and *The Incoherence of the Incoherence* (*Tahāfut at-tahāfut*) in defense of philosophy. In the two first named Averroës stakes a bold claim: only the metaphysician employing certain proof (syllogism) is capable and competent (as well as obliged) to interpret the doctrines contained in the prophetically revealed law (Shar' or Sharī'ah), and not the Muslim *mutakallimūn* (dialectic theologians), who rely on dialectical arguments. To establish the true, inner meaning of religious beliefs and convictions is the aim of philosophy in its quest for truth. This inner meaning must not be divulged to the masses, who must accept the plain, external meaning of Scripture contained in stories, similes, and metaphors. Averroës applied Aristotle's three arguments (demonstrative, dialectical, and persuasive—*i.e.*, rhetorical and poetical) to the philosophers, the theologians, and the masses. The third work is devoted to a defense of philosophy against his predecessor al-Ghazālī's telling attack directed against Avicenna and al-Fārābī in particular.

Defense of philosophy against theologians

Spirited and successful as Averroës' defense was, it could not restore philosophy to its former position, quite apart from the fact that the atmosphere in Muslim Spain and North Africa was most unfavourable to the unhindered pursuit of speculation. As a result of the reforming activity of Ibn Tūmart (*c.* 1078–1130), aimed at restoring pure monotheism, power was wrested from the ruling Almoravids, and the new Berber dynasty of the Almohads was founded, under whom Averroës served. In jurisprudence the emphasis then shifted from the practical application of Muslim law by appeal to previous authority to an equal stress on the study of its principles and the revival of independent legal decision on the basis of Ibn Tūmart's teaching. Of perhaps even more far-reaching significance was Ibn Tūmart's idea of instructing the heretofore ignorant masses in the plain meaning of the Sharī'ah so that practice would be informed with knowledge. These developments were accompanied by the encouragement of the *falāsifah*—"those who," according to Averroës' *Faṣl*, "follow the way of speculation and are eager for a knowledge of the truth"—to apply demonstrative arguments to the interpretation of the theoretical teaching of the Sharī'ah. But with the hands of both jurists and theologians thus strengthened, Averroës' defense of philosophy continued to be conducted within an unfavourable atmosphere.

Averroës himself acknowledged the support of Abū Ya'qūb, to whom he dedicated his *Commentary on Plato's Republic*. Yet Averroës pursued his philosophical

quest in the face of strong opposition from the *mutakallimūn*, who, together with the jurists, occupied a position of eminence and of great influence over the fanatical masses. This may explain why he suddenly fell from grace when Abū Yūsuf—on the occasion of a "holy war" (*jihād*) against Christian Spain—dismissed him from high office and banished him to Lucena in 1195. To appease the theologians in this way at a time when the caliph needed the undivided loyalty and support of the people seems a more convincing reason than what the Arabic sources tell us (attacks on Averroës by the mob, probably at the instigation of jurists and theologians). But Averroës' disgrace was only short-lived—though long enough to cause him acute suffering—since the caliph recalled Averroës to his presence after his return to Marrakesh. After his death, Averroës was first buried at Marrakesh, and later his body was transferred to the family tomb at Córdoba. It is not rare in the history of Islām that the rulers' private attachment to philosophy and their friendship with philosophers goes hand in hand with official disapproval of philosophy and persecution of its adherents, accompanied by the burning of their philosophical writings and the prohibition of the study of secular sciences other than those required for the observance of the religious law. Without caliphal encouragement Averroës could hardly have persisted all his life in his fight for philosophy against the theologians, as reflected in his *Commentary on Plato's Republic*, in such works as the *Faṣl* and *Tahāfut at-tahāfut*, and in original philosophical treatises (*e.g.*, about the union of the active intellect with the human intellect). It is likely that the gradual estrangement of his two masters and patrons from Ibn Tūmart's theology and their preoccupation with Islāmic law also helped him. That Averroës found it difficult to pursue his philosophical studies alongside the conscientious performance of his official duties he himself reveals in a few remarks scattered over his commentaries; *e.g.*, in that on Aristotle's *De partibus animalium*.

Contents and significance of works. To arrive at a balanced appraisal of Averroës' thought it is essential to view his literary work as a whole. In particular, a comparison of his religious–philosophical treatises with his *Commentary on Plato's Republic* shows the basic unity of his attitude to the Sharīʿah dictated by Islām and therefore determining his attitude to philosophy, more precisely to the *nomos*, the law of Plato's philosopher-king. It will then become apparent that there is only one truth for Averroës, that of the religious law, which is the same truth that the metaphysician is seeking. The theory of the double truth was definitely not formulated by Averroës, but rather by the Latin Averroists. Nor is it justifiable to say that philosophy is for the metaphysician what religion is for the masses. Averroës stated explicitly and unequivocally that religion is for all three classes; that the contents of the Sharīʿah are the whole and only truth for all believers; and that religion's teachings about reward and punishment and the hereafter must be accepted in their plain meaning by the elite no less than by the masses. The philosopher must choose the best religion, which, for a Muslim, is Islām as preached by Muḥammad, the last of the prophets, just as Christianity was the best religion at the time of Jesus, and Judaism at the time of Moses.

Unity of religious and philosophical truth

It is significant that Averroës could say in his *Commentary on Plato's Republic* that religious law and philosophy have the same aim and in the *Faṣl* that "philosophy is the companion and foster-sister of the Sharīʿah." Accepting Aristotle's division of philosophy into theoretical (physics and metaphysics) and practical (ethics and politics), he finds that the Sharīʿah teaches both to perfection: abstract knowledge commanded as the perception of God, and practice—the ethical virtues the law enjoins (*Commentary on Plato's Republic*). In the *Tahāfut* he maintains that "the religious laws conform to the truth and impart a knowledge of those actions by which the happiness of the whole creation is guaranteed." There is no reason to question the sincerity of Averroës. These statements reflect the same attitude to law and the same emphasis on happiness. Happiness as the highest good is the aim of political science. As a Muslim, Averroës insists on the attainment of happiness in this and the next life by all believers. This is, however, qualified by Averroës as the disciple of Plato: the highest intellectual perfection is reserved for the metaphysician, as in Plato's ideal state. But the Muslim's ideal state provides for the happiness of the masses as well because of its prophetically revealed law, which is superior to the Greek *nomos* (law) for this reason. The philosopher Averroës distinguishes between degrees of happiness and assigns every believer the happiness that corresponds to his intellectual capacity. He takes Plato to task for his neglect of the third estate because Averroës believes that everyone is entitled to his share of happiness. Only the Sharīʿah of Islām cares for all believers. It legitimates speculation because it demands that the believer should know God. This knowledge is accessible to the naïve believer in metaphors, the inner meaning of which is intelligible only to the metaphysician with the help of demonstration. On this point all *falāsifah* are agreed, and all recognize the excellence of the Sharīʿah stemming from its divinely revealed character. But only Averroës insists on its superiority over the *nomos*.

The Sharīʿah's concern for the masses

Insisting on the prerogative of the metaphysician—understood as a duty laid upon him by God—to interpret the doctrines of religion in the form of right beliefs and convictions (like Plato's philosopher-king), he admits that the Sharīʿah contains teachings that surpass human understanding but that must be accepted by all believers because they contain divinely revealed truths. The philosopher is definitely bound by the religious law just as much as the masses and the theologians, who occupy a position somewhere in between. In his search for truth the metaphysician is bound by Arabic usage, as is the jurist in his legal interpretations, though the jurist uses subjective reasoning only, in contrast to the metaphysician's certain proof. This means that the philosopher is not bound to accept what is contradicted by demonstration. He can, thus, abandon belief in the creation out of nothing since Aristotle demonstrated the eternity of matter. Hence creation is a continuing process. Averroës sought justification for such an attitude in the fact that a Muslim is bound only by consensus (*ijmāʿ*) of the learned in a strictly legal context where actual laws and regulations are concerned. Yet, since there is no consensus on certain theoretical statements, such as creation, he is not bound to conform. Similarly, anthropomorphism is unacceptable, and metaphorical interpretation of those passages in Scripture that describe God in bodily terms is necessary. And the question whether God knows only the universals, but not the particulars, is neatly parried by Averroës in his statement that God has knowledge of particulars but that his knowledge is different from human knowledge. These few examples suffice to indicate that ambiguities and inconsistencies are not absent in Averroës' statements.

The *Commentary on Plato's Republic* reveals a side of Averroës that is not to be found in his other commentaries. While he carried on a long tradition of attempted synthesis between religious law and Greek philosophy, he went beyond his predecessors in spite of large-scale dependence upon them. He made Plato's political philosophy, modified by Aristotle, his own and considered it valid for the Islāmic state as well. Consequently, he applied Platonic ideas to the contemporary Almoravid and Almohad states in a sustained critique in Platonic terms, convinced that if the philosopher cannot rule, he must try to influence policy in the direction of the ideal state. For Plato's ideal state is the best after the ideal state of Islām based on and centred in the Sharīʿah as the ideal constitution. Thus, he regrets the position of women in Islām compared with their civic equality in Plato's *Republic*. That women are used only for childbearing and the rearing of offspring is detrimental to the economy and responsible for the poverty of the state. This is most unorthodox.

Platonic political philosophy

Of greater importance is his acceptance of Plato's idea of the transformation and deterioration of the ideal, perfect state into the four imperfect states. Muʿāwiyah I, who in Muslim tradition perverted the ideal state of the

first four caliphs into a dynastic power state, is viewed by Averroës in the Platonic sense as having turned the ideal state into a timocracy—a government based on love of honour. Similarly, the Almoravid and Almohad states are shown to have deteriorated from a state that resembled the original perfect Sharīʿah state into timocracy, oligarchy, democracy, and tyranny. Averroës here combines Islāmic notions with Platonic concepts. In the same vein he likens the false philosophers of his time, and especially the *mutakallimūn*, to Plato's sophists. In declaring them a real danger to the purity of Islām and to the security of the state, he appeals to the ruling power to forbid dialectic theologians to explain their beliefs and convictions to the masses, thus confusing them and causing heresy, schism, and unbelief. The study of *The Republic* and the *Nicomachean Ethics* enabled the *falāsifah* to see more clearly the political character and content of the Sharīʿah in the context of the classical Muslim theory of the religious and political unity of Islām.

Use of Plato's works

Leaning heavily on the treatment of Plato's political philosophy by al-Fārābī, a 10th-century philosopher, Averroës looks at *The Republic* with the eyes of Aristotle, whose *Nicomachean Ethics* constitutes for Averroës the first, theoretical part of political science. He is, therefore, only interested in Plato's theoretical statements. Thus he concentrates on a detailed commentary on Books II–IX of *The Republic* and ignores Plato's dialectical statements and especially his tales and myths, principally the myth of Er. He explains Plato, whose *Laws* and *Politikos* he also knows and uses, with the help, and in the light, of Aristotle's *Analytica posteriora*, *De anima*, *Physica*, and *Nicomachean Ethics*. Naturally, Greek pagan ideas and institutions are replaced by Islāmic ones. Thus Plato's criticism of poetry (Homer) is applied to Arab pre-Islāmic poetry, which he condemns. Averroës sees much common ground between the Sharīʿah and Plato's general laws (interpreted with the help of Aristotle), notwithstanding his conviction that the Sharīʿah is superior to the *nomos*. He accepts al-Fārābī's equation of Plato's philosopher-king with the Islāmic *imām*, or leader and lawgiver, but leaves it open whether the ideal ruler must also be a prophet. The reason for this may well be that, as a sincere Muslim, Averroës holds that Muḥammad was "the seal of the prophets" who promulgated the divinely revealed Sharīʿah once and for all. Moreover, Averroës exempts Muḥammad from the general run of prophets, thus clearly rejecting the psychological explanation of prophecy through the theory of emanation adopted by the other *falāsifah*. No trace of this theory can be discovered in Averroës' writings, just as his theory of the intellect is strictly and purely Aristotelian and free from the theory of emanation. In conclusion, it may be reiterated that the unity of outlook in Averroës' religious–philosophical writings and his commentary on *The Republic* gives his political philosophy a distinctly Islāmic character and tone, thereby adding to his significance as a religious philosopher.

BIBLIOGRAPHY. LEON GAUTHIER, *Ibn Rochd (Averroès)* (1948), a balanced overall picture of his life and works, and *La Théorie d'Ibn Rochd (Averroès) sur les rapports de la religion et de la philosophie* (1909), an indispensable work that was the basis for the author's later study cited above; G.F. HOURANI, *(Averroës) On the Harmony of Religion and Philosophy* (Eng. trans., 1961); E.I.J. ROSENTHAL, "The Place of Politics in the Philosophy of Ibn Rushd," *Bulletin of the School of Oriental and African Studies*, vol. 15, no. 2 (1953), a discussion of Averroës' political philosophy within the context of his life and writings, *Averroës' Commentary on Plato's Republic*, 3rd ed. (1969), Hebrew text, Eng. trans. and notes, giving Averroës' Greek and Arabic sources, and "Ibn Rushd: The Consummation," in *Political Thought in Medieval Islam*, 3rd ed. ch. 9 (1968); S. VAN DEN BERGH, *Averroës' Tahāfut al-tahāfut: The Incoherence of the Incoherence*, 2 vol. (1954), Eng. trans. with important notes tracing Averroës' sources, especially Aristotle.

(E.I.J.R.)

Avicenna

Abū ʿAlī al-Ḥusayn ibn ʿAbd Allāh ibn Sīnā, known in the West as Avicenna, the most influential of all Muslim philosopher-scientists, was especially famous for his contributions in the fields of Aristotelian philosophy and medicine. Because of his dominating influence and authority in these fields, he has been given the honorific titles of ash-Shaykh ar-Raʾīs ("the Leading Wise Man") in the East and Prince of the Physicians in the West.

Early years

Avicenna, a Persian who spent his whole life in the eastern and central regions of Persia, was born in Bukhara (now in Uzbek S.S.R.) in 980 and received his earliest education in that city under the direction of his father, who was an Ismāʿīlī (a member of an Islāmic religious and political movement, the theology of which drew on a popularized form of Neoplatonism). Avicenna himself, however, was never attracted to the Ismāʿīlīyah. Since the house of his father was a meeting place for learned men, from his earliest childhood Avicenna was able to profit from the company of the outstanding masters of his day. A precocious child with an exceptional memory that he retained throughout his life, he had memorized the Qurʾān and much Arabic poetry by the age of ten. Thereafter, he studied logic and metaphysics under teachers whom he soon outgrew, and then spent the few years until he reached the age of 18 in his own self-education. He read avidly and mastered Islāmic law, then medicine, and finally metaphysics. Particularly helpful in his intellectual development was his gaining access to the rich royal library of the Sāmānids—the first great native dynasty that arose in Persia after the Arab conquest—as the result of his successful cure of the Sāmānid prince, Nūḥ ibn Manṣūr. By the time he was 21 he was accomplished in all branches of formal learning and had already gained a wide reputation as an outstanding physician. His services were also sought as an administrator, and for a while he even entered government service as a clerk.

But suddenly the whole pattern of his life changed. His father died; the Sāmānid house was defeated by Maḥmūd of Ghazna, the Turkish leader and legendary hero who established Ghaznavid rule in Khorāsān (northeastern Iran and modern western Afghanistan); and Avicenna began a period of wandering and turmoil, which was to last to the end of his life with the exception of a few unusual intervals of tranquillity. Destiny had plunged Avicenna into one of the tumultuous periods of Persian history, when new Turkish elements were replacing Persian domination in Central Asia and local Persian dynasties were trying to gain political independence from the ʿAbbāsid caliphate in Baghdad (in modern Iraq). But the power of concentration and the intellectual prowess of Avicenna was such that he was able to continue his intellectual work with remarkable consistency and continuity and was not at all influenced by the outward disturbances.

Avicenna wandered for a while in different cities of Khorāsān and then left for the court of the Būyid princes, who were ruling over central Persia, first going to Rayy (near modern Tehrān) and then to Qazvīn, where as usual he made his livelihood as a physician. But in these cities also he found neither sufficient social and economic support nor the necessary peace and calm to continue his work. He went, therefore, to Hamadan in west central Persia, where Shams ad-Dawlah, another Būyid prince, was ruling. This journey marked a new phase in Avicenna's life. He became court physician and enjoyed the favour of the ruler to the extent that twice he was appointed vizier. As was the order of the day, he also suffered political reactions and intrigues against him and was forced into hiding for some time; he was even imprisoned.

Life in the court at Hamadan

This was the period when he began his two most famous works. *Kitāb ash-shifāʾ* ("The Book of Healing") is a vast philosophical and scientific encyclopaedia, probably the largest work of its kind ever written by one man. It treats of logic, the natural sciences, including psychology, the *quadrivium* (geometry, astronomy, arithmetic, and music), and metaphysics, but there is no real exposition of ethics or of politics. His thought in this work owes a great deal to Aristotle but also to other Greek influences and to Neoplatonism. His system rests on the conception of God as the necessary existent: in God alone essence,

what he is, and existence, that he is, coincide. There is a gradual multiplication of beings through a timeless emanation from God as a result of his self-knowledge. The *Canon of Medicine* (*Qānūn fī aṭ-ṭibb*) is the most famous single book in the history of medicine in both East and West. It is a systematic encyclopaedia based for the most part on the achievements of Greek physicians of the Roman imperial age and on other Arabic works and, to a lesser extent, on his own experience (his own clinical notes were lost during his journeys). Occupied during the day with his duties at court as both physician and administrator, Avicenna spent almost every night with his students composing these and other works and carrying out general philosophical and scientific discussions related to them. These sessions were often combined with musical performances and gaiety and lasted until late hours of the night. Even in hiding and in prison he continued to write. The great physical strength of Avicenna enabled him to carry out a program that would have been unimaginable for a person of a feebler constitution.

Life in the court at Isfahan

The last phase of Avicenna's life began with his move to Isfahan (about 250 miles south of Tehrān). In 1022 Shams ad-Dawlah died, and Avicenna, after a period of difficulty that included imprisonment, fled to Isfahan with a small entourage. In Isfahan, Avicenna was to spend the last 14 years of his life in relative peace. He was esteemed highly by 'Alā' ad-Dawlah, the ruler, and his court. Here he finished the two major works he began in Hamadan and wrote most of his nearly 200 treatises; he also composed the first work on Aristotelian philosophy in the Persian language and the masterly summary of his "Book of Healing" called *Kitāb an-najāt* ("The Book of Salvation"), written partly during the military campaigns in which he had to accompany 'Alā' ad-Dawlah to the field of battle. During this time he composed his last major philosophical opus and the most "personal" testament of his thought, *Kitāb al-ishārāt wa at-tanbīhāt* ("The Book of Directives and Remarks"). In this work he described the mystic's spiritual journey from the beginnings of faith to the final stage of direct and uninterrupted vision of God. Also in Isfahan, when an authority on Arabic philology criticized him for his lack of mastery in the subject, he spent three years studying it and composed a vast work called *Lisān al-'arab* ("The Arabic Language"), which remained in rough draft until his death. Accompanying 'Alā' ad-Dawlah on a campaign, Avicenna fell ill and, despite his attempts to treat himself, died in Hamadan in 1037 from colic and from exhaustion.

Besides fulfilling the role of the master of the Muslim Aristotelians, Avicenna also sought in later life to found an "oriental philosophy" (*al-ḥikmat al-mashriqīyah*). Most of his works directly concerning this have been lost, but enough remains in some of his other works to give an indication of the direction he was following. He took the first steps upon a path toward mystical theosophy that marked the direction that Islāmic philosophy was to follow in the future, especially in Persia and the other eastern lands of Islām.

Avicenna's influence

In the Western world, Avicenna's influence was felt, though no distinct school of "Latin Avicennism" can be discerned as can with Averroës, the great Spanish-Arabic philosopher. Avicenna's "Book of Healing" was translated partially into Latin in the 12th century, and the complete *Canon* appeared in the same century. These translations and others spread the thought of Avicenna far and wide in the West. His thought, blended with that of St. Augustine, the Christian philosopher and theologian, was a basic ingredient in the thought of many of the medieval Scholastics, especially in the Franciscan schools. In medicine the *Canon* became *the* medical authority for several centuries, and Avicenna enjoyed an undisputed place of honour equalled only by the early Greek physicians Hippocrates and Galen. In the East his dominating influence in medicine, philosophy, and theology has lasted over the ages and is still alive within the circles of Islāmic thought.

BIBLIOGRAPHY. Translations and commentaries on Avicenna's works include: M. ACHENA and H. MASSE, *Le Livre de science*, 2 vol. (1955–58); A.M. GOICHON, *Livre des directives et remarques* (1951); O.C. GRUNER, *A Treatise on the Canon of Medicine of Avicenna* (1930); M. HORTEN (ed.), *Das Buch der Genesung der Seele: Eine philosophische Enzyklopädie Avicennas*, vol. 4, *Die Metaphysik, Theologie, Kosmologie und Ethik* (1908); H. JAHIER and A. NOUREDDINE, *Poème de la médicine* (1956); A.F. MEHREN, *Traités mystiques . . . d'Avicenne*, 3 vol. (1889–91); F. RAHMAN, *Avicenna's Psychology* (1952).

General studies include: S.M. AFNAN, *Avicenna: His Life and Works* (1958); H. CORBIN, *Avicenne et le récit visionnaire*, 2nd ed., 2 vol. (1954; Eng. trans., *Avicenna and the Visionary Recital*, 1960); M. CRUZ HERNANDEZ, *La metafísica de Avicenna* (1949); L. GARDET, *La Pensée religieuse d'Avicenne* (1951); S.H. NASR, *An Introduction to Islamic Cosmological Doctrines* (1964), and *Three Muslim Sages* (1964), locating Avicenna within the context of Islāmic intellectual tradition; M.H. SHAH, *The General Principles of Avicenna's Canon of Medicine* (1966), an analysis from the point of view of modern medical theory and practice; G.W. WICKENS (ed.), *Avicenna: Scientist and Philosopher* (1952), a collection of essays; G.C. ANAWATI, *Essai de bibliographie avicennienne* (1950); Y. MAHDAVI, *Bibliographie d'Ibn Sina* (1954).

(S.H.N.)

Avoidance Behaviour

In one of its major meanings "avoidance" is any behaviour induced by adverse stimuli. The underlying implication that a single neural mechanism is involved (such as a specific part of the brain, which, under electrical stimulation, seems to inflict punishment) remains only a hypothesis. Clearly, the same kinds of avoidance behaviour might result from different underlying physiological mechanisms. Thus, although the various dichotomies, or polarities, of behaviour such as positive and negative, psychoanalytic life and death instincts, and approach and withdrawal concepts may be logical or philosophical conveniences, they seem, nevertheless, to lack clear meaning physiologically.

Alternative usage defines avoidance behaviour by describing a number of patterns: active avoidance (fleeing), passive avoidance (freezing stock-still or hiding), and a pattern of protective reflexes, as seen in the startle response. There is good reason to suppose that, in cats, for example, each of these patterns is coordinated separately by the brain. One kind of fleeing, in which the cat moves continuously and shows much upward climbing, is produced by electrical stimulation of specific parts of the brain (hypothalamic sites). Stimulation of other sites (in the thalamus) generates other types of fleeing movements, causing the animal to crouch, look around, move, slink close to the floor, and hide, if possible. In general, among birds and mammals, brain sites for fleeing of the first type occur in hypothalamic and mesencephalic zones.

Brain function in coordination of patterns

Protective reflexes in mammals include ear retraction to a position of safety—pressed against and somewhat behind the skull—as when a horse is seen to lay its ears back. Among the monkey-like bush babies (*Galagos*) the outer ear folds up laterally and longitudinally at the same time, under threat. The eyes are closed, and the muscles around the eye are contracted, adding to the protection. During this so-called startle reflex, breathing is checked, and the mouth corners are pulled back to expose the teeth; this prepares both for biting in defense and also for movements of the tongue and for head shaking to free the mouth of any dangerous or distasteful substance that may have been taken in. In most mammals, the limbs flex as if ready for a leap; in the human startle reflex, the arms are thrust outward as if ready to grasp at a support.

It is helpful to consider avoidance behaviour in terms of factors that elicit it (*e.g.*, specific stimuli) and regulate it (*e.g.*, hormones).

Factors in avoidance behaviour. *Specific stimuli.* Warning calls and visual signals that are unique to different species of birds and mammals effectively and specifically evoke avoidance patterns. In some cases, learning clearly emerges as a factor; thus, members of a colony of birds seem to learn to respond to the alarm calls of all species present in the colony. Among ducklings, a visual model

to evoke fleeing and hiding can be fashioned as a cardboard cutout. When moved overhead in one direction, the model resembles a short-necked, long-tailed hawk, and the ducklings flee from it; when moved in the other direction, the model looks like a harmless, long-necked goose, and the ducklings tend to stay calm. The model is effective, however, in eliciting the two kinds of behaviour only when the ducklings are accustomed to geese flying over but not hawks.

Innate factors also contribute to such responses (see INSTINCT). Domestic chicks, for example, show crouching and freezing in response to the long alarm call of their species. Many of the perching birds (passerines) will gather to mob when stimulated by the sight of an owl. The eyes in the characteristic owl face have been found to be especially important; even birds reared in isolation respond to man-made models with appropriate eyespots painted on. It has been suggested that many human beings are specifically (and perhaps instinctively) disturbed by the sight of snakes—the notion of a legless object perhaps being a key stimulus. Human responses to spiders and centipedes with conspicuous legs also may be intense. In the reaction to snakes at least, notwithstanding Freudian explanations that they symbolize male sex organs, the behaviour of people may be compared with owl mobbing among passerine birds.

Specific chemical signals can induce avoidance behaviour; some are released by minnows and tadpoles when their skin is damaged (usually indicating to fellows that there is danger). These chemicals appear to be specific for each species of fish and are highly effective in producing fleeing (see CHEMORECEPTION). Many ants produce volatile alarm substances (terpenes) that are attractants to other ants at low concentrations and, in high concentrations near their source, produce rapid locomotion, defense postures, and, sometimes, fleeing. Some invertebrate avoidance responses are reflexes evoked by very specific stimuli; rapid swimming by cockles clapping their shells, for example, is elicited by starfish extract. Shell jerking is produced in a freshwater snail (*Physa*) by contact with a leech, another specific response to a major predator.

Pain, startle, and novelty. Painful stimuli are pre-eminent among those that produce avoidance. Among mammals (including man) many such responses are patently inborn, as is the reflex withdrawal of one's finger from a hot griddle.

To classify a stimulus as startling or novel requires some comparison with previous stimulation. Human responses (orientation reflex) to startling or interesting stimuli may be studied by presenting a series of repeated tones; the orientation reflex tends to appear at the moment at which some change in the usual sequence (such as a longer or shorter tone) occurs. There is some evidence that the hippocampus (a brain structure) is involved in the human experience of novelty. Surgical removal of the hippocampus in many animals makes avoidance responses to strange objects far more persistent; a comparable operation in small parrots (lovebirds) greatly increases the persistence of calls that gather others for mobbing. Probably the hippocampus takes part in establishing memory of any new stimulus, and once this has occurred, the stimulus is no longer novel. Removal of other brain structures (the amygdala) reduces avoidance of strange objects (*e.g.*, in lovebirds) and also makes fleeing and defensive attack less likely.

Probable role of the hippo campus

Passive and active avoidance. Passive avoidance is achieved by the inhibition of a previously exhibited response. Thus, after a laboratory animal has learned to approach a food dish, it may then be punished by an electric shock whenever a selected visual or auditory stimulus is present. In passive avoidance, the animal may freeze as soon as the stimulus is given; in active avoidance, the animal is given the opportunity of fleeing.

Freezing proper entails general motor inhibition, which, if sustained, may pass into signs of reduced arousal. States of considerable loss of muscle tonus, of eye closure, and many signs of deep sleep have been variously termed feigning death or animal hypnosis. In very young fowl, such signs can be induced simply by holding the animal firmly if the experience is novel (and thus presumably frightening). Such states tend to occur as an alternative to fleeing when the apparently frightening stimulus is difficult to locate or to escape. Among social mammals (*e.g.*, cats or dogs) the status and confidence of an animal may be inferred from its degree of leg extension, arched back (vertebral tonus), and cocky tail elevation. Threat from which there seems no escape may induce a progressive approach to immobility and to general motor inhibition.

Punishment. Inhibitory interconnections have been postulated between the punishment and reward systems within the brain. One line of evidence suggesting a single punishment system rather than a number of them includes behavioral and neurological resemblances in the responses of animals to fear-inducing and to frustrating circumstances. If either fear or frustration is induced during conditioning, both produce resistance to extinction. Both are specifically opposed by barbiturate drugs.

Whatever its physiological basis, negative reinforcement (punishment) can induce in an animal both the inhibition of the response that produced the punishment and the avoidance of the location at which it occurred. Sometimes the tendency to show avoidance behaviour develops further with time, even without additional training. Thus, when being conditioned to discriminate between stimuli (*e.g.*, two tones), some breeds of dog (*e.g.*, Alsatians), if made to wait for food reward or given an impossible discrimination to perform, will howl and show great excitement. On later days, they may first resist mounting the conditioning stand and finally resist approaching the room to which earlier they ran eagerly, presumably for the rewards of food. Stimuli associated with the training room are sometimes said to act as secondary negative reinforcers (secondary punishments) in such a case.

Even a piece of cockroach nervous system (metathoracic ganglion) and the leg it controls have been shown to be capable of avoidance conditioning. If each contact of the leg with a water surface is paired with an electric shock, the leg comes to be retracted on contact with the water; no such change occurs in a control leg receiving the same number of shocks at random. The conditioning is accompanied by a very marked decrease in a chemical (acetylcholinesterase) found in the nervous system; since it greatly facilitates transmission of some nerve impulses, such a chemical may well be basic to this primitive kind of learning.

Avoidance conditioning of the leg of a cockroach

Hormonal effects. Male hormones (androgens) cause the performance of new mobbing calls in the breeding season by many male passerine birds (*e.g.*, chaffinch) and also some other birds (*e.g.*, farmyard cock); it is not certain whether the effect is specific to the vocalization or whether the hormone produces a general change in responsiveness to frightening stimuli. Female hamsters are initially faster than males to emerge from a box and also move about more in a strange place; perhaps females innately tend to be less nervous. Females behave more like male hamsters if given a small injection of male hormone (testosterone) in the second day of life; the adult difference survives castration, so it probably rests on sexual differentiation of the nervous system rather than on adult hormone levels.

The adrenocorticotropic hormone (ACTH) from the pituitary glands of many animals may facilitate avoidance behaviour. ACTH has other direct effects on the nervous system (*e.g.*, facilitating male sexual behaviour).

Functions of avoidance behaviour. *Fleeing and escape.* Most animals capable of locomotion show a rapid locomotor reflex to painful or startling stimuli. Such a reflex is very ancient in an evolutionary sense; it is present even in such primitive marine animals as the slender, tiny, translucent amphioxus. The rapid propulsion of an octopus or squid by its own jet of water or of a crayfish by a blow of its tail, the sudden leap and flight of a grasshopper, and the retraction of a worm into its hole—all are examples of such avoidance behaviour.

Reflex escape movements

Many invertebrates commonly compete in speed against

their vertebrate predators, which tend to have faster conducting individual nerve cells; in order to compete successfully, the invertebrates seem to have evolved giant nerves (bundle of individual cell fibres), for the broader a nerve is, the faster it conducts. Among such lower animals, perhaps one-third or more of the nerve cord running the length of the body is made up of fibres responsible for initiating the escape response of the species. The fibres of a cockroach, for example, activate a mechanism that produces rapid running when the rear end (anal cerci) is disturbed by air movements. Bony fishes also have such structures, the Mauthner cells, that initiate escape swimming when stimulated.

Escape may be facilitated not only by speed of response but also by its explosive onset (*e.g.*, after a period of shamming death), making it difficult for a predator to predict the behaviour of a prospective meal. Escape movements may stop as suddenly as they start. Many animals may even be especially conspicuous in escape (*e.g.*, showing coloured hind wings, as do some grasshoppers and moths), so that their disappearance appears even more sudden. Presumably, the predator, engaged in pursuing and tracking a moving prey, finds its difficult to shift quickly enough to a different kind of search and so is unable to localize the exact point of disappearance.

In many instances, rapid locomotion is enough to frustrate a predator; in others, direction is crucial (*e.g.*, a fish moving upward to the water surface or downward to the bottom or, among birds, a more elaborate celestial orientation). Under threat, insects such as pond skaters (*Vellia*) flee toward the nearest shore; beach fleas (amphipods) flee to the sea; and particular populations of ducks have a preferred compass direction for escape (so-called nonsense orientation).

Freezing. Immobility usually makes detection less likely. For stick insects and other animals resembling twigs or leaves, when immobility itself becomes conspicuous against moving foliage, the animals' compensatory swaying increases the camouflage effect. There seems to be an evolutionary conflict between camouflage and the need for conspicuous signals in communication. Social groups commonly keep in touch by calls or by movements such as tail flicks, which are inhibited during freezing or even under incipient immobility. Movements may be made conspicuous by patches of white or colour on a bird's outer tail feathers, or under a mammal's tail. The well-known white rear patch of hair among antelopes, for example, is hidden when the tail is folded or lowered under conditions of safety.

Protection reflexes, armour, and spines. Facial protective reflexes are usually well developed in flat-faced mammalian predators like cats and tarsiers, whose eyes and ears are especially exposed to injury by prey. The reflexes also are exaggerated in social species for use in communication; thus ear flattening in horse and dog displays has a counterpart in scalp retraction among Old World monkeys. The scalp movements and raised brows are effectively used in communication, despite the greatly reduced mobility of ears among monkeys. Limbs and other appendages (*e.g.*, antennae) are withdrawn or used to protect sensitive areas by both vertebrates and invertebrates. Among mollusks and such groups as sea squirts and barnacles, the whole soft body can be retracted into a protective shell, or carapace; a kind of door (operculum) may be used to stop the entrance (*e.g.*, among snails and barnacles), and trap-door spiders pull the stopper in place behind them. Bone may have evolved in fossil vertebrates as protective armour in jawless ancient fishes (ostracoderms), probably as a result of natural selection in the face of dominant arthropod predators (eurypterids). With the evolution of jaws, the vertebrates themselves gave rise to nearly all later large predators. Evolutionary advantage then apparently came with complex sense organs and behaviour; in most vertebrate lines there is evidence of a progressive reduction in body armour (dermal bone). Thus, although such cartilaginous fishes as sharks and rays do not exhibit such bony skins, they may well have evolved from heavily armoured ancient fishes (placoderms).

The completely armoured turtles and tortoises

Armour nevertheless has evolved repeatedly, particularly among animals incapable of fast locomotion; trunkfish (boxfish), for example, have a body entirely boxed by bony plates; and tortoises and turtles are perhaps the most completely armoured of four-legged animals. The turtles seem to have evolved early from the basal stock of the reptiles; thanks to the shell (carapace) within which they can withdraw head, limbs, and tail, they represent one of the few reptilian orders that have remained consistently successful. The turtle's dorsal carapace appears to consist of newly evolved plates of dermal bones, but the belly plates (ventral plastron) may well be retained in part from fish ancestors. Reptilian land vegetarians usually tended to evolve armour, as in the fossil dinosaurs such as stegosaurs and ankylosaurs.

South American toothless animals (edentates) such as anteaters are probably survivors of a comparable early development in mammals. The armour of armadillos and the presence of bony plates in the skin of the extinct sloths suggest that the whole group may derive from an armoured ancestor. The appearance of hair in the mammal line, partly inferred from the presence of fleas in early evolutionary periods, seems to have led to the evolution of a light, spiny type of armour. Such modern mammals as hedgehogs, echidnas, insect-eating tenrecs, and some rodents and their relatives (lagomorphs) all possess defensive spines that are commonly erectile and are often able to roll into a ball like an armadillo.

Chemical means of defense may be widely distributed in the body, making the animal distasteful to predators. Some of these chemical compounds may be derived from plants eaten or synthesized by the animal itself (*e.g.*, bufotoxin in toads). Although the animal attacked may be killed and thus not benefit, his fellows do since they are likely to be avoided by the predator. Poison or distasteful substances may also be ejected from a bodily reservoir and squirted at the enemy (*e.g.*, the skunk, some ants) or inserted into a puncture made by a spine (*e.g.*, triggerfish) or teeth (*e.g.*, certain snakes). Many poisons act to paralyze muscles by blocking nerve transmission at the neuromuscular junction (*e.g.*, cobra venom).

Warning behaviour. Mobbing behaviour apparently advertises the presence of a predator that is potentially but not immediately dangerous; thus mammalian nest predators can be safely mobbed by flying birds, as can owls in the daytime. From the safety of trees, such mammals as monkeys and squirrels mob predators on the ground. Mobbing calls are typically easy to locate, the calls being short and staccato, and they provide excellent cues of distance and direction (see SPACE PERCEPTION: *Auditory cues;* PERCEPTION OF MOVEMENT: *Auditory*). Conspicuous movements, such as tail flicks among small birds and squirrels, accompany the calls.

More urgently, intense warning behaviour is given in response to sources of immediate danger (*e.g.*, hawks actively hunting). Under these circumstances, warning calls are usually long whistles that make location difficult because of their gradual onset and termination and their narrow ranges of pitch. The evolution of warning behaviour that puts the displaying animal in danger (such as these intense warning calls) seems likely to come about only if the benefit to offspring and other members of the species is great. Indeed, it has been calculated that if an individual loses his life as the result of his warning behaviour, increased transmission of his family's genes will result only if the reproductive rate of relatives (as close as sisters) is doubled as a result of his sacrifice.

BIBLIOGRAPHY. R.A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology*, 2nd ed. (1970), comprehensive and detailed coverage of the theory of animal behaviour that may be too difficult in parts for the lay reader; S.P. GROSSMAN, *A Textbook of Physiological Psychology* (1967), an unusually clear treatment of this field, although now somewhat outdated.

(R.J.A.)

Azerbaijan Soviet Socialist Republic

One of the 15 union republics of the Soviet Union, the Azerbaijan Soviet Socialist Republic, or Azerbaijan

S.S.R., is the easternmost of the three such entities—the others are Armenia and Georgia—that occupy the area fringing the southern flanks of the Caucasus Mountains. To the south lies Iranian Azerbaijan, to the east the waters of the Caspian Sea. In addition to a variegated and often strikingly beautiful natural environment, the Azerbaijan of the later 20th century offers a blend of traditions and modern development. The proud and ancient people of its remoter areas retain many distinctive folk traditions, but the lives of its 5,600,000 or so inhabitants (the population has more than doubled since the 1920s) have been touched by an accelerating modernism characterized by industrialization, the development of power resources, and the growth of the cities, in which half the people (though only 42 percent of the Azerbaijanis) now live. Industry dominates the economy, and more diversified pursuits have supplemented the exploitation of oil, which made Azerbaijan the world's leading producer at the beginning of the 20th century. Fine horses and caviar continue as some of the more distinctive traditional exports of the republic.

Administratively, the Azerbaijan S.S.R., which was established on April 28, 1920, now includes the geographically separate Nakhichevan Autonomous S.S.R., which, with its capital, Nakhichevan, lies beyond an intervening strip of Armenian territory. The Nagorno-Karabakh Autonomous Oblast, of which Stepanakert is the administrative centre, is also an administrative division. The territory of Azerbaijan covers 33,400 square miles (86,600 square kilometres), larger than Sri Lanka and about the same size as Portugal. Its capital is the ancient and economically important city of Baku, whose harbour is the best on the Caspian Sea. For further details see CAUCASUS MOUNTAINS; CASPIAN SEA; RUSSIA AND THE SOVIET UNION, HISTORY OF.

THE LAND

Landscapes. As a result of its broken relief, its drainage patterns, and its climate, Azerbaijan is characterized by a variety of landscapes. More than 40 percent of its territory is taken up by lowlands, about half lies at 1,300 to 4,900 feet (400 to 1,500 metres), and areas above 4,900 feet (1,500 metres) occupy a little more than 10 percent of the total area.

The mountain regions. The highest peaks are Bazardyuzi, Shakhdag, and Tufan, all part of the Great Caucasus, which forms a natural northern boundary for the republic. Magnificent spurs and ridges, cut into by the deep gorges of mountain streams, make this part of Azerbaijan a region of great natural beauty.

Forests and wildlife

The slopes of the mountains are covered with beech, oak, and pine forests, and the animal life includes Caucasian deer, roe deer, wild boar, brown bear, lynx, European bison (wisent), chamois, and leopard, though the latter is rare. Typical birds include the Caucasian grouse and the stone partridge.

The spurs of the Little Caucasus, in southwest Azerbaijan, form the second important mountain system, which includes the Shakhdag, Murovdag, and Zangezur ranges and also the Karabakh upland. The large and scenic Lake Gyoygyol lies at an altitude of 5,138 feet.

The southeast part of Azerbaijan is bordered by the Talysh Mountains, with Kyumyurkyoy as the highest.

The Kura-Araks Lowland. This vast territory is named after the main river and its tributary. The Shirvan, Milskaya, and Mugan plains are part of this lowland and have similar soils and climate, the difference in names reflecting purely historical considerations. Plant life is that of the steppe and semidesert, and gray soils and saline solonchaks and, in higher regions, gray alkaline solonetz and chestnut soils prevail. A well-developed network of canals between the Kura and Araks rivers makes it possible to irrigate a major part of the lowland. The Upper Karabakh Canal, 109 miles (175 kilometres) long, provides a vital link between the Araks River and the Mingechaur Reservoir on the Kura River. The reservoir, constructed in 1953, contains 565,000,000,000 cubic feet (16,000,000,000 cubic metres) of water and has a surface area of 234 square miles (605 square kilometres) and a maximum depth of 246 feet (75 metres). An associated hydroelectric power plant has a capacity of 360,000 kilowatts. The Upper Karabakh Canal alone irrigates about 250,000 acres (100,000 hectares) of fertile land, and in addition supplies the Araks River with water during dry summer periods.

The Upper Shirvan Canal, the second most important, is 75 miles long and also irrigates about 250,000 acres.

Climate. The dry subtropical climate prevailing in central and eastern Azerbaijan is characterized by a mild winter and a long (four to five months) and very hot summer, with an average temperature of 81° F (27° C) and a maximum temperature of 109° F (43° C).

Southeast Azerbaijan is characterized by a humid subtropical climate and has the highest precipitation in the republic, reaching 47–55 inches (1,200–1,400 millimetres) a year, most of which falls in the cold months.

A dry continental climate, with a cold winter and a dry, hot summer, prevails in the Nakhichevan Autonomous S.S.R. at altitudes of 2,300 to 3,300 feet (700 to 1,000 metres). Moderately warm, dry or humid types of climate are to be found in other parts of Azerbaijan. The mountain forest zone has a moderately cold climate, while an upland tundra climate characterizes the altitude of 9,850 feet (3,000 metres) and above. Frosts and heavy snowfalls make the passes at such altitudes inaccessible for three or four months of the year.

THE PEOPLE

Cosmopolitan ethnic heritage

Ethnic composition. Today's Turkic Azerbaijanis, who comprise about 74 percent (up from 68 percent in 1959) of the republic's population, combine in themselves the predominantly Turkic strain, which flooded Azerbaijan especially during the Oğuz Seljuq migrations of the 11th century, with mixtures of older inhabitants—Iranians and others—who had lived in Transcaucasia since ancient times. In the Nakhichevan A.S.S.R., about 94 percent of the 219,000 (1974) inhabitants are Azerbaijanis, whereas more than 80 percent of the 154,000 people of the Nagorno-Karabakh Autonomous Oblast are Armenian. These subunits of the union republic were established to minimize friction between the two nationalities.

Demographic trends. Russians make up the largest minority in Azerbaijan, in 1970 comprising 10 percent, down from 13.6 percent in 1959, an unusual trend not found in most of the Soviet Union. Population estimates showed the republic's population to have risen from 5,117,081 at the census of 1970 to about 5,606,000 in 1975. The average population density reached 168 per square mile (65 per square kilometre) by the mid-1970s. Urban population constitutes half the republic's total, and 42 percent of the Azerbaijanis themselves live in cities (as do 92 percent of the 510,000 Russians). The 42 percent of Azerbaijanis urbanized is a larger proportion than is common in non-Slavic republics; of the 1,359,000 people (1974 estimate) in the Greater Baku metropolitan area, about 46 percent are Azerbaijanis. Other cities are Kirovabad, formerly Ganja (190,000 in 1970), Sheki (43,000), Lenkoran (36,000), and Nakhichevan (33,000). New cities established during the 1950s and '60s include Sumgait (124,000), Mingechaur (43,000), Ali-Bayramly (34,000), and Stepanakert (30,000).

ECONOMY

Azerbaijan is a developed industrial and agrarian republic, with the balance between the gross output of agriculture and that of industry gradually tilting in favour of the latter; in 1960 gross industrial output accounted for 71 percent and agriculture for 29 percent of the total output; in 1970 the corresponding figures were 83 and 17.

New industrialization

The emphasis on heavy industry has considerably expanded two traditional industries—petroleum and natural gas—but engineering and light and food industries are also of growing importance.

Natural resources. At the turn of the 20th century Azerbaijan was the world's leading petroleum producer, and it was also the birthplace of the oil-refining industry. In 1901, for example, Azerbaijan produced 11,400,000 tons of oil, more than the United States; it accounted

for 95 percent of Russian and 50 percent of the world production. As the 20th century progressed, however, Azerbaijan's role in oil production decreased as the industry developed in other regions of the Soviet Union and the world, though by the mid-1970s the annual output of oil in Azerbaijan amounted to 19,500,000 tons.

Azerbaijan also has other natural resources, including gas, iodo-bromide waters, lead, zinc, iron, and copper ores, nepheline syenites utilized in the production of aluminum, common salt, and a great variety of building materials, including marl, limestone, and marble.

Fuel and power. The development of Azerbaijan's industry created a demand for fuel and power supplies. By the mid-1970s, the yearly electric-power output of the republic had reached 14,200,000,000 kilowatt-hours. Almost 90 percent of the electricity is produced at thermoelectric power stations, which have been built throughout the republic. The largest of these are Ali-Bayramly (capacity 1,120,000 kilowatts); the northern station in Baku (319,000 kilowatts); and Sumgait (470,000 kilowatts). Hydroelectric stations, among which is that at Mingechaur (2,400,000 kilowatts), produce a further 1,500,000,000 kilowatt-hours, and new hydroelectric stations are planned for the 1970s and '80s on the Kura and Araks rivers.

Industry. Azerbaijan has a diversified industrial base, with heavy industry and its leading branches—power, manufacturing, and chemicals—predominating.

Branches of the processing industry, producing mineral fertilizers, gasoline, and kerosine, herbicides, industrial oils, synthetic rubber, plastics, etc., are developing. Sumgait has emerged as the major centre of this industry, as well as of ferrous metallurgy.

The republic's share in the total output of Soviet manufacturing has increased considerably in the third quarter of the 20th century. Azerbaijan manufactures equipment for the oil and gas industry, electrical equipment of all kinds, and many appliances and instruments. Some of these goods are exported to other countries. This type of industry is located mostly in Baku, Kirovabad, and Mingechaur, and there are plans to extend it also to Agdam, Kazakh, and Sumgait.

Light industrial manufactures include cotton and woollen textiles, knitwear, traditional household items and souvenirs, footwear, and other consumer goods. The cities of Sheki, Stepanakert, Kirovabad, Mingechaur, and Baku are the main centres of this industry. Food processing plants are distributed fairly evenly throughout the republic. Azerbaijan fisheries are of particular importance because of the sturgeon of the Caspian Sea; sturgeon roe is made into caviar of world renown.

Agriculture. Azerbaijan's agriculture has also been developing: although arable land in Azerbaijan constitutes only 7 percent of the total, the republic accounts for 10 percent of the gross agricultural output of the Soviet Union. Raw cotton is the leading agricultural product. In 1913 cotton production amounted to 64,000 tons, in 1940 to 154,000 tons; the 1974 figure of 584,000 tons made Azerbaijan one of the leading cotton-producing areas in the Soviet Union. Tobacco is the second most valuable crop, and the republic's annual output, about 42,000 tons (1973), constitutes 8 percent of the total Soviet yield. Favourable conditions for grapes have contributed to the development of this branch of agriculture: Azerbaijan produced 5,500 tons of grapes in 1913 and 100,000 tons in 1960. The ensuing years witnessed a further increase, with the yield reaching 174,000 tons by 1965 and 480,000 tons by 1973. During the 1970s, production of grapes is expected to reach 770,000 tons annually. Most of the grape varieties grown in Azerbaijan are used for wine making. Tea was introduced into the republic fairly recently, and the necessary prerequisites exist that make it possible to enlarge the output. Vegetables, particularly early varieties, fruit, walnuts, and hazelnuts are also promising crops. Special attention is being paid to extending the areas under pomegranates. Some districts, particularly those of Sheki, Zakataly, and Geokchay, are—as they have been traditionally—engaged in silkworm breeding.

Development of viticulture

High commodity output is not characteristic of Azerbaijan's animal husbandry. More than 5,000,000 sheep and goats form its main livestock, while cattle number some 1,600,000. Riding horses, especially the valuable Karabakh breeds, are exported.

Economic regions. *The Apsheron region.* This includes the Apsheron Peninsula and several other areas of eastern Azerbaijan. As a result of its advantageous geographical position, it is crossed by freight routes connecting Azerbaijan and the whole of Transcaucasia with the North Caucasus and Central Asia. Numerous highways also run from the peninsula to every corner of the republic.

On the shores of the Caspian Sea, the Apsheron region nevertheless remains one of the most arid parts of Azerbaijan. Its main natural wealth is mineral, including oil, natural gas, iodo-bromide waters, and limestone, used in building and cement production.

Such cities as Baku, the capital of the republic, Sumgait, and other industrial centres make the Apsheron region one of the most highly industrialized and densely populated areas of Azerbaijan.

Baku

Baku itself owes its modern growth to the development of the oil industry; oil derricks encircle the city, and it acts as a magnet to workers from many areas. In 1886 the population of Baku had already reached 86,000, against 6,000 in 1830, and by 1974 it exceeded 910,000; its metropolitan area in that year had a population of 1,360,000, or a quarter of the total for the entire republic. Contemporary Baku is one of the largest and most attractive cities in the Soviet Union. Situated on natural terraces running down to a gulf of the Caspian Sea, the city has a two-mile-long picturesque boulevard and many historic sites.

Modern Sumgait, 22 miles northwest of Baku, is currently a centre of the iron and steel, nonferrous metallurgical, and chemical industries, although the development of light engineering is envisaged.

The Lenkoran region. This area of southern Azerbaijan is well endowed by nature, with evergreen vegetation and thick beech and oak forests. Warm climate crops, such as tea, feijoa (a fruit-bearing shrub), rice, grapes, tobacco, and citrus trees, flourish there. The region is also becoming one of the largest producers of spring and winter vegetables in the Soviet Union.

Mild winters draw many birds to the Caspian coast, and a reservation provides a resting home for flamingos, swans, pelicans, herons, and buzzards.

The towns of Lenkoran, Astara, and Massaly are small, and the Talysh, or Talīshī—Iranian people who form the bulk of the local population—have preserved their old customs and traditions. Industry is mostly concerned with the processing of agricultural goods, while in the mountains the Talysh make colourful rugs and carpets.

The Kuba-Khachmas region. This lies to the north of Apsheron. Its coastal lowlands specialize in grain and vegetable production, while vast orchards surround the towns of Kuba and Kusary. The mountain slopes are used for grazing. Special kinds of sheep are bred in order that their skins may be utilized in the fur industry.

The Shirvan region. An industrially and agriculturally developed part of Azerbaijan, this area is centred on the Shirvan Plain. The Mingechaur hydroelectric station is located here, and a new thermoelectric station was under construction in the 1970s. The area also has a well-developed network of highways. Industry is generally engaged in the processing of such agricultural products as cotton, grapes, and fruit. The most important vineyards lie in the vicinity of Shemakha, a town famed for its wines, notably Matrasa and Shemakha, respectively dry red and sweet. In Kyurdamir a fragrant dessert wine of the same name is produced. The best varieties of pomegranates are grown near Geokchay.

The Mugano-Salyany region. Lying south of the Kura River and within the boundaries of the Mili and Mugan plains, this area specializes in cotton growing, producing about 70 percent of the gross cotton output of the entire republic. Cotton-ginning plants are located in Barda, Salyany, and Ali-Bayramly, all of which, in addition to

Cotton production

being located on the Kura River, have the advantage of being junctions of railways and motor roads. A thermal power station stands near Ali-Bayramly.

The Southwestern region. This includes the Nagorno-Karabakh Autonomous Oblast, as well as Lachin, Fizuli, and Kubatly administrative districts. As the average altitude is 4,900 feet (1,500 metres), it is one of the areas in the republic where broken relief impedes the development of transport, industry, and agriculture. Agricultural production is concentrated in the mountain valleys. Animal husbandry constitutes 60 percent of the gross agricultural output, the leading branches being sheep and pig raising. Grapes, tobacco, and grain are the main crops.

Armenians form the greater part of the population. The Nagorno-Karabakh Autonomous Oblast is well known in the Soviet Union for the longevity of its people; many have a life-span of more than 100 years.

The Kirovabad-Kazakh region. Occupying a special place in the plans for the future development of the republic, this region is situated in the centre of Transcaucasia at the junction of the Azerbaijan, Armenian, and Georgian republics. The region has conditions favourable both for human life and for intensive agriculture. Trade routes have crossed this part of Azerbaijan from time immemorial, and the ancient town of Ganja (now Kirovabad) was founded here. It is an important industrial centre, with light, food, engineering, chemical, and nonferrous metallurgical industries. Naftalan is a health resort.

The Sheki-Zakataly region. This area includes the towns of Sheki (formerly Nukha), Zakataly, and Belokany. Its territory borders on the Great Caucasus, which shields it from cold northern winds. Numerous mountain rivers provide it with an ample supply of water, and the region is densely populated. Agricultural products include tobacco, aromatic plants (mint, basil, and roses), rice, corn (maize), and various fruits. The area is also the largest Soviet producer of hazelnuts and walnuts.

The Nakhichevan region. This is a typical semidesert, although a system of irrigation makes it possible to cultivate grapes, cotton, and grain. There are several sources of mineral water in the foothill areas of the Nakhichevan A.S.S.R.

TRANSPORTATION

Railways

Most of the rivers of Azerbaijan are not navigable, and most freight—including that sent out of the republic—is carried by rail. Considerable portions of the rail network are electrified, and total track mileage is about 1,150. Annual freight turnover is about 66,000,000 tons, the principal goods carried being oil products, building materials, timber, and grain. The major railway lines go through the Kura Valley and connect Baku with Tbilisi and Batumi in the Georgian S.S.R.

Motor transport is used extensively for both freight and passengers within the republic. The total length of highways has reached 13,400 miles, and the annual freight turnover by truck is some 3,090,000,000 ton-miles (4,520,000,000 ton-kilometres). Highways connect various parts of the republic and are often the only means of land communication between some of the remote mountain districts and the administrative centres and large cities.

Baku, on the Caspian, is one of the busiest seaports in the Soviet Union, handling such important and vital goods as oil, timber, grain, and cotton. A ferry link (1962) between Baku and Krasnovodsk (also on the Caspian, in the Turkmen S.S.R.) has increased considerably the amount of cargo passing through Azerbaijan.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Constitutional and political framework. The Azerbaijan constitution (1937) declares it to be one of the 15 constituent republics forming the Soviet Union. Though sovereignty and equality are claimed for it, both are limited by the fact that the republic lacks independence in foreign affairs, military matters, economic planning, domestic political control, and cultural and ideological questions, in all of which it must defer to the central Soviet authorities in Moscow. At the top of the governmental structure within Azerbaijan today are the republic's Supreme Soviet, selected from a single slate of candidates for a four-year term, and its Presidium, selected by the Supreme Soviet. The duties of both are largely ceremonial rather than legislative. The highest executive and administrative body of the republic's government is the Council of Ministers. Local branches of government are the *oblast* and city soviets. The judicial system includes a Supreme Court and lower courts.

Communist Party

The Communist Party of Azerbaijan, a subordinate branch of the Communist Party of the Soviet Union (CPSU), is the only authorized political organization. Among the 249,670 members (1970) in the republic, Azerbaijanis comprised 66 percent, noticeably less than their proportion (almost three-quarters) of the whole population of the republic. Armenians made up almost 14 percent of the Azerbaijan party membership, and Russians 12.4 percent, though they constituted only 9.4 and 10 percent of the population, respectively. Small minorities made up the remainder. The closely connected Azerbaijan Komsomol (Young Communist League) has a membership of 432,000. Labour union membership totals about 1,242,000; the unions are not concerned with relations between workers and management except in connection with such matters as incentives to increase production.

Social services. The republic has a well-established health service with large specialized clinics and medical research institutes. More than 1,000 hospitals and polyclinics are active. Medical service, provided without direct charge to patients, is supported by general taxation on individual employees and workers and by taxes from income of factories and other firms.

The network of medical and prevention establishments is growing: by the mid-1970s there were 27 doctors, 44 paramedical personnel, and 95 hospital beds for every 10,000 people.

Science and culture. In the course of its long history, Azerbaijan has given the world a number of outstanding thinkers, poets, and scientists. Among the medieval scientists and philosophers, Abul Hasan Bakhmanyar (11th century), the author of numerous works on mathematics and philosophy, and Abul Hasan Shirvani (11th–12th centuries), the author of *Astronomy*, may be noted. The poet and philosopher Nẹzāmī, called Ganjavī after his place of birth, Ganja (now Kirovabad), was the author of *Khamseh* ("The Quintuplet"), composed of five romantic poems, including "The Treasure of Mysteries," "Khosrow and Shīrīn," and "Leyli and Mejnūn."

The people of Azerbaijan have kept their ancient musical tradition. The art of *ashug*s, who improvise songs to their own accompaniment on a stringed instrument called a *kobuz*, remains extremely popular. *Mugam*s, vocal and instrumental compositions, are widely known, the town of Shusha being particularly famous for this art.

In the Soviet period, illiteracy has been virtually eradicated, and a network of institutes of higher education, research centres, theatres, and similar bodies has been established. By 1971 there were 13 institutes for higher education in Azerbaijan with more than 100,000 students. The largest is the Azerbaijan Institute of Petroleum and Chemistry, in whose seven departments more than 17,000 students attend classes.

Education at all levels is supported by taxes levied upon working people and firms and is available without direct tuition charge.

The Azerbaijan Academy of Sciences was established in 1945 and in 1975 had 51 full members and 44 corresponding members. The academy coordinates the activity of 30 research centres, including institutes of cybernetics, physics, theoretical problems of chemical technology, petrochemical processes, and genetics.

Since World War II a number of new cultural institutions have been established. By the mid-1970s the republic had 13 state theatres, half of them in Baku, about 2,500 clubs, 38 museums, and 3,200 libraries with about 25,300,000 volumes. The opera and ballet are widely at-

tended. Some of Azerbaijan's composers, notably Uzeir Hajjibekov (the operas *Ker-Ogly* and *Leyli and Mejnūn* and the operetta *Arshin Mal'Alan*) and Kara Karayev (the ballets *Seven Beauties* and *The Path of Thunder*), have international reputations. The latter's symphonic music is also well known abroad.

Publishing

Azerbaijan has no private publishing; several government firms publish scientific books and magazines as well as books and magazines about art and literature in Azerbaijani, Russian, and other languages. Of the 1,283 book titles issued in 1973, 70.5 percent appeared in Azerbaijani and most of the rest in Russian. Copies of books in Azerbaijani per person fell from 2.8 annually around 1959 to 2 about 1970; the reason for this probably was the striking increase (51.4 percent) in the republic's Azerbaijani population and a failure to provide more copies to meet the increase.

Magazines

The magazines *Literaturny Azerbaidzhan* (in Russian), *Azerbaijan Gadïnï* ("Azerbaijan Woman," in Azerbaijani), and *Azerbaidzhanskoye neftyanoye khozyaistvo* ("Azerbaijan Petroleum Economy," in Russian) have the highest circulation.

Newspapers of the republic appearing in the Azerbaijani language numbered 92 in 1973 out of a total 116 in all languages.

Baku has several radio stations, a television studio, and a film studio.

PROSPECTS

Important changes have taken place in the structure of the economy of the republic over the past decades, with the industrial rate of growth surpassing that of agriculture. According to the 1971–75 plan of economic development, adopted by the 24th Congress of the Communist Party of the Soviet Union, a more rapid development of engineering and chemical industries was envisaged, and the latter's output was to rise steeply with the completion of a new chemical plant in Sumgait. A new microbiological industry was also to come into existence, making artificial fodder out of by-products of oil processing. In agriculture, new irrigation and drainage schemes were planned.

New residential quarters were to be built in Baku and to be connected with the city centre by new subway lines already partially in operation. The water supply of the capital was to be improved and increased with the building of a new water-supply system bringing water from the Kura River and the Samur-Apsheron Canal.

Among the problems facing the republic is the difficulty found by Azerbaijani intellectuals and executives in placing their educated, talented students in employment, commensurate with their abilities, in the urban areas, especially Baku. Despite the fact that the percentage of non-Azerbaijanis living in the republic dropped between 1959 and 1970, Russians and Armenians, the two largest minorities, grew in absolute numbers in the period. In Azerbaijan, Russians, Armenians, Jews, and Tatars live and work largely in the cities. Their presence in the urban work force creates a dilemma for Azerbaijani authorities seeking to advance their own graduates and at the same time to avoid ethnic tension and discrimination.

BIBLIOGRAPHY. For further information on Azerbaijani life, see FIRUZ KAZEMZADEH, *The Struggle for Transcaucasia, 1917–1921* (1951); EDWARD ALLWORTH, "La Rivalité entre le Russe et les langues orientales dans les territoires asiatiques de l'U.R.S.S.," *Cahiers du Monde Russe et Soviétique*, vol. 7, no. 4 (1966), pp. 531–563; DAVID M. LANG, "Religion and Nationalism; A Case Study: The Caucasus," *Survey* No. 66 (January 1968), pp. 33–47; ALEC NOVE and J.A. NEWTH, *The Soviet Middle East: A Model for Development?* (1967); ZEV KATZ *et al.* (eds.), *Handbook of Major Soviet Nationalities* (1975).

(E.D.S./E.Al.)

Azov, Sea of

An inland sea of the Atlantic Basin, the Sea of Azov (Azovskoye More in Russian, spelled Azovskoje More in the transliteration system of the Akademiya Nauk) is situated on the southern shores of the European portion of the Soviet Union. Separated on the south from the Black Sea by the Kerch and Taman peninsulas, it meets the Black Sea at the Kerch Strait.

Physical characteristics. Its area is 15,000 square miles (38,000 square kilometres), its mean depth 26 feet (eight metres), and its maximum depth 46 feet. It is thus the world's shallowest sea. Mean water volume is 77 cubic miles (320 cubic kilometres). The west, north, and east shores are low-lying, with sandstone deposits containing mussel fossils; the south shores are mainly high and jagged and composed of limestone. Characteristic features of the shores are alluvial mussel-fossilized spits, which separate many shallow bays and the estuaries from the sea. Some of the estuaries are connected directly with the sea; others are separated from it by sandy overspills. Into the Sea of Azov flow the great rivers Don and Kuban and many lesser ones such as the Mius, the Berda, the Obitochnaya, and the Yeya.

The topography of the bottom is even but has a minute incline toward the centre. Bottom deposits are composed of sand, mussel fossils, and silt. Certain parts of the southern shore are marked by rocky soil. The geological history of the Sea of Azov is inseparably connected with that of the Black Sea. The Sea of Azov formed in the Middle Miocene Epoch (16,000,000 years ago) as part of a wide basin. In the Early Pliocene (beginning 7,000,000 years ago), the basin of the Black Sea and the Sea of Azov was isolated from the ocean. In the subsequent period, connections with the Caspian Sea, through the Manych Trench, and with the Mediterranean Sea were formed and broken several times. In the Middle and Late Pliocene the Cimmerian and Kuyalnits basins formed and evolved into the Chaudin Basin on the sites of the Sea of Azov and the Black Sea; they had a connection with the Caspian Sea at the time but evidently were isolated from the Mediterranean. At the beginning of the Quaternary Period (2,500,000 years ago) the Sea of Azov acquired roughly the outline it has today. Toward the end of this period a connection between the Black Sea and the Mediterranean Sea was established (a stage known as the Karangats Basin).

Former connection with the Caspian Sea

Climate. The climate of the Sea of Azov is continental and temperate. In the winter, northeast and east winds predominate, bringing cold air from the mainland. The mean January–February air temperature ranges from 30° F (−1° C) in the south to 21° F (−6° C) in the north, the minima being around −27° F (−33° C). Under the influence of cyclones, heavy frosts often alternate with thaws, and there are frequent fogs. July is the warmest month, with a mean temperature of 75° F (24° C). The maximum temperature is around 104° F (40° C).

Hydrology. The hydrological regime is determined by the continental situation of the sea, the climate, the influx from the rivers, and the water exchange with the Black Sea, as well as by the water-related activity of man in the basin. The yearly water balance of the Sea of Azov in the centre consists of 9.5 cubic miles from river inflow and 3.2 cubic miles of precipitation, minus evaporation of 8.1 cubic miles; 4.2 cubic miles of surplus freshwater flows out through the Kerch Strait. As a result of the freshwater surplus the mean salinity is only 11 parts per thousand in the south, 9–10 in the rest of the sea, and 2–4 in the inlet areas. In certain years, because of changes in the water turnover through the Kerch Strait and changes in the river influx, salinity may rise to 12–13 parts per thousand in the south, and to 11–11.5 in the north, or may sink to 10.5 in the south and 8–9.5 in the north. Sivash Lake is particularly noted for salinity. Here, as a result of heavy evaporation and a weak link with the sea (through the Tonky Strait), the salinity reaches 124–166 parts per thousand.

Extensive water freshening and low air and water temperatures cause an annual freezing. The ice forms in November–December in Taganrog Bay and spreads southwest, reaching its fullest extent in February. Thawing occurs in March and April.

Currents and tides

Currents flow in a counterclockwise rotation, but under the influence of winds they may reverse. They may reach a speed of 10 to 12 inches per second. The mean water

level of the sea varies as much as 13 inches from year to year, according to the river influx. Tidal oscillations of the water level may reach 18 feet. The oxygen content of the water is ten to 11 millilitres per litre in winter and five to six millilitres per litre or more in summer. In summer, during extended calms, the oxygen on the bottom may dissipate entirely through oxidation, causing the fish to die. The Sea of Azov is rich in nutrient salts; but throughout the year the nutrient-salt content varies radically with river influx and consumption of salts by living organisms.

Animal life and vegetation. There is an extraordinarily high level of biological productivity, resulting from the sea's shallowness, the excellent mixing and even warming of the water, and the input of great quantities of nutrient material by the rivers. Its varieties of phytoplankton total 188, and 25–30 varieties of macrophytes are found. Among flowering plants zostera is common. The fauna includes 324 invertebrate species and 79 species of fish, including sturgeon, perch, bream, herring, sea-roach, gray mullet, minnow, shemaja, and bullheads. Sardines and anchovies are particularly abundant. Local fisheries catch an average of 1,500,000 tons per year.

Navigation. The Sea of Azov handles much freight and passenger traffic. The progress of heavy oceangoing craft is hampered by shallowness at some points. Icebreakers assist in winter navigation. Principal ports are Taganrog, Zhdanov, Yeysk, and Berdyansk. (A.M.Mu.)

Aztec Religion

Aztec religion consists of the mythology, beliefs, and practices of the people of the Aztec Empire in pre-Spanish Mexico. According to their own traditional history, the Aztecs, a Nahuatl-speaking tribe, departed in AD 1168 from Aztlán, somewhere in northwestern Mexico or the southwestern United States, where they had dwelt for more than a thousand years. When they tried to settle in the central valley of Mexico, among civilized states such as Colhuacán, Azcapotzalco, Texcoco, in the 13th century, they were treated as semibarbarous intruders. Poor and landless, they managed to build Mexico-Tenochtitlán, a village of reed huts, on some islands and sandflats in the lakes. It was not until the reign of Itzcóatl (1428–40) that they began their ascent to the leadership of the League of the Three Cities (Mexico, Texcoco, Tlacopan), which eventually became what is now called the Aztec Empire.

In the process of building that empire through war, diplomacy, and trade, the Aztecs became involved with various aboriginal ethnic groups, whose godheads, myths, rites, and beliefs they readily borrowed. Their religion appeared as a synthesis that combined many features from different cultures.

SOURCES

Aztec religion is known through a large number of documents, which may be divided into five categories:

Early accounts written by the conquistadors. Important among these are the *Cartas de Relación* ("Letters of Information"), sent by Hernán Cortés to his emperor, Charles V, and the *Historia verdadera de la conquista de la Nueva España* ("True History of the Conquest of New Spain") by Bernal Díaz del Castillo. Rites, ceremonies, temples, paraphernalia of the cults, are often described in these accounts. Their value, however, is lessened by the writers' ignorance of the Aztec language, their lack of understanding of the Indian way of thinking, and their deep hostility to the native religion, which they considered to be inspired by the devil. These documents, therefore, should be interpreted with utmost caution.

Early accounts by Roman Catholic missionaries. Paradoxically enough, the priests generally showed more understanding and tolerance than did the laymen. Thanks to their training and theological knowledge, they were able to analyze the Indian mind and to gain insight into the meaning of the myths and ritual. They believed that the native gods were devils and that the whole religion was a trap set by Satan to destroy the souls of the Indians. Yet most of them (for example, Toribio de Motolinía) were moved by a sincere compassion toward the aboriginal population and did their best to observe and describe its way of life. The missionaries, as a rule, learned the native languages, especially Aztec.

The most illustrious of these priests was Father Bernardino de Sahagún. He arrived in Mexico very early (1529), learned the Nahuatl tongue, and spent his life building a wonderful monument, a real encyclopaedia called *Historia general de las cosas de Nueva España* ("General History of the Things of New Spain"). Sahagún deeply loved "his sons the Indians." Most of his work was devoted to an incredibly detailed account of the beliefs and rites, dictated to him in Nahuatl by Aztec noblemen and priests. Although as a Catholic monk he severely condemned their religion, he admired the Aztecs' high standards of morality, the virtues of the priests, the ideals of the rulers, the knowledge of the native thinkers. Thus the Nahuatl text of Sahagún's "History" may be said to provide a truly boundless treasury of mythological lore, descriptions of the rites, theological concepts, and religious poetry.

The Aztec encyclopaedia

Aztec sacred books. These works, known as codices, which were kept in the temples, were written (or rather, painted) on deerskin or agave-fibre paper, by scribes *(tlacuiloanime)*, who used a combination of pictography, ideograms, and phonetic symbols, and dealt with the ritual calendar, divination, ceremonies, and speculations on the gods and the universe. Most of the texts were destroyed after the Conquest, but some outstanding specimens have survived, such as the Codex Borbonicus, the Codex Borgia, the Codex Fejérváry-Mayer, and the Codex Cospiano.

The interpretation of these manuscripts is far from easy. Only a few of them, such as the Borbonicus, are truly Aztec, while others, of the utmost importance, such as the Borgia, seem to emanate from the priestly colleges of the "Mexica-Puebla" area, between the central highlands and the Oaxaca Mountains.

Another kind of book, either pre-Cortesian or post-Cortesian, affords valuable material. Examples are such historical manuscripts as the Codex Telleriano-Remensis, the Azcatitlán, the Codex of 1576, and the Codex Mendoza; they describe the history of the Aztec tribe and state, and occasionally depict religious scenes and events.

Books written by Aztec chroniclers. These works were written in the Latin alphabet, either in Nahuatl or in Spanish, by learned natives, such as Tezozómoc, Ixtlilxóchitl, and Chimalpahin, who used ancient pictographic manuscripts as their basis. Some of these books, such as the *Anales de Cuauhtitlán* (in Nahuatl; "Annals of Cuauhtitlán"), or the *Historia de los Mexicanos por sus pinturas* (in Spanish; "History of the Mexicans through their Paintings"), are anonymous compilations.

Archaeological materials. These include statues of deities, religious bas-reliefs and mural paintings, clay statuettes and vases, and stone or wooden masks. Since most Aztec art is of a highly symbolic nature, such objects can convey important information. Their interpretation, of course, often raises delicate problems, which may be solved only through careful checking against written sources.

HISTORY

Early religious life in central Mexico. Early religious phenomena can only be deduced from archaeological remains. Numerous clay figurines found in tombs afford little evidence of religious beliefs during the agricultural Pre-Classic periods of Zacatenco and Ticomán (roughly 1500 to the 1st century BC). It is possible, however, that terra-cotta statuettes of women were meant to represent an agricultural deity, a goddess of the crops. Two-headed figurines found at Tlatilco, a site of the late Pre-Classic, may portray a supernatural being. Clay idols of a fire god in the form of an old man with an incense burner on his back date from the same period.

The first stone monument on the Mexican plateau is the pyramid of Cuicuilco, near Mexico City. In fact, it is rather a truncated cone, with a stone core; the rest is made of sun-dried brick with a stone facing. It shows the main features of the Mexican pyramids as they were developed in later times. It was doubtless a religious monument, crowned by a temple built on the terminal platform and surrounded with tombs. The building of such a structure obviously required a protracted and organized effort under the command of the priests.

The final phase of the Pre-Classic cultures of the central highland forms a transition from the village to the city, from rural to urban life. This was a far-reaching social and intellectual revolution, bringing about new religious ideas together with new art forms and theocratic regimes. It is significant that Olmec statuettes have been found at Tlatilco with late Pre-Classic material.

Olmec civilization

The Olmec civilization, whose heartland is generally held to have been in the Gulf of Mexico area in the states of Tabasco and Veracruz (La Venta, Tres Zapotes), may be considered, in the present condition of scholarly knowledge, as the first of the higher native cultures in Middle America. Its beginnings were contemporary with the late Pre-Classic, as early as the 3rd century BC or even the 9th century BC; its late phase extended down to the 5th century AD, contemporaneous with the Classic "golden age." Traces of the characteristic Olmec style are to be found throughout the Mexican territory from the Gulf Coast to the Oaxaca Valley and the Morelos and Guerrero areas.

The Olmecs, a native people whose origin and language remain unknown, developed the arts of architecture and sculpture. They built ceremonial centres, erected carved monoliths and altars, and devised a system of hieroglyphic writing. Their bas-reliefs and carved jadeites (sodium aluminum silicates, a form of jade) clearly show gods, priests and mythical scenes. A jaguar cult seems to have held a central place in the religious beliefs and practices, coupled with that of a peculiar "baby," an infant whose features (especially the mouth) suggest a tiger.

Classic religious life: Teotihuacán. Although jaguars occasionally appear in religious frescoes at Teotihuacán, the main Classic site of the central plateau, Olmec religion does not seem to have taken root in the highlands. Archaeological discoveries show that the influence of the large ceremonial centre of Teotihuacán was felt as far as the Mayan area of Guatemala. The Toltecs worshipped earth and water deities: the rain god, known in later times as Tlaloc in Nahuatl, the water goddess (Chalchiuhtlicue in Nahuatl); and the Feathered Snake (Quetzalcóatl in Nahuatl), a symbol of the fertility of the soil.

The Classic civilizations of the 1st millennium AD—Teotihuacán on the central mesa, El Tajín near the Gulf Coast, the Zapotec city of Monte Albán in Oaxaca, the Mayan cities of Chiapas, Yucatán, Guatemala, and Honduras—had many religious features in common, such as the rain god called Cocijo by the Zapotecs and Chac by the Maya.

Other Classic centres

At Palenque and other Mayan centres, a maize god was worshipped under the form of the Foliated Cross, a stylized cornstalk; there was also a sun cult, as evidenced by the carved reliefs of the Temple of the Sun at Palenque. Classic religion revolved mainly around deities linked with agriculture.

The Teotihuacán frescoes show that the benevolent rain god was believed to provide men, after death, with an eternal paradise, depicted as a tropical garden where they sang, danced, and played among flowers and fruit trees. There is no evidence there of human sacrifice. The priests are shown offering incense, rubber balls, and jade plates. Some features of ornamental art (*e.g.*, the scroll motif) also strongly suggest that the Teotihuacán theocracy was led by a priestly caste that may have come from the "hot lands," the rain-forest lowlands between the central plateau and the Gulf Coast. The bulk of the population, who worked on the land and provided the manpower for building monuments, may have belonged to agricultural tribes such as the Otomi, who in later times worshipped both an earth and moon goddess of fertility and a rain god.

Teotihuacán declined in the 8th century. External influences mark the so-called Mazapan transitional phase. A god, Xipe Totec, whose cult seems to have originated in the Yopi area of the Pacific slopes of Mexico, appears for the first time on the highland. Human victims were sacrificed to him.

Toltec religion. In the 9th century, a Nahuatl-speaking people from northern Mexico invaded the central area and founded a city called Tollan or Tula. Those warlike Toltecs worshipped astral deities such as Tezcatlipoca, the god of the night sky. The whole history of Tula until the 10th century is made up of the struggle between the old Teotihuacán religion symbolized by Quetzalcóatl (Feathered Snake) and the new one brought by the invaders. Vanquished by Tezcatlipoca's sorcery, the benevolent Feathered Snake fled from Tula. Thus central Mexico submitted to the cruel rites of the sky and war gods, who demanded the blood of human victims.

Rise of the Aztecs. The disintegration of the Toltec civilization after the fall of Tula in 1168 opened the central highland to the wandering Chichimec (barbarous) tribes of northern Mexico. Some cities such as Xochimilco, Colhuacán, Cholula remained as strongholds of the Toltec religious tradition. Others were founded by the newcomers like Azcapotzalco, Texcoco, Tlaxcala, and eventually Tenochtitlán (Mexico). After the 14th century, in spite of chronic warfare among 28 city-states, a community of culture developed through trading relations, a common language (Nahuatl), and intermarriage between the ruling families. Each city-state worshipped its particular tribal god: Quetzalcóatl (Cholula), Camaxtli (Tlaxcala), Huitzilopochtli (Mexico), but there was a general interchange of myths and rites.

Aztec synthesis of tribal cults

The Aztec religion, therefore, combined the astral cults of the northern tribes with those of the earth and rain gods of the sedentary peasants. Thus the pyramid of the main *teocalli* (temple) supported two sanctuaries of equal size, one devoted to Huitzilopochtli, god of the sun and of war, the other consecrated to Tlaloc, the ancient rain god. The priestly hierarchy, likewise, was headed by two high priests of equal rank, the Feathered Snakes: one of them bore the title of "priest of Tlaloc," the other that of "priest of Our Lord" (Huitzilopochtli).

The tendency of the Aztec mind to syncretism is illustrated by their pantheon, which at the time of the Spanish conquest included the Otomi fire god Otontecuhtli, the Huaxtec love goddess Tlazoltéotl, the Yopi god "Our Lord the Flayed One" (Xipe Totec), the earth goddess of the northern steppes Itzpapalotl, and the deity of anointments and medicine Tzapotlatenan from the Zapotec area. Some religious hymns were sung in foreign ("Chichimec") languages; rites connected with the Venus worship had been borrowed from the Mazatecs of Oaxaca.

Survivals after the Spanish Conquest. This traditional tendency to religious syncretism did not disappear altogether as a result of the Spanish Conquest and the mass conversion of the Indians to Christianity. Although virtually all the native people (with a few exceptions, such as the unevangelized Maya Lacandones of Chiapas) became devout Catholics, their conversion did not prevent them from retaining both magical and religious rites and beliefs down to this day. Most Indian groups still make offerings to the rain gods on mountaintops, "feed" the earth by sacrificing and burying animals, and cure sickness by magical means. The belief in the *nahual*, an individual totemic animal, is widespread. The *peyote* cult is still of the utmost importance among the Huichol of northwestern Mexico.

In many cases the ancient religion has combined with Christianity. Thus the pre-Spanish pilgrimage to the earth and moon goddess Tonantzin at Tepeyac has now taken the form of the cult devoted to Our Lady of Guadalupe, whom the Nahuatl-speaking Indians still call Tonantzin and whom the Otomi call by the name of the ancient moon deity. In most villages purely Indian brotherhoods

combine the cult of the Roman Catholic saints with pre-Cortesian ceremonies. In rural areas Mexican Catholicism is permeated by such survivals.

MYTHS

Cosmogony and eschatology. The Aztecs believed, as did other Middle American Indians, such as the Maya-Quiché, that four worlds had existed before the present universe. Those worlds or "Suns" had been destroyed by catastrophes. Mankind had been entirely wiped out at the end of each Sun. The present world was the fifth Sun.

The myth of the 'Suns'

The first Sun was called *nahui-ocelotl*, "Four-Jaguar," a date of the ritual calendar. Mankind was destroyed by jaguars. The animal was considered by the Aztecs as the *nahualli* (animal disguise) of Tezcatlipoca.

At the end of the Second Sun, *nahui-ehécatl*, "Four-Wind," a magical hurricane had transformed all men into monkeys. That disaster was caused by Quetzalcóatl in the form of Ehécatl, the wind god.

A rain of fire had put an end to the third Sun, *nahui-quiahuitl*, "Four-Rain." Tlaloc as the god of thunder and lightning presided over that period.

The fourth Sun, *nahui-atl*, "Four-Water," ended in a gigantic flood that lasted for 52 years. Only one man and one woman survived, sheltered in a huge cypress. But they were changed into dogs by Tezcatlipoca, whose orders they had disobeyed.

Present mankind has been created by Quetzalcóatl. The Feathered Snake with the help of his twin Xolotl, the dog-headed god, succeeded in reviving the dried bones of the old dead by sprinkling them with his own blood. The present Sun is called *nahui-ollin*, "Four-Earthquake," and is doomed to disappear in a tremendous earthquake. The skeleton-like monsters of the west, the *tzitzimime*, will then appear and kill all mankind.

Two deeply rooted concepts are revealed by these myths. One is the belief that the universe is unstable, that death and destruction continually threaten it. The other emphasizes the necessity of the sacrifice of the gods. Thanks to Quetzalcóatl's self-sacrifice, the ancient bones of Mictlan, "the Place of Death," have given birth to men. In the same way the sun and moon have been created: the gods, assembled in the darkness at Teotihuacán, built a huge fire; two of them, Nanahuatzin, a small deity covered with ulcers, and Tecciztécatl, a richly bejewelled god, threw themselves into the flames, from which the former emerged as the sun and the latter as the moon. Then the sun refused to move unless the other gods gave him their blood; they were compelled to sacrifice themselves to feed the sun.

Cosmology. According to the Aztec cosmological ideas, the universe has the general shape of a cross. To each of the four world directions are attached five of the 20 day-signs, one of them being a Year-Bearer (east, *acatl*, reed; west, *calli*, house; north, *tecpatl*, flint knife; south, *tochtli*, rabbit), a colour (east, red or green; west, white; north, black; south, blue), and certain gods. The fifth cardinal point, the centre, is attributed to the fire god Huehueteotl, because the hearth stands at the centre of the house.

Above the earth, which is surrounded by the "heavenly water" (*ilhuicáatl*) of the ocean, are 13 heavens, the uppermost of which, "where the air is delicate and frozen," is the abode of the Supreme Couple. Under the "divine earth," *teotlalli*, are the 9 Hells of Mictlan, with 9 rivers that the souls of the dead must cross. Thirteen was considered a favourable number, 9 extremely unlucky.

All the heavenly bodies and constellations were divinized, such as the Great Bear (Tezcatlipoca), Venus (Quetzalcóatl), the stars of the north (Centzon Mimixcoa, "the 400 Cloud-Serpents"), the stars of the south (Centzon Huitznáua, "the 400 Southerners"). The solar disk, Tonatiuh, was supposed to be borne on a litter from the east to the zenith, surrounded by the souls of dead warriors, and from the zenith to the west among a retinue of divinized women, the Cihuateteo. When the night began on the earth, day dawned in Mictlan, the abode of the dead.

Deities. The ancient tribes of central Mexico had worshipped fertility gods for many centuries when the Aztecs invaded the valley. The cult of these gods remained extremely important in Aztec religion. Tlaloc, the giver of rain but also the wrathful deity of lightning, was the leader of a group of rain gods, the Tlaloques, who dwelt on mountaintops. Chalchiuhtlicue (the One Who Wears a Jade Skirt) presided over fresh waters, Huixtocihuatl over salt waters and the sea. Numerous earth goddesses were associated with the fertility of the soil and with the fecundity of women, as Teteoinnan (Mother of the Gods), Coatlicue (the One Who Wears a Snake Skirt), Cihuacóatl (Snake-Woman), Itzpapalotl (Obsidian-Butterfly). Their significance is twofold: as fertility deities, they give birth to the young gods of maize, Centéotl, and of flowers, Xochipilli; as symbols of the earth that devour the bodies and drink the blood, they appear as warlike godheads. Tlazoltéotl, a Huaxtec goddess, presided over carnal love and over the confession of sins.

Vegetation and fertility gods

Xipe Totec, borrowed from the faraway Yopi tribe, was a god of the spring, of the renewal of vegetation, and at the same time the god of the corporation of goldsmiths. Human victims were killed and flayed to honour him.

The concept of a Supreme Couple played an important role in the religion of the old sedentary peoples such as the Otomi. Among the Aztecs it took the form of *intonan intota*, "Our Mother, our Father," the Earth and the Sun. But the Fire-god Huehueteotl was also associated with the Earth. In addition, Ometecuhtli (the Lord of the Duality) and Omecihuatl (the Lady of the Duality) were held to abide in the 13th heaven: they decided on which date a human being would be born, thus determining his or her destiny.

Among the fertility gods are to be counted the "400 Rabbits" (Centzon Totochtin), little gods of the crops, of *octli* (a fermented drink) and drunkenness, such as Ometochtli and Tepoztécatl.

The Aztecs brought with them the cult of their tribal sun and war god, Huitzilopochtli, "the Hummingbird of the Left," who was considered "the reincarnated Warrior of the South," the conquering sun of midday. According to a legend probably borrowed from the Toltecs, he was said to have been born near Tula. His mother, the earth goddess Coatlicue, had already given birth to the 400 Southerners and to the night goddess Coyolxauhqui, whom the newborn god exterminated with his *xiuhcoatl* (turquoise-snake).

Tezcatlipoca, god of the night sky, was the protector of the young warriors. Quetzalcóatl, the ancient Teotihuacán deity of vegetation and fertility, had been "astralized" and transformed into a god of the morning star. He was also revered as a wind god and as the ancient priest-king of the Toltec golden age: the discoveries of writing, the calendar, and the arts were attributed to him.

People of the Sun. The Aztecs thought of themselves as "the People of the Sun." Their duty was to wage cosmic war in order to provide the sun with his *tlaxcaltiliztli* (nourishment). Without it the sun would disappear from the heavens. Thus the welfare and the very survival of the universe depended upon the offerings of blood and hearts to the sun, a notion that the Aztecs extended to all the deities of their pantheon. Human sacrifice, as a result, became the most important feature of ritual.

Mythology of death and afterlife. The beliefs of the Aztecs concerning the other world and life after death show the same syncretism already observed. The old paradise of the rain god Tlaloc, depicted in the Teotihuacán frescoes, opened its gardens to those who died by drowning, lightning, or as a result of leprosy, dropsy, gout, or lung diseases. He was supposed to have caused their death and to have sent their souls to paradise.

Two categories of dead persons went up to the heavens as companions of the Sun: the Quauhteca (Eagle People), who comprised the warriors who died on the battlefield or on the sacrificial stone, and the merchants who were killed while travelling in faraway countries; and the women who died while giving birth to their first child and thus became Cihuateteo, "Divine Women."

Quauhteca and Cihuateteo

All the other dead went down to Mictlan, under the northern deserts, the abode of Mictlantecuhtli, the skeleton-masked god of death. There they travelled for four years until they arrived at the ninth Hell, where they disappeared altogether.

Offerings were made to the dead 80 days after the funeral, then one year, two, three and four years later. Then all link between the dead and the living had been severed. But the warriors who crossed the heavens in the retinue of the Sun came back to earth after four years as hummingbirds. The Cihuateteo, "Divine Women," used to appear at night at the crossroads and strike the passers-by with palsy.

World view. The world vision, or *Weltanschauung*, of the Aztecs conceded only a small part to man in the scheme of things. His destiny was submitted to the all-powerful *tonalpohualli* (the calendrical round); his life in the other world did not result from any moral judgment. His duty was to fight and die for the gods and for the preservation of the world order. Moreover, witchcraft, omens, portents dominated everyday life. That such a pessimistic outlook should have coexisted with the wonderful dynamism of Aztec civilization is in itself a remarkable achievement.

PRACTICES AND INSTITUTIONS

Aztec ritual calendar. *Tonalpohualli*, an Aztec term meaning "the count of days," is the name of the ritual calendar of 260 days. It runs parallel to the solar calendar of 365 days, which was divided into 18 months of 20 days and 5 supplementary unlucky days. The word *tonalli* means "day" and "destiny": the 260-day calendar was mainly used for purposes of divination. The days were named by the combination of 20 signs (natural phenomena such as *ehecatl*, wind, *ollin*, earthquake; animals like *tochtli*, rabbit, *ocelotl*, jaguar; plants as *acatl*, reed; objects like *tecpatl*, flint knife, *calli*, house) with the numbers 1 to 13. Thus the calendrical round included 20 series of 13 days.

Astrology and divination

Specialized priests called *tonalpouhque* interpreted the signs and numbers on such occasions as birth, marriage, departure of traders to faraway countries, and election of rulers. Each day and each 13-day series was deemed lucky, unlucky, or indifferent according to the deities presiding over it. Thus *ce-coatl* (one-snake) was held as favourable to the traders, *chicome-xochitl* (seven-flower) to the scribes and the weavers, *nahui-ehecatl* (four-wind) to the magicians. The men who were born during the *ce-ocelotl* (one-jaguar) series would die on the sacrificial stone, those whose birth took place on the day *ome-tochtli* (two-rabbit) would be drunkards, etc. The *tonalpohualli* dominated every aspect of public and private life. The same system can be found among all the Indian cultures of Middle America; its origin is at least as ancient as the Olmec civilization.

Temples and ceremonial centres. The ceremonial centres, such as the holy city built at the heart of Mexico, Tenochtitlán, consisted mainly of temples (*teocalli*), pyramids whose terminal platform supported the sanctuary proper. Some temples dedicated to the wind god were round. Other buildings, the *calmecac*, were used as the residence of the priests and as colleges of higher education. To the temple were annexed courts for the ceremonial ball game (*tlachtli*), sacrificial stones (*techcatl*), skull racks (*tzompantli*), and ritual bathhouses (*temazcalli*). All the districts (*calpulli*) of the city had their own temples, as well as the corporations of traders and craftsmen.

Human sacrifice. The outstanding feature of Mexican ritual from Toltec times was human sacrifice. The victims were either war prisoners or slaves bought for that purpose. In some cases the victim was chosen from among a certain category (women, young men). Death by sacrifice was considered a certain way of gaining a happy eternal life. It was therefore stoically accepted or even voluntarily sought. The victim wore the dress and ornaments of the god and was called *ixiptla*, the god's "image." The priests forced the victim backward on the sacrificial stone; one of them opened the victim's breast with a stroke of his flint knife and tore out the heart, which was burned in a stone urn (*cuauhxicalli*). In some ceremonies victims were decapitated, drowned, or burned. Part of their flesh was also ritually eaten.



Sacrificing a victim and offering his heart to the sun god. Florentine Codex, by Aztec Indian artists, c. 1550. In the American Museum of Natural History.
By courtesy of the American Museum of Natural History

At the end of each 52-year cycle, the ceremony called tying up the years was performed on the Huixachtécatl mountaintop. The priests lit the New Fire on the breast of a victim. The last renewal of the Fire took place in 1507.

Priesthood and rites. The clergy were extremely numerous. Most priests came from noble families, but sons and daughters of "plebeians" could accede to priesthood. Under the two high priests, the *Mexicatl teohuatzin* was at the head of all religious activities in the city and the provinces, with two assistants, one in charge of ritual, the other of education. The huge estates of the temples were administered by a general treasurer. Each deity had his own college of priests. Both male and female priests remained celibate.

A special category of priests, the *tonalpouhque*, interpreted the sacred books to predict the future. Those who consulted them used to pay them with clothes and food.

Feasts and festivals

As each month of 20 days was marked by a feast, sacrifices, pageants, flower offerings, and dances and songs dedicated to a particular god or group of gods, the activities of the clergy were manifold and unceasing. There were, in addition, ceremonies on certain days of the ritual calendar, and ceremonies performed by various corporations: traders, goldsmiths, feather workers, salt-makers, water bearers, healers, and midwives.

The clergy did not directly intervene in affairs of state but its influence was doubtless extremely strong. Priests of high rank were members of the electoral body that designated the rulers.

Sorcerers. The Aztecs believed in and greatly feared sorcery. The sorcerers (*nahualli*, a word meaning "disguise") were supposed to be endowed with the power of transforming themselves into animals, such as dogs and owls. They could cause sickness and death by burning a wooden figurine representing the intended victim of their witchcraft. Some sorcerers secretly opened the tomb of a recently buried woman who had died in childbirth (*i.e.*, of a "divinized woman") and severed her left forearm. Using the arm as a magical wand, they would put the members of a household to sleep and make off with their belongings.

Sorcerers, also called owl-men, prepared herb love potions as well as poisonous drinks.

Although witchcraft was punished by death, there is evidence that its practice was widespread and that sorcerers exacted gifts of considerable value from those who had them cast spells on their enemies.

Other practices. The physicians and midwives—both influential and esteemed professions—made wide use of religious rites and formulas. Several deities, it was believed, could cause illnesses or cure them. Skin diseases, ulcers, leprosy, dropsy, were attributed to Tlaloc, eye afflictions to Xipe Totec, and venereal diseases to Xochipilli. The healers addressed their prayers to these gods. Specialized deities were invoked, such as Ixtlilton for the sicknesses of children, the goddesses Quato and Caxoch, who cured headaches, the earth goddess who protected pregnant women and presided over the steam bath (*temazcalli*).

Rites like incense burning, rubbing with tobacco—considered a living being to which prayers were addressed—and ceremonial drinking of hallucinatory potions were performed by the physicians both in diagnosis and in treatment.

PLACE OF AZTEC RELIGION IN ANCIENT AMERICA

While the Aztec Empire brought to Mexico a new principle of political unity, religion remained a mixture of local beliefs and practices. The priesthood tried to introduce some order into the theological chaos of a religion that included many different traditions and cults.

The great sky god Tezcatlipoca, for instance, was said to assume several personalities and names. As the "black Tezcatlipoca," he remained the traditional northern deity of the night stars. As red, he was identified with Xipe Totec. As blue, he became Huitzilopochtli himself. Among the other gods, only Quetzalcóatl was considered Tezcatlipoca's brother and equal; all the other deities had been created by those two brother gods.

Nezahualcóyotl, the king of Texcoco (1431–72), had erected a temple in the form of a high tower without any statue or idol to the "unknown god, creator of all things," a faceless, mythless deity called Ipalnemohuani, "the One By Whom We Live." But it is unlikely that such a metaphysical concept could have spread among the population, which remained attached to local and traditional ideas and rites.

Aztec religion is only one form of Middle American native religion, or rather the form it assumed in later times. Even outside the Aztec area, among the Yucatec Maya for instance, many myths and practices, such as human sacrifice, closely parallel those of central Mexico.

Andean religion, on the contrary, as observed in the Inca Empire, shows many important differences. Even though some features may have been common to both the Andean and Mexican areas in a distant past (*e.g.*, the feline cult), separate developments had taken place for many centuries. It is significant that the ritual calendar so prevalent in Middle America, with its highly peculiar combination of 13 numbers and 20 day-signs, is not found anywhere in the Andes.

BIBLIOGRAPHY. C.A. BURLAND, *Magic Books from Mexico* (1953), graphic documents and interpretations; A. CASO, *The Religion of the Aztecs* (1937), a short but extremely accurate description by one of the most prominent Mexican scholars; HERNANDO CORTES, *Letters of Cortés*, trans. and ed. by F.A. MCNUTT, 2 vol. (1908), the famous reports of the conqueror to his sovereign; BERNAL DIAZ DEL CASTILLO, *The True History of the Conquest of New Spain*, trans. by A.P. MAUDSLAY, 5 vol. (1908–16), the testimony of an intelligent witness on native life in 1519–21; T.A. JOYCE, *Mexican Archaeology* (1914), a somewhat dated, but still valuable account of Aztec and Maya religion; J.E.S. LINNE, *Archaeological Researches at Teotihuacán, Mexico* (1934), the Teotihuacan civilization as seen through its ancient art and architecture; FR. TORIBIO BENAVENTE MOTOLINIA, *Memoriales* (1903), the report of a priest who was among the first missionaries in Mexico (in Spanish); FR. BERNARDINO DE SAHAGUN, *Historia general de las cosas de Nueva España*, 5 vol. (1938), the main source of our knowledge of Aztec culture (in Spanish), and *Florentine Codex: General History of the Things of New Spain*, trans. from the Aztec into English by A.J.O. ANDERSON and C.E. DIBBLE, Book 1, *The Gods* (1950); Book 2, *The Ceremonies* (1951); and Book 3, *The Origin of the Gods* (1952), a valuable document on the myths and descriptions of rites as dictated to Sahagún by his Aztec informants, accurately translated into English; EDUARD SELER, *Gesammelte Abhandlungen zur amerikanischen Sprach- und Alterthumskunde* (1903–04), a series of important papers based on a profound knowledge of Aztec language and religion (in German); JACQUES SOUSTELLE, *The Daily Life of the Aztecs* (1963), *Art of Ancient Mexico* (1967), *The Ancient Civilizations of Mexico* (1969); FR. JUAN DE TORQUEMADA, *Veynte y un libros rituales y Monarchia Indiana*, 3 pt. (1723), a generally reliable compilation based on early information (in Spanish).

(Ja.S.)

Ba'al Shem Ṭov

Israel ben Eliezer, called the Ba'al Shem Tov (Master of the Good Name) and also known by the acronym Besht, was the charismatic founder of Ḥasidism, a Jewish spiritual movement originated in the 18th century that demands the total hallowing of human existence. He was also responsible for divesting Kabbala (esoteric Jewish mysticism) of the rigid asceticism imposed on it by Isaac ben Solomon Luria in the 16th century. The Besht's life has been so adorned with fables and legends that a biography in the ordinary historical sense is not possible.

Early life

The Besht was born about 1700, probably at Tluste (Polish Tłuste, now Tolstoye), in southern Ukraine. Contemporary Jews called the village Okop or Akuf, depending on the Hebrew vocalization. As a young orphan he held various semi-menial posts connected with synagogues and Hebrew elementary religious schools. After marrying the daughter of the wealthy and learned Ephraim of Kuty, he retired to the Carpathian Mountains to engage in mystical speculation, meanwhile eking out his living as a lime digger. During this period his reputation as a healer, or *ba'al shem*, who worked wonders by means of herbs, talismans, and amulets inscribed with the divine name, began to spread. He later became an innkeeper and a ritual slaughterer and, about 1736, settled in the village of Medzhibozh, in Podolia. From this time until his death, he devoted himself almost entirely to spiritual pursuits.

Though the Besht gained no special renown as a scholar or preacher during his lifetime, he made a deep impression on his fellow Jews by going to the marketplace to converse with simple people and by dressing like them. Such conduct by a holy man was fiercely condemned in some quarters but enthusiastically applauded in others. The Besht defended his actions as a necessary "descent for the sake of ascent," a concept that eventually evolved into a socio-theological theory that placed great value on this type of spiritual ministration.

While still a young man, the Besht had become acquainted with such figures as Rabbi Naḥman of Gorodënka and Rabbi Naḥman of Kosov, already spoken of as creators of a new life, and with them he regularly celebrated the ritual of the three sabbath meals. In time it became customary for them to deliver pious homilies and discourses after the third meal, and the Besht took his turn along with the others. Many of these discourses were later recorded and have been preserved as the core of Ḥasidic literature. When the Besht's spiritual powers were put to a test by other members of the group—an indication that he probably was not yet recognized as the "first among equals"—he reportedly recognized a *mezuza* (ritual object affixed to a doorpost) as ritually "unfit" by means of his clairvoyant powers.

Renunciation of asceticism

The Besht gradually reached the point where he was prepared to renounce the strict asceticism of his companions. In words recorded by his grandson Rabbi Baruch of Medzhibozh, he announced:

> I came into this world to point a new way, to prevail upon men to live by the light of these three things: love of God, love of Israel, and love of Torah. And there is no need to perform mortifications of the flesh.

By renouncing mortification in favour of new rituals, the Besht in effect had taken the first step toward initiating a new religious movement within Judaism. The teaching of

the Besht centred on three main points: communion with God, the highest of all values; service in ordinary bodily existence, proclaiming that every human deed done "for the sake of heaven" (even stitching shoes and eating) was equal in value to observing formal commandments; and rescue of the "sparks" of divinity that, according to the Kabbala, were trapped in the material world. He believed that such sparks are related to the soul of every individual. It was the Besht's sensitivity to the spiritual needs of the unsophisticated and his assurance that redemption could be attained without retreat from the world that found a ready response among his listeners, the common Jewish folk. He declared that they were, one and all, "limbs of the divine presence."

The Besht and his followers were fiercely attacked by rabbinical leaders for "dancing, drinking, and making merry all their lives." They were called licentious, indifferent, and contemptuous of tradition—epithets and accusations that were wild exaggerations, to say the least.

Assessment

An understanding of the Besht's view of the coming of the Messiah depends to a great extent on the interpretation of a letter attributed to, but not signed by, the Besht. It affirms that the author made "the ascent of the soul," met the Messiah in heaven, and asked him when he would come. The answer he received was: "when your well-springs shall overflow far and wide"—meaning that the Besht had first to disseminate the teaching of Ḥasidism. According to one view, the story indicates that the messianic advent was central in the Besht's belief; according to another, it effectively removes messianic redemption from central spiritual concern in the life that must be lived here and now.

Among the Besht's most outstanding pupils was Rabbi Jacob Joseph of Polonnoye, whose books preserve many of the master's teachings. He speaks with holy awe of his religious teacher in tones that were echoed by other disciples, such as Dov Baer of Mezrechye, Rabbi Nahum of Chernobyl, Aryeh Leib of Polonnoye, and a second grandson, Rabbi Ephraim of Sydoluvka, who was but one of many to embellish the image of his grandfather with numerous legends.

During his lifetime, the Besht brought about a great social and religious upheaval and permanently altered many traditional values. In an atmosphere marked by joy, new rituals, and ecstasy, he created a new religious climate in small houses of prayer outside the synagogues. The changes that had occurred were further emphasized by the wearing of distinctive garb and the telling of stories. Though the Besht never did visit Israel and left no writings, by the time he died, in 1760, he had given to Judaism a new religious dimension in Ḥasidism that continues to flourish to this day. (R.S-U.)

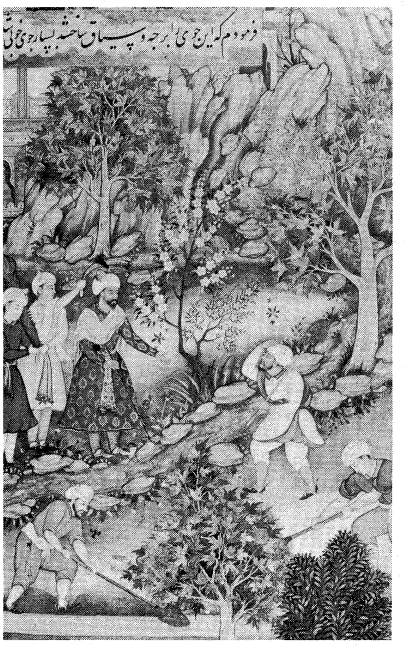
BIBLIOGRAPHY. DOV BAER, *Shivḥé ha-Besht* (1814), is the earliest collection of legends (in Hebrew) about the Ba'al Shem Ṭov. SAMUEL A. HORODEZKY, *Shivḥé ha-Besht* (1947), arranges systematically Dov Baer's desultory legends, with preface and commentary. DAN BEN-AMOS and JEROME R. MINTZ (eds. and trans.), *In Praise of Baal Shem Tov* (1970), offers an English translation of Dov Baer's legends, based upon the 1814 edition. MEYER LEVIN, *The Golden Mountain* (1932); and MARTIN BUBER, *Die Legende des Baalschem* (1932; Eng. trans., *The Legend of the Baal-Shem*, 1955), retell with literary flair the legends of the Ba'al Shem Ṭov. SALOMO BIRNBAUM, *Leben und Worte des Baalschem* (1920; Eng. trans., *The Life and Sayings of the Baal Shem*, 1933), contains excerpts from the writings and teachings of the Ba'al Shem Ṭov. See also the selected bibliography in DAN BEN-AMOS and JEROME R. MINTZ (*op. cit.*), pp. 273–279.

(S.Z.L.)

Bābur

Bābur (Ẓahīr-ud-Dīn Muḥammad), the first Mughal (or Mongol) emperor of India and founder of the Mughal dynasty there, was a many-sided man, a poet and diarist of genius, a soldier of distinction, a statesman, and an adventurer. He came from the Barlas tribe of Mongol origin to which the conqueror Timur belonged, but isolated members of the tribe had become Turks in language and manners through long residence in Turkish regions. Hence Bābur, though called a Mughal, drew most of his support from Turks, and the empire he founded was Turkish in character. His family had become members of the Chagatai clan, by which name they are known.

By courtesy of the trustees of the British Museum



Bābur, inspecting a garden, portrait miniature from the *Bābur-nameh*, 16th century. In the British Museum (MS. Or 3714).

Early years

Bābur (authorities disagree over whether the name means "lion," "tiger," or "panther") was born on February 15, 1483. He was fifth in direct male descent from Timur and 13th, through the female line, from Genghis Khan, the first of the great Mongol conquerors. Bābur's father, 'Umar Shaykh Mīrzā, ruled the small principality of Fergana to the north of the Hindu Kush (mountains). Because there was no fixed law of succession among the Turks, every prince of the Timurids—the dynasty founded by Timur—considered it his right to rule the whole of Timur's dominions. These territories were vast, and, hence, the princes' claims led to unending wars. The Timurid princes, moreover, considered themselves kings by profession, their business being to rule others without observing too precisely whether any particular region had actually formed a part of Timur's empire. Bābur's father, true to this tradition, spent his life trying to recover Timur's old capital of Samarkand, and Bābur followed in his footsteps. The qualities needed in this jungle of dynastic warfare were the ability to inspire loyalty and devotion, to manage the turbulent factions often rent by family feuds, and to draw revenue from the trading and agricultural classes. Bābur eventually mastered them all, but he was also a commander of genius.

For ten years (1494–1504) Bābur sought to recover Samarkand and twice occupied it briefly (1497 and 1501). But, in Muḥammad Shaybānī Khān, a descendant of Genghis Khan and ruler of the Uzbeks beyond the Syrdarya (Jaxartes) river, he had an opponent more powerful than even his close relatives. In 1501 Bābur was decisively defeated at Sar-e Pol and within three years had lost both Samarkand and his principality of Fergana. There was always hope at that time, however, for a prince of ability with engaging qualities and powers of leadership. In 1504 Bābur seized Kābul with his personal followers, maintaining himself there against all rebellions and intrigues. His last unsuccessful attempt on Samarkand (1511–12) induced him to give up a hopeless quest and to concentrate on expansion elsewhere. In 1522, when he was already turning his attention to Sind and India, he finally secured Qandahār, a strategic site on the road to Sind.

When Bābur made his first raid into India in 1519, the

Punjab was part of the dominions of Sultan Ibrāhīm Lodī of Delhi, but the governor, Dawlat Khān Lodī, resented Ibrāhīm's attempts to diminish his authority. By 1524 Bābur had invaded the Punjab three more times but was unable to master the tangled course of Punjab and Delhi politics sufficiently to achieve a firm foothold. Yet it was clear that the Delhi sultanate was rent with dissension and ripe for overthrow. After mounting a full-scale attack there, Bābur was recalled by an Uzbek attack on his Kābul kingdom, but a joint request for help from ʿĀlam Khān, Ibrāhīm's uncle, and Dawlat Khān encouraged Bābur to attempt his fifth, and first successful, raid.

Setting out in November 1525, Bābur met Ibrāhīm at Pānīpat, 50 miles (80 kilometres) north of Delhi, on April 21, 1526. Bābur's army was estimated at no more than 12,000, but they were seasoned followers, adept at cavalry tactics, and were aided by new artillery acquired from the Ottoman Turks. Ibrāhīm's army was said to number 100,000 with 100 elephants, but its tactics were antiquated, and it was rent with dissension. Bābur won the battle by coolness under fire, his use of artillery, and effective Turkish wheeling tactics on a divided, dispirited enemy. Ibrāhīm was killed. With his usual speed Bābur occupied Delhi three days later and reached Āgra on May 4. His first action there was to lay out a garden by the river Yamuna, now known as the Ram Bagh.

First victory in India

This brilliant success must have seemed at the time to be little more than one of his former forays on Samarkand. His small force, with the unaccustomed Indian hot weather upon them and 800 miles from their base at Kābul, was surrounded by powerful foes. All down the Ganges Valley were militant Afghan chiefs, in disarray at the moment but with a formidable military potential. To the south were the kingdoms of Mālwa and Gujarāt, both with large resources, while in Rājasthān, Rānā Sāngā of Mewār (Udaipur) was head of a powerful confederacy threatening the whole Muslim position in north India. Bābur's first problem was that his own followers, who, suffering from the heat and disheartened by the hostile surroundings, wished to return home as Timur had done. By employing threats, reproaches, promises, and appeals, vividly described in his memoirs, Bābur diverted them. He then dealt with Rānā Sāngā, who, when he found that Bābur was not retiring as his Turkish ancestor had done, advanced to the attack with an estimated 100,000 horses and 500 elephants. With most of the neighbouring strongholds still held by his foes, Bābur was virtually surrounded. He sought divine favour by abjuring liquor, breaking the wine vessels, and pouring the wine down a well. His followers responded both to this act and his stirring exhortations and stood their ground at Khānua, 37 miles west of Āgra, on March 16, 1527. Bābur used his customary tactics—a barrier of wagons for his centre, with gaps for the artillery and for cavalry sallies, and wheeling cavalry charges on the wings. The artillery stampeded the elephants, and the flank charges bewildered the Rājputs, who, after ten hours, broke, never to rally under a single leader again.

Bābur had now to deal with the defiant Afghans to the east, who had captured Lucknow while he was facing Rānā Sāngā. Other Afghans had rallied to Sultan Ibrāhīm's brother Maḥmūd Lodī, who had occupied Bihār. There were also Rājput chiefs still defying him, principally the ruler of Chanderi. After capturing that fortress in January 1528, Bābur turned to the east. Crossing the Ganges, he drove the Afghan captor of Lucknow into Bengal. He then turned on Maḥmūd Lodī, whose army was scattered in Bābur's third great victory of the Ghāghara, where that river joins the Ganges, on May 6, 1529. Artillery was again decisive, helped by the skillful handling of boats. Bābur's dominions were now secure from Qandahār to the borders of Bengal, with a southern limit marked by the Rājput desert and the forts of Ranthambhor, Gwalior, and Chanderi. Within this great area, however, there was no settled administration, only a congeries of quarrelling chiefs. An empire had been gained but had still to be pacified and organized. It was thus a precarious heritage that Bābur passed on to his son Humāyūn.

Consolidation of his conquests

In 1530, when Humāyūn became so ill that his life was despaired of, Bābur is said to have offered his life to God in exchange for Humāyūn's, walking seven times round the bed to complete the vow. Humāyūn recovered, and, from that time, Bābur declined, to die in Āgra on December 26, 1530.

Bābur is rightly considered the founder of the Indian Mughal Empire, even though the work of consolidating the empire was performed by his grandson Akbar. Bābur, moreover, provided the glamour of magnetic leadership that inspired the next two generations.

Assessment

Bābur was a military adventurer of genius, an empire builder of good fortune, and an engaging personality. He was also a Turki poet of considerable gifts that would have won him distinction apart from his political career. He was a lover of nature who constructed gardens wherever he went and complemented beautiful spots by holding convivial parties. Finally, his prose memoirs, the *Bābur-nāmeh*, have become a world classic of autobiography. They were translated from Turki into Persian in Akbar's reign (1589) and into English (1921–22). They portray a ruler unusually magnanimous for his age, cultured, witty, convivial, and full of good fellowship and adventurous spirit, with a sensitive eye for natural beauty.

BIBLIOGRAPHY. Bābur's own autobiography, the *Bābur-nāmeh* in the original Turki, is available in English as the *Memoirs of Bābur*, 2 vol. (1921–22). L.F. RUSHBROOK WILLIAMS, *An Empire-Builder of the Sixteenth Century* (1918), provides a good introduction both to Bābur and to the contemporary Indian situation. J.F. GRENARD, *Baber, fondateur de l'empire des Indes* (1930; Eng. trans., *Baber, First of the Moguls*, 1931), is a good interpretative study. The standard work in English is WILLIAM ERSKINE, *A History of India Under The Two First Sovereigns of the House of Taimur, Báber and Humáyun*, 2 vol. (1854). A condensed but valuable study is by E. DENISON ROSS in *The Cambridge History of India*, vol. 4, ch. 1 (1963). For further information, see *The Encyclopaedia of Islam*, new ed., vol. 1 (1960).

(T.G.P.S.)

Babylon

Babylon (Bab-ilu) was one of the most famous cities of antiquity, the capital of southern Mesopotamia (Babylonia) from the early 2nd until the late 1st millennium BC, and of the Neo-Babylonian Empire in the 7th and 6th centuries BC. Situated on the Euphrates River about 55 miles (88 kilometres) south of Baghdad, the city exists only as extensive ruins near the modern town of al-Ḥillah, Iraq.

History. Though traces of prehistoric settlement exist, Babylon's development as a major city was late by Mesopotamian standards, no mention of it occurring before the 23rd century BC. After the fall of the 3rd dynasty of Ur, under which Babylon had been a provincial centre, it became the nucleus of a small kingdom established in 1894 BC by the Amorite king Sumuabum, whose successors consolidated its status. The sixth and best known of the Amorite dynasts, Hammurabi (1792–1750 BC), conquered the surrounding city-states and raised Babylon to the capital of a kingdom comprising all southern Mesopotamia and part of Assyria (northern Iraq). Its political importance, together with its favourable geographical position, made it henceforth the main commercial and administrative centre of Babylonia, while its wealth and prestige made it a target for foreign conquerors.

Old Babylonian kingdom

After a Hittite raid in 1595 BC, the city passed to the control of the Kassites (*c.* 1570), who established a dynasty lasting over four centuries. Later in this period, Babylon became a literary and religious centre, the prestige of which was reflected in the elevation of Marduk, its chief god, to supremacy in Mesopotamia. In 1234 Tukulti-Ninurta I of Assyria took Babylon, though subsequently the Kassite dynasty reasserted itself until 1158, when the city was sacked by the Elamites. Babylon's acknowledged political supremacy is shown by the fact that the dynasty of Nebuchadrezzar I (1124–1103) made it their capital, although they did not originate there. This dynasty endured for over a century.

Just before 1000, pressure from Aramaean immigrants

from northern Syria brought administrative dislocation inside Babylon. From this period to the fall of Assyria in the late 7th century BC, there was a continual struggle between Aramaean or associated Chaldean tribesmen and the Assyrians for political control of the city. Its citizens claimed privileges, such as exemption from forced labour, certain taxes, and imprisonment, which the Assyrians, with a similar background, were usually readier to recognize than were immigrant tribesmen. Furthermore, the citizens, grown wealthy by commerce, benefitted by an imperial power able to protect international trade, but suffered economically from disruptive tribesmen. Such circumstances made Babylon usually prefer Assyrian to Aramaean or Chaldean rule.

From the 9th to the late 7th centuries Babylon was almost continuously under Assyrian suzerainty, usually wielded through native kings, though sometimes Assyrian kings ruled in person. Close Assyrian involvement in Babylon began with Tiglath-pileser III (744–727 BC) as a result of Chaldean tribesmen pressing into city territories, several times usurping the kingship. Disorders accompanying increasing tribal occupation finally persuaded Sennacherib (704–681 BC) that peaceful control of Babylon was impossible, and in 689 he ordered destruction of the city. Esarhaddon (680–669 BC) rescinded Sennacherib's policy, and, after expelling the tribesmen and returning the property of the Babylonians to them, undertook the rebuilding of the city; but the image of Marduk, removed by Sennacherib, was retained in Assyria throughout his reign, probably to prevent any potential usurper from using it to claim the kingship. In the mid-7th century, civil war broke out between the Assyrian king Ashurbanipal and his brother who ruled in Babylonia as subking. Ashurbanipal laid siege to the city, which fell to him in 648 after famine had driven the defenders to cannibalism.

Neo-Babylonian Empire

After Ashurbanipal's death, a Chaldean leader, Nabopolassar, in 626 made Babylon the capital of a kingdom that under his son Nebuchadrezzar II became a major imperial power. Nebuchadrezzar undertook a vast program of rebuilding and fortification in Babylon, labour gangs from many lands increasing the mixture of the population. Nebuchadrezzar's most important successor, Nabonidus, campaigned in Arabia for a decade, leaving his son Belshazzar as regent in Babylon. Nabonidus failed to protect property rights or religious traditions of the capital, and attempted building operations elsewhere to rival Marduk's great temple of Esagila. When the Persians under Cyrus attacked in 539 BC, the capital fell almost without resistance; a legend (accepted by some as historical) that Cyrus achieved entry by diverting the Euphrates is unconfirmed in contemporary sources.

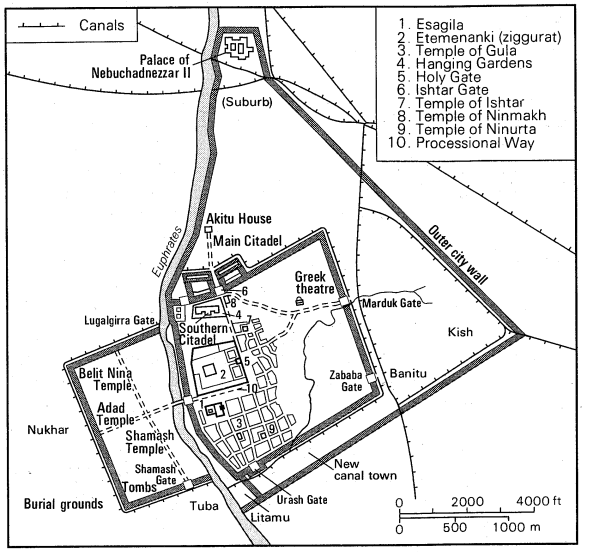
Under the Persians, Babylon retained most of its institutions, became capital of the richest satrapy in the empire, and, according to Herodotus, the world's most splendid city. A revolt against Xerxes I (482) led to destruction of its fortifications and temples, and the melting down of the golden image of Marduk.

Conquest by Alexander the Great

In 331 Babylon surrendered to Alexander the Great, who confirmed its privileges and ordered the restoration of the temples. Alexander, recognizing the commercial importance of the city, allowed its satrap to issue coinage and began construction of a harbour to foster trade. In 323 Alexander died in the palace of Nebuchadrezzar; he had planned to make Babylon his imperial capital. Alexander's conquest brought Babylon into the orbit of Greek culture, and Hellenistic science was greatly enriched by the contributions of Babylonian astronomy. After a power struggle among Alexander's generals, Babylon passed to the Seleucid dynasty in 312. The city's importance was much reduced by the building of a new capital, Seleucia, on the Tigris, to which part of Babylon's population was transferred in 275.

The ancient city. Evidence of the topography of ancient Babylon is provided by excavations, cuneiform texts, and descriptions by the 5th-century Greek historian Herodotus and other classical authors. The extensive rebuilding by Nebuchadrezzar has left relatively little archaeological data in the central area earlier than his time, while elsewhere the water table has limited excavation in early strata. The reports of Herodotus largely relate to the Babylon built by Nebuchadrezzar.

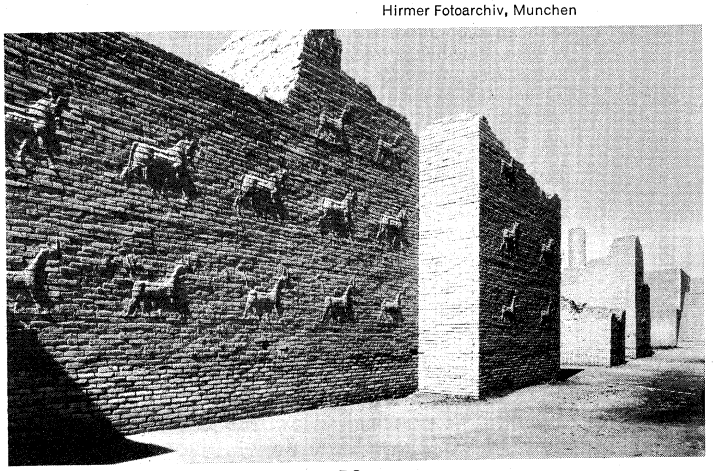
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*, Georg Westermann Verlag, Braunschweig



Babylon during the reign of Nebuchadrezzar II.

Nebuchadrezzar's Babylon was the largest city in the world, covering 2,500 acres (10,000 hectares). The Euphrates, which has since shifted its course, flowed through it, the older part of the city being on the east bank. There the central feature was Esagila, the great temple of Marduk, with its associated ziggurat (a tower built in several stages) Etemenanki. The latter, popularly known as the Tower of Babel, had a base 100 yards on a side, and its seven stages, the uppermost a temple in blue glaze, reached to a height of 300 feet (91 metres). Four other temples in the eastern half of the city are known from excavations and a larger number from texts. Along the Euphrates, particularly in the neighbourhood of Esagila, were quays for trading vessels, and textual evidence that Babylon was an entrepôt for trade with south Babylonia points to the existence of warehouses. The river was spanned by a bridge on brick piles, with stone capping, to the western half of the city. The streets were laid out on a grid, with the main axis parallel to the river. From Esagila northward passed the paved Processional Way, its walls decorated with enamelled lions. Passing through the Ishtar Gate, adorned with enamelled bulls and dragons, it led to the Akitu House, a small temple outside the city, visited by Marduk at the New Year festival. West of the Ishtar Gate, one of eight fortified gates, were two palace complexes that covered about 40 acres with their fortifications.

Hirmer Fotoarchiv, Munchen



The Ishtar Gate, Babylon, 6th century BC.

East of the Processional Way lay an area that since the time of Hammurabi had contained private dwellings built around central courtyards. A powerful double wall, reinforced by a fosse (ditch), enclosed the city on both sides of the Euphrates. Beyond the city walls to the east an outer rampart of triple construction, 11 miles long, met the Euphrates south and north of the city, at its northern junction enclosing another palace. Between the inner and outer defenses was irrigated land with a network of canals, some going back to the time of Hammurabi. Greek tradition refers to the Hanging Gardens, a simulated hill of vegetation-clad terracing over a vaulted substructure that in Hellenistic times was deemed one of the Seven Wonders of the World. The German archaeologist R. Koldewey identified the base of this early in the 20th century AD with a part of the palace complex, though the tradition could have arisen from the existence of trees on the ziggurat.

Hanging Gardens

The present site. The present site, an extensive field of ruins, contains several prominent mounds. The main mounds are (1) Babil, the remains of Nebuchadrezzar's palace in the northern corner of the outer rampart; (2) Qasr, comprising the palace complex (with a building added in Persian times), the Ishtar Gate, and the Emakh temple; (3) Amran ibn Ali, the ruins of Esagila; (4) Merkez, marking the ancient residential area east of Esagila; (5) Humra, containing rubble removed by Alexander from the ziggurat in preparation for rebuilding, and a theatre he built with material from the ziggurat; and (6) Ishin Aswad, where there are two further temples. A depression called Sahn marks the former site of the ziggurat Etemenanki. An over-life-size basalt lion, probably of Hittite origin and brought to Babylon in antiquity, stands north of the Ishtar Gate.

Archaeology. After minor surveys and excavations by the British scholar C.J. Rich (1811 and 1817), and the English diplomat A.H. Layard (1850), the French orientalist F. Fresnel, the German assyriologist J. Oppert (1852–54), and others, a major archaeological operation began under R. Koldewey of the Deutsche Orient-Gesellschaft in 1899, continuing unbroken until 1917. In the course of his excavation of the structures mentioned, Koldewey also discovered cuneiform inscriptions, statues, steles (pillars), terra-cotta reliefs, cylinder seals, pottery, glassware, and jewellery. Further brief investigations were made by the Deutsches Archäologisches Institut in 1956 under H.J. Lenzen at the Greek theatre, and in 1966 under H.J. Schmidt at the site of Etemenanki. Restoration of the Emakh temple, and of part of the Ishtar Gate, the Processional Way, and the palace complex was begun in 1958 by the Iraq Department of Antiquities, which has also built a half-size model of the complete Ishtar Gate at the entrance to the site.

BIBLIOGRAPHY. The principal excavator's own account of his results is given in a good popular form by R. KOLDEWEY, *Das wieder erstehende Babylon*, 4th ed. (1925; Eng. trans., *The Excavations at Babylon*, 1914). H.W.F. SAGGS, "Babylon," in D. WINTON THOMAS (ed.), *Archaeology and Old Testament Study*, pp. 39–56 (1967), gives a summary of medieval and early modern investigations of the site of Babylon, together with an account of the results of the excavations there, and related evidence, oriented to its biblical interest. H.W.F. SAGGS, *Everyday Life in Babylonia and Assyria* (1965), popular account of particular aspects of Babylonian life, pp. 156–180 dealing with the Babylon of Nebuchadrezzar.

(H.W.F.S.)

Bach, Johann Sebastian

Although he was admired by his contemporaries primarily as an outstanding harpsichordist, organist, and expert on organ building, Johann Sebastian Bach is now generally regarded as one of the greatest composers of all time and is celebrated as the creator of the *Brandenburg Concertos*, *The Well-Tempered Clavier*, the *Mass in B Minor*, and numerous other masterpieces of church and instrumental music. Appearing at a propitious moment in the history of music, Bach was able to survey and bring together the principal styles, forms, and national traditions that had developed during preceding generations and, by virtue of his synthesis, enrich them all.



Bach, lithograph by Rudolph Hoffmann.
By courtesy of the Osterreichische Nationalbibliothek, Vienna

He was a member of a remarkable family of musicians who were proud of their achievements, and about 1735 he drafted a genealogy, *Ursprung der musicalisch-Bachischen Familie* ("Origin of the Musical Bach Family"), in which he traced his ancestry back to his great-great-grandfather Veit Bach, a Lutheran baker (or miller), who was driven from Hungary to Wechmar in Thuringia, a historic region of Germany, by religious persecution late in the 16th century and died in 1619. There were Bachs in the area before that, and it may be that, when Veit moved to Wechmar, he was returning to his birthplace. He used to take his cittern to the mill and play it while grinding was going on. Johann Sebastian remarked, "A pretty noise they must have made together! However, he learnt to keep time, and this apparently was the beginning of music in our family."

Until the birth of Johann Sebastian, his was the least distinguished branch of the family; its members had been competent practical musicians, but not composers such as Johann Christoph, Johann Michael, and Johann Ludwig. In later days the most important musicians in the family were Johann Sebastian's sons, Wilhelm Friedemann, Carl Philipp Emanuel, and Johann Christian (the "English Bach").

Early years. J.S. Bach was born at Eisenach, Thuringia (now in East Germany), on March 21, 1685, the youngest child of Johann Ambrosius Bach and Elisabeth Lämmerhirt. Ambrosius was a string player, employed by the town council and the ducal court of Eisenach. Johann Sebastian started school in 1692 or 1693 and did well in spite of frequent absences. Of his musical education at this time, nothing definite is known; but he may have picked up the rudiments of string playing from his father, and no doubt he attended the Georgenkirche, where Johann Christoph Bach was organist until 1703.

By 1695 both his parents were dead, and he was looked after by his eldest brother, also named Johann Christoph (1671–1721), organist at Ohrdruf. This Christoph had been a pupil of the influential keyboard composer Johann Pachelbel, and he apparently gave Johann Sebastian his first formal keyboard lessons. The young Bach again did well at school, until in 1700 his voice secured him a place in a select choir of poor boys at the school at the Michaelskirche, Lüneburg (now in West Germany).

First keyboard lessons

His voice must have broken soon after this, but he remained at Lüneburg for a time, making himself generally useful. No doubt he studied in the school library, which had a large and up-to-date collection of church music; he probably heard Georg Böhm, organist of the Johanneskirche; and he visited Hamburg to hear the renowned organist and composer Johann Adam Reinken at the Katharinenkirche, contriving also to hear the French orchestra maintained by the Duke of Celle.

He seems to have returned to Thuringia in the late summer of 1702, for, at some time between July and No-

vember, he applied for the post of organist at Sangerhausen (in modern East Germany). But for ducal interference, he would have obtained it; and, when every allowance is made for the low technical standards of that area, this means that he must already have been a reasonably proficient organist. His experience at Lüneburg, if not at Ohrdruf, had turned him away from the secular string-playing tradition of his immediate ancestors; thenceforth, he was, chiefly, though not exclusively, a composer and performer of keyboard and sacred music. The next few months are wrapped in mystery, but, by March 4, 1703, he was a member of the orchestra employed by Johann Ernst, Herzog von Weimar (brother of Wilhelm Ernst, whose service Bach entered in 1708). This post was a mere stopgap; he probably already had his eye on the organ then being built at the Neukirche in Arnstadt; for, when it was finished, he helped to test it, and in August 1703 he was appointed organist—all this at the age of 18. Arnstadt documents imply that he had been court organist at Weimar; this is incredible, though it is likely enough that he had occasionally played there.

The Arnstadt period. At Arnstadt, on the northern edge of the Thuringian forest, where he remained until 1707, Bach devoted himself to keyboard music, the organ in particular. While at Lüneburg, he had apparently had no opportunity of becoming directly acquainted with the spectacular, flamboyant playing and compositions of Dietrich Buxtehude, the most significant exponent of the north German school of organ music. In October 1705 he repaired this gap in his knowledge by obtaining a month's leave and walking to Lübeck (over 200 miles [300 kilometres]). His visit must have been profitable, for he did not return until about the middle of January 1706. In February his employers complained about his absence and about other things as well: he had harmonized the hymn tunes so freely that the congregation could not sing to his accompaniment, and, above all, he had produced no cantatas. Perhaps the real reasons for his neglect were that he was temporarily obsessed with the organ and was on bad terms with the local singers and instrumentalists, who were not under his control and did not come up to his standards. In the summer of 1705 he had made some offensive remark about a bassoon player, which led to an unseemly scuffle in the street. His replies to these complaints were neither satisfactory nor even accommodating; and the fact that he was not dismissed out of hand suggests that his employers were as well aware of his exceptional ability as he was himself and were reluctant to lose him.

Acquaintance with Buxtehude's music

During these early years, Bach inherited the musical culture of the Thuringian area, a thorough familiarity with the traditional forms and hymns (chorales) of the orthodox Lutheran service, and, in keyboard music, perhaps (through his brother, Johann Christoph) a bias toward the formalistic styles of the south. But he also learned eagerly from the northern rhapsodists, Buxtehude above all. By 1708 he had probably learned all that his German predecessors could teach him and arrived at a first synthesis of northern and southern German styles. He had also studied, on his own and during his presumed excursions to Celle, some French organ and instrumental music.

Among the few works that can be ascribed to these early years with anything more than a show of plausibility are the *Capriccio sopra la lontananza del suo fratello dilettissimo* (*Capriccio on the Departure of His Most Beloved Brother*, 1704, BWV 992); the chorale prelude on *Wie schön leuchtet* (*How Brightly Shines*, *c.* 1705, BWV 739); the fragmentary early version of the organ *Prelude and Fugue in G Minor* (before 1707, BWV 535a). (The "BWV" numbers provided are the standard catalog numbers of Bach's works as established in the *Bach-Werke-Verzeichnis*, prepared by the German musicologist Wolfgang Schmieder.)

The Mühlhausen period. In June 1707 Bach obtained a post at the Blasiuskirche in Mühlhausen in Thuringia. He moved there soon after and married his cousin Maria Barbara Bach at Dornheim on October 17. At Mühlhausen things seem, for a time, to have gone more smoothly. He produced several church cantatas at this time; all of these works are cast in a conservative mold, based on biblical and chorale texts and displaying no influence of the "modern" Italian operatic forms that were to appear in Bach's later cantatas. The famous organ *Toccata and Fugue in D Minor* (BWV 565), written in the rhapsodic northern style, and the *Prelude and Fugue in D Major* (BWV 532) may also have been composed during the Mühlhausen period, as well as the organ *Passacaglia in C Minor* (BWV 582), an early example of Bach's instinct for large-scale organization. Cantata No. 71, *Gott ist mein König* (*God is my King*), of February 4, 1708, was printed at the expense of the city council and was the first of Bach's compositions to be published. While at Mühlhausen, Bach copied music to enlarge the choir library, tried to encourage music in the surrounding villages, and was in sufficient favour to be able to interest his employers in a scheme for rebuilding the organ (February 1708). His real reason for resigning on June 25, 1708, is not known. He himself said that his plans for a "well-regulated [concerted] church music" had been hindered by conditions in Mühlhausen and that his salary was inadequate. It is generally supposed that he had become involved in a theological controversy between his own pastor Frohne and Archdeacon Eilmar of the Marienkirche. Certainly, he was friendly with Eilmar, who provided him with librettos and became godfather to Bach's first child; and it is likely enough that he was not in sympathy with Frohne, who, as a Pietist, would have frowned on elaborate church music. It is just as possible, however, that it was the dismal state of musical life in Mühlhausen that prompted Bach to seek employment elsewhere. At all events, his resignation was accepted, and shortly afterward he moved to Weimar, some miles west of Jena on the Ilm River. He continued, nevertheless, to be on good terms with Mühlhausen personalities, for he supervised the rebuilding of the organ, is supposed to have inaugurated it on October 31, 1709, and composed a cantata for February 4, 1709, which was printed but has disappeared.

Cantatas of the Mülhausen period

The Weimar period. Bach was, from the outset, court organist at Weimar and a member of the orchestra. Encouraged by Wilhelm Ernst, he concentrated on the organ during the first few years of his tenure. From Weimar, Bach occasionally visited Weissenfels; in February 1713 he took part in a court celebration there that included a performance of his first secular cantata, *Was mir behagt*, or the *Hunt Cantata* (BWV 208).

Late in 1713 Bach had the opportunity of succeeding Friedrich Wilhelm Zachow at the Liebfrauenkirche, Halle; but the Herzog raised his salary, and he stayed on at Weimar, becoming concertmaster on March 2, 1714, with the duty of composing a cantata every month. He became friendly with a relative, Johann Gottfried Walther, a music lexicographer and composer who was organist of the town church, and, like Walther, Bach took part in the musical activities at the Gelbes Schloss (Yellow Castle), then occupied by Herzog Wilhelm's two nephews, Ernst August and Johann Ernst, both of whom he taught. The latter was a talented composer who wrote concertos in the Italian manner, some of which Bach arranged for keyboard instruments; the boy died in 1715, in his 19th year.

Unfortunately, Bach's development cannot be traced in detail during the vital years 1708–14, when his style underwent a profound change. There are too few datable works. From the series of cantatas written in 1714–16, however, it is obvious that he had been decisively influenced by the new styles and forms of the contemporary Italian opera and by the innovations of such Italian concerto composers as Antonio Vivaldi. The results of this encounter can be seen in such cantatas as numbers 182, 199, and 61 in 1714; 31 and 161 in 1715; and 70 and 147 in 1716. His favourite forms appropriated from the Italians were those based on refrain (ritornello) or da capo schemes in which wholesale repetition—literal or with modifications—of entire sections of a piece permitted him to create coherent musical forms with much larger dimensions than had hitherto been possible. These newly acquired techniques henceforth gov-

Italian influence on Bach

erned a host of Bach's arias and concerto movements, as well as many of his larger fugues (especially the mature ones for organ) and profoundly affected his treatment of chorales.

Among other works almost certainly composed at Weimar are most of the *Orgelbüchlein* (*Little Organ Book*); all but the last of the so-called 18 "Great" chorale preludes; the earliest organ trios; and most of the organ preludes and fugues. The "Great" *Prelude and Fugue in G Major* for organ (BWV 541) was finally revised about 1715, and the *Toccata and Fugue in F Major* (BWV 540) may have been played at Weissenfels.

On December 1, 1716, Johann Samuel Drese, musical director at Weimar, died. He was then succeeded by his son, who was rather a nonentity. Bach presumably resented being thus passed over; and in due course he accepted an appointment as musical director to Prince Leopold of Köthen, which was confirmed in August 1717. Herzog Wilhelm, however, refused to accept his resignation—partly, perhaps, because of Bach's friendship with the Herzog's nephews, with whom the Herzog was on the worst of terms. About September a contest between Bach and the famous French organist Louis Marchand was arranged at Dresden. The exact circumstances are not known; but Marchand avoided the contest by leaving Dresden a few hours before it should have taken place. By implication, Bach won. Perhaps this emboldened him to renew his request for permission to leave Weimar; at all events he did so but in such terms that the duke imprisoned him for a month (November 6–December 2). A few days after his release, Bach moved to Köthen (in modern East Germany), some 30 miles (50 kilometres) north of Halle.

The Köthen period. There, as musical director, he was concerned chiefly with chamber and orchestral music. Even though some of the works may have been composed earlier and revised later, it was at Köthen that the sonatas for violin and clavier, viola da gamba and clavier, and the works for unaccompanied violin and cello were put into something like their present form. The *Brandenburg Concertos* were finished by March 24, 1721; in the sixth concerto—so it has been suggested—Bach bore in mind the technical limitations of the Prince, who played the gamba. Bach played the viola by choice; he liked to be "in the middle of the harmony." He also wrote a few cantatas for the Prince's birthday and other such occasions; most of these seem to have survived only in later versions, adapted to more generally useful words. And he found time to compile pedagogical keyboard works: the *Clavierbüchlein* for W.F. Bach (begun January 22, 1720); some of the *French Suites;* the *Inventions* (1720); and the first book (1722) of *Das Wohltemperierte Klavier* (*The Well-Tempered Clavier*, eventually consisting of two books, each of 24 preludes and fugues in all keys and known as the Forty-eight). This remarkable collection systematically explores both the potentials of a newly established tuning procedure, which, for the first time in the history of keyboard music, made all the keys equally usable, and the possibilities for musical organization afforded by the system of "functional tonality," a kind of musical syntax consolidated in the music of the Italian concerto composers of the preceding generation and a system that was to prevail for the next 200 years. At the same time *The Well-Tempered Clavier* is a compendium of the most popular forms and styles of the era: dance types, arias, motets, concertos, etc., presented within the unified aspect of a single compositional technique: the rigorously logical and venerable fugue.

Influence of The *Well-Tempered Clavier* on keyboard music

Maria Barbara Bach died unexpectedly and was buried on July 7, 1720. About November, Bach visited Hamburg; his wife's death may have unsettled him and led him to inquire after a vacant post at the Jacobikirche. Nothing came of this, but he played at the Katharinenkirche in the presence of Reinken. After hearing Bach improvise variations on a chorale tune, the old man said, "I thought this art was dead; but I see it still lives in you."

On December 3, 1721, Bach married Anna Magdalena Wilcken, daughter of a trumpeter at Weissenfels. Apart from his first wife's death, these first four years at Köthen were probably the happiest of Bach's life. He was on the best terms with the Prince, who was genuinely musical; and in 1730 Bach said that he had expected to end his days there. But the Prince married on December 11, 1721, and conditions deteriorated. The Princess—described by Bach as "an *amusa*" (that is to say, opposed to the muses)—required so much of her husband's attention that Bach began to feel neglected. He also had to think of the education of his elder sons, born in 1710 and 1714, and he probably began to think of moving to Leipzig as soon as the cantorate fell vacant with the death of Johann Kuhnau on June 5, 1722. Bach applied in December, but the post—already turned down by Bach's friend, Georg Philipp Telemann—was offered to another prominent composer of the day, Christoph Graupner, the musical director at Darmstadt. As the latter was not sure that he would be able to accept, Bach gave a trial performance (Cantata No. 22, *Jesu nahm zu sich die Zwölfe* [*Jesus called unto Him the Twelve*]) on February 7, 1723; and, when Graupner withdrew (April 9), Bach was so deeply committed to Leipzig that, although the Princess had died on April 4, he applied for permission to leave Köthen. This he obtained on April 13, and on May 13 he was sworn in at Leipzig.

He was appointed honorary musical director to Köthen, and both he and Anna were employed there from time to time until the Prince died, on November 19, 1728.

Years at Leipzig. As Director of Church Music for the city of Leipzig, Bach had to supply performers for four churches. At the Peterskirche the choir merely led the hymns. At the Neukirche, Nikolaikirche, and Thomaskirche, part singing was required; but Bach himself conducted, and his own church music was performed, only at the last two. His first official performance was on May 30, 1723, the first Sunday after Trinity Sunday, with Cantata No. 75, *Die Elenden sollen essen.* New works produced during this year include many cantatas and the *Magnificat* in its first version. The first half of 1724 saw the production of the *St. John Passion,* which was subsequently revised. The total number of cantatas produced during this ecclesiastical year was about 62, of which about 39 were new works.

Production of the *St. John Passion*

On June 11, 1724, the first Sunday after Trinity, Bach began a fresh annual cycle of cantatas, and within the year he wrote 52 of the so-called chorale cantatas, formerly supposed to have been composed over the nine-year period 1735–44. The Sanctus of the *Mass in B Minor* was produced at Christmas.

During his first two or three years at Leipzig, Bach had produced a large number of new cantatas, sometimes, as recent research has revealed, at the rate of one a week. This phenomenal pace raises the question of Bach's approach to composition. Bach and his contemporaries, subject to the hectic pace of production, had to invent or discover their ideas quickly and could not rely on the unpredictable arrival of "inspiration." Nor did the musical conventions and techniques or the generally rationalistic outlook of the time necessitate this reliance, as long as the composer was willing to accept them. The Baroque composer who submitted to the regimen inevitably had to be a traditionalist who willingly embraced the conventions.

Symbolism. A repertory of melody types existed, for example, that was generated by an explicit "doctrine of figures" that created musical equivalents for the figures of speech in the art of rhetoric. Closely related to these "figures" are such examples of pictorial symbolism in which the composer writes, say, a rising scale to match words that speak of rising from the dead or a descending chromatic scale (depicting a howl of pain) to sorrowful words. Pictorial symbolism of this kind occurs only in connection with words—in vocal music and in chorale preludes, where the words of the chorale are in the listener's mind. There is no point in looking for resurrection motifs in *The Well-Tempered Clavier.* Pictorialism, even when not codified into a doctrine, seems to be a fundamental musical instinct and essentially an expressive device. It can, however, become more abstract, as in the case of number symbolism, a phenomenon observed

Use of pictorial symbolism

too often in the works of Bach to be dismissed out of hand. Number symbolism is sometimes pictorial; in the *St. Matthew Passion* it is reasonable that the question "Lord, is it I?" should be asked 11 times, once by each of the faithful disciples. But the deliberate search for such symbolism in Bach's music can be taken too far. Almost any number may be called "symbolic" (3, 6, 7, 10, 11, 12, 14, and 41 are only a few examples); any multiple of such a number is itself symbolic; and the number of sharps in a key signature, notes in a melody, measures in a piece, and so on may all be considered significant. As a result, it is easy to find symbolic numbers anywhere, but ridiculous to suppose that such discoveries invariably have a meaning.

Besides the melody types, the Baroque composer also had at his disposal similar stereotypes regarding the further elaboration of these themes into complete compositions, so that the arias and choruses of a cantata almost seem to have been spun out "automatically." One is reminded of Bach's delightfully innocent remark "I have had to work hard; anyone who works just as hard will get just as far," with its implication that everything in the "craft" of music is teachable and learnable. The fact that no other composer of the period, with the arguable exception of Handel, even remotely approached Bach's achievement indicates clearly enough that the application of the "mechanical" procedures was not literally "automatic" but was controlled throughout by something else —artistic discrimination, or taste. "Taste," a most respected attribute in the culture of the 18th century, is an utterly individual compound of raw talent, imagination, psychological disposition, judgment, skill, and experience. It is unteachable and unlearnable.

As a result of his intense activity in cantata production during his first three years in Leipzig, Bach had created a supply of church music to meet his future needs for the regular Sunday and feast-day services. After 1726, therefore, he turned his attention to other projects. He did, however, produce the *St. Matthew Passion* in 1729, a work that inaugurated a renewed interest in the mid-1730s for vocal works on a larger scale than the cantata: the now-lost *St. Mark Passion* (1731), the *Christmas Oratorio*, BWV 248 (1734), and the *Ascension Oratorio* (Cantata No. 11, *Lobet Gott in seinen Reichen;* 1735).

St. Matthew Passion and *Christmas Oratorio*

Nonmusical duties. In addition to his responsibilities as musical director, Bach also had various nonmusical duties in his capacity as the cantor of the school at the Thomaskirche. Since he resented these latter obligations, Bach frequently absented himself without leave, playing or examining organs, taking his son Friedemann to hear the "pretty tunes," as he called them, at the Dresden opera and fulfilling the duties of the honorary court posts that he contrived to hold all his life. To some extent, no doubt, he accepted engagements because he needed money; he complained in 1730 that his income was less than he had been led to expect (he remarked that there were not enough funerals); but, obviously, his routine work must have suffered. Friction between Bach and his employers thus developed almost at once. On the one hand, Bach's initial understanding of the fees and prerogatives accruing to his position—particularly regarding his responsibility for musical activities in the University of Leipzig's Paulinerkirche—differed from that of the town council and the university organist, Johann Gottlieb Görner. On the other hand, Bach remained, in the eyes of his employers, their third (and unenthusiastic) choice for the post, behind Telemann and Graupner. Furthermore, the authorities insisted on admitting unmusical boys to the school, thus making it difficult for Bach to keep his churches supplied with competent singers; they also refused to spend enough money to keep a decent orchestra together. The resulting ill feeling had become serious by 1730. It was temporarily dispelled by the tact of the new rector, Johann Matthias Gesner, who admired Bach and had known him at Weimar; but Gesner stayed only until 1734 and was succeeded by Johann August Ernesti, a young man with up-to-date ideas on education, one of which was that music was not one of the humanities but a time-wasting sideline. Trouble flared up again in July 1736; it then took the form of a dispute over Bach's right to appoint prefects and became a public scandal. Fortunately for Bach, he became court composer to the Elector of Saxony in November 1736. As such, after some delay, he was able to induce his friends at court to hold an official inquiry, and his dispute with Ernesti was settled in 1738. The exact terms of the settlement are not known; but, thereafter, Bach did as he liked.

Instrumental works. In 1726, after he had completed the bulk of his cantata production, Bach began to publish the clavier *Partitas* singly, with a collected edition in 1731, perhaps with the intention of attracting recognition beyond Leipzig and thus securing a more amenable appointment elsewhere. The second part of the *Clavierübung*, containing the *Concerto in the Italian Style* and the *French Overture (Partita) in B Minor* appeared in 1735. The third part, consisting of the *Organ Mass* with the *Prelude and Fugue* ["St. Anne"] *in E Flat Major* (BWV 552), appeared in 1739. From *c.* 1729–36 Bach was honorary musical director to Weissenfels; and, from 1729 to 1737 and again from 1739 for a year or two, he directed the Leipzig Collegium Musicum. For these concerts he adapted some of his earlier concertos as harpsichord concertos, thus becoming one of the first composers in history—if not the very first—of concertos for keyboard instrument and orchestra, just as he was one of the first to use the harpsichordist's right hand as a true melodic part in chamber music. These are just two of several respects in which the basically conservative and traditional Bach, as is becoming increasingly recognized, was a significant innovator as well.

Innovations as a keyboard composer

About 1733 Bach began to produce cantatas in honour of the Elector of Saxony and his family, evidently with a view to the court appointment he secured in 1736; many of these secular movements were adapted to sacred words and re-used in the *Christmas Oratorio*. The Kyrie and Gloria of the *Mass in B Minor*, written in 1733, were also dedicated to the Elector, but the rest of the *Mass* was not put together until Bach's last years. On his visits to Dresden, Bach had won the regard of Graf Hermann Karl von Keyserlingk, the Russian envoy, who commissioned the so-called *Goldberg Variations;* these were published as part four of the *Clavierübung* about 1742, and Book Two of the "Forty-eight" seems to have been compiled about the same time. In addition, he wrote a few cantatas, revised some of his Weimar organ works, and published the so-called *Schübler Chorale Preludes* in or after 1746.

Last years. In May 1747 he visited his son Emanuel at Potsdam and played before Frederick II the Great of Prussia; in July his improvisations, on a theme proposed by the King, took shape as *The Musical Offering*. In June 1747 he joined a Sozietät der Musikalischen Wissenschaften (Society of the Musical Sciences) that had been founded by his former pupil Lorenz Christoph Mizler; he presented the canonic variations on the chorale *Vom Himmel hoch da komm' ich her* (*From Heaven Above to Earth I Come*) to the society, in manuscript, and afterward published them.

Of Bach's last illness little is known, except that it lasted several months and prevented him from finishing *The Art of the Fugue*. His constitution was undermined by two unsuccessful eye operations performed by John Taylor, the itinerant English quack who numbered Handel among his other failures; and he died on July 28, 1750, at Leipzig. His employers proceeded with relief to appoint a successor; Burgomaster Stieglitz remarked, "The school needs a cantor, not a musical director—though certainly he ought to understand music." Anna Magdalena was left badly off. For some reason, her stepsons did nothing to help her, and her own sons were too young to do so. She died on February 27, 1760, and was given a pauper's funeral.

Unfinished as it was, *The Art of the Fugue* was published in 1751. It attracted little attention and was reissued in 1752 with a laudatory preface by Friedrich Wilhelm Marpurg, a well-known Berlin musician, who later became director of the royal lottery. In spite of

Publication of *The Art of the Fugue*

Marpurg and of some appreciative remarks by Johann Mattheson, the influential Hamburg critic and composer, only about 30 copies had been sold by 1756, when Emanuel Bach offered the plates for sale. As far as is known, they were sold for scrap.

Emanuel Bach and the organist-composer Johann Friedrich Agricola (a pupil of Sebastian's) wrote an obituary; Mizler added a few closing words and published the result in the journal of his society (1754). There is an English translation of it in *The Bach Reader*. Though incomplete and inaccurate, the obituary is of very great importance as a firsthand source of information.

Bach appears to have been a good husband and father. Indeed, he was the father of 20 children, only ten of whom survived to maturity. There is amusing evidence of a certain thriftiness, a necessary virtue; for he was never more than moderately well off, and he delighted in hospitality. Living as he did at a time when music was beginning to be regarded as no occupation for a gentleman, he occasionally had to stand up for his rights both as a man and as a musician; he was then obstinate in the extreme. But no sympathetic employer had any trouble with Bach, and with his professional brethren he was modest and friendly. He was also a good teacher and from his Mühlhausen days onward was never without pupils.

REPUTATION AND INFLUENCE

For about 50 years after Bach's death, his music was neglected. This was only natural; in the days of Haydn and Mozart, no one could be expected to take much interest in a composer who had been considered old-fashioned even in his lifetime—especially since his music was not readily available, and half of it (the church cantatas) was fast becoming useless as a result of changes in religious thought.

At the same time, musicians of the late 18th century were neither so ignorant of Bach's music nor so insensitive to its influence as some modern authors have suggested. Emanuel Bach's debt to his father was considerable and Bach exercised a profound and acknowledged influence directly on Haydn, Mozart, and Beethoven.

After 1800 the revival of Bach's music gained momentum. The German writer Johann Nikolaus Forkel published a *Life, Genius and Works* in 1802 and acted as adviser to the publishers Hoffmeister and Kühnel, whose collected edition, begun in 1801, was cut short by the activities of Napoleon. By 1829 a representative selection of keyboard music was nonetheless available, although very few of the vocal works were published. But in that year the German musician Eduard Devrient and the German composer Felix Mendelssohn took the next step with the centenary performance of the *St. Matthew Passion*. It and the *St. John Passion* were both published in 1830; the *Mass in B Minor* followed (1832–45). The Leipzig publisher Peters began a collected edition of "piano" and instrumental works in 1837; the organ works followed in 1844–52.

Publication of the *Mass in B Minor*

Encouraged by Robert Schumann, the Bach-Gesellschaft (BG) was founded in the centenary year 1850, with the purpose of publishing the complete works. By 1900 all the known works had been printed, and the BG was succeeded by the Neue Bach-Gesellschaft (NBG), which exists still, organizing festivals and publishing popular editions. Its chief publication is its research journal, the *Bach-Jahrbuch* (from 1904). By 1950 the deficiencies of the BG edition had become painfully obvious, and the Bach-Institut was founded with headquarters at Göttingen (West Germany) and Leipzig, to produce a new standard edition (the *Neue Bach-Ausgabe* or NBA) expected to comprise 84 volumes.

In retrospect, the Bach revival, reaching back to 1800, can be recognized as the first conspicuous example of the deliberate exhumation of old music, accompanied by biographical and critical studies; and it served as an inspiration and a model for subsequent work of that kind.

Among the biographical and critical works on Bach, the most important was the monumental study *Johann Sebastian Bach* (2 vols., Leipzig, 1873–80), by the German musicologist Philipp Spitta, covering not only Bach's life and works but also a good deal of the historical background. Although wrong in many details, the book is still indispensable to the Bach student.

Editions of Bach's works. The word *Urtext* (original text) may lead the uninitiated to suppose that they are being offered an exact reproduction of what Bach wrote. It must be understood that the autographs of many important works no longer exist. Therefore, Bach's intentions often have to be pieced together from anything up to 20 sources, all different. Even first editions and facsimiles of autograph manuscripts are not infallible guides to Bach's intentions. In fact, they are often dangerously misleading, and practical musicians should take expert advice before consulting them. Editions published between 1752 and *c.* 1840 are little more than curiosities, chiefly interesting for the light they throw on the progress of the revival.

No comprehensive edition is trustworthy throughout: neither Peters nor the BG nor even the NBA. Nevertheless, it is advisable to begin by finding out whether the music desired has been published by the NBA.

MAJOR WORKS

Vocal music (sacred)

MASSES: *Mass in B Minor*, BWV 232 (1724–46); 4 Lutheran masses (*i.e.*, containing only settings of the Kyrie and the Gloria).

ORATORIOS: *Christmas Oratorio*, BWV 248 (1734); *Easter Oratorio* (*Kommt, eilet und laufet*, BWV 249; 1725); *Ascension Oratorio* (1735).

PASSIONS: *Passion According to St. John*, BWV 245 (1724); *Passion According to St. Matthew*, BWV 244 (1729).

CANTATAS: About 200 for different Sundays in the church year (1707–after 1735; mainly 1714–16, 1723–27), mostly for soloist(s), chorus, and orchestra.

OTHER WORKS: *Magnificat in D Major*, BWV 243; 7 motets; 2 Sanctus settings (3 others based on works by other composers); 186 independent chorale harmonizations.

Vocal music (secular)

CANTATAS: 24, mostly for soloists, chorus, and orchestra—all on German texts, except two Italian. They include the *Coffee Cantata* (*Schweigt stille, plaudert nicht*, BWV 211; *c.* 1732) and the *Peasant Cantata* (*Mer hahn en neue Oberkeet*, BWV 212; 1742).

OTHER WORKS: 5 songs for voice and continuo and 1 quodlibet for four voices and continuo.

Orchestral music

CONCERTOS: 6 *Brandenburg Concertos* (pre-1721); 2 concertos for violin and orchestra and one for two violins (1717–23); 7 for one harpsichord, 3 for two harpsichords, 2 for three and 1 for four harpsichords; 1 concerto for harpsichord, flute, and violin.

OTHER ORCHESTRAL WORKS: 4 overtures (suites); *Sinfonia in D Major* (incomplete).

Chamber music

SONATAS: 2 for violin and continuo; 2 for flute and continuo; 1 for two flutes and harpsichord; 2 for flute, violin and continuo; 3 for harpsichord and flute; 3 for harpsichord and viola da gamba; 6 for harpsichord and violin.

OTHER CHAMBER MUSIC: *Das musikalisches Opfer* (1747) for strings, flute, and continuo; 6 unaccompanied sonatas (partitas) for violin (*c.* 1720); 6 unaccompanied suites (sonatas) for cello (*c.* 1720).

Organ music

CHORALE PRELUDES: 140 chorale preludes including the *Orgelbüchlein* (mainly 1714–16); *Clavierübung*, vol. 3 (1739), and *Schübler Chorale Preludes* (1746 or later).

FUGUES: 18 preludes and fugues (1708–17, 1729–39), including the "St. Anne" in E flat major and the "Wedge" in E minor; 5 toccatas and fugues (1700–17), including the "Dorian" in D minor; 3 fantasies and fugues; 4 other fugues.

OTHER ORGAN COMPOSITIONS: Variations on the chorale *Vom Himmel hoch* (1747); *Passacaglia in C Minor*, BWV 582 (1708–17); 4 concertos; 7 fantasies; 4 preludes; 6 sonatas (trios); 3 trios.

Harpsichord music

COLLECTIONS: *Clavierübung*: vol. 1 (1726–31), 6 partitas; vol. 2 (1735), *French Overture in B Minor* and *Concerto in the Italian Style;* vol. 3 (1739) is organ music with 4 "duets" for harpsichord; and vol. 4 (1742), *Goldberg Variations; The Well-Tempered Clavier*, 2 vol. (1722 and 1742), containing 48 preludes and fugues, one in each key in each book; *Clavierbüchlein* (1720), for Wilhelm Friedemann Bach, con-

taining 15 two-part and 15 three-part inventions, 20 preludes, 2 chorale preludes, 2 allemandes, 4 minuets, a fugue, and an "applicatio"; *Clavierbüchlein* (1722) and *Notenbuch* (1725), both for Anna Magdalena Bach, containing marches, minuets, a musette, polonaises, etc.; 6 *French Suites* and 6 *English Suites.*

OTHER HARPSICHORD WORKS: *Aria variata* in A minor; 2 capriccios; *Chromatic Fantasy and Fugue;* 5 fantasies, 2 with fugues; 12 *Little Preludes;* 4 preludes and 6 for beginners; 4 preludes and fughettas, 3 preludes and fugues; 2 sonatas; 4 miscellaneous suites; 7 toccatas and arrangements.

For unspecified instrument(s)

Die Kunst der Fuge (1749); 16 fugues and 4 canons.

BIBLIOGRAPHY

Catalogs: WOLFGANG SCHMIEDER, *Thematisch-systematisches Verzeichnis der musikalischen Werke von Johann Sebastian Bach. Bach-Werke-Verzeichnis* (BWV; 1950), the standard catalog of Bach's music, including a comprehensive bibliography (up to 1950) for each work; PAUL KAST, *Die Bach-Handschriften der Berliner Staatsbibliothek* (1958), a descriptive catalog of the Bach manuscripts in the possession of the Deutsche Staatsbibliothek, Berlin, the largest single repository with over 75 percent of the surviving Bach sources.

Collections of correspondence, sketchbooks, and reminiscences: Bach-Dokumente I: Schriftstücke von der Hand Johann Sebastian Bachs (1963), a critical edition of all surviving nonmusical documents, such as letters and receipts, in Bach's hand; and *Bach-Dokumente II: Fremdschriftliche und gedruckte Dokumente zur Lebensgeschichte Johann Sebastian Bachs 1685–1750* (1969), a critical edition of all known printed and handwritten discussions of and references to Bach dating from his lifetime—both volumes of the *Dokumente* are edited by WERNER NEUMANN and HANS-JOACHIM SCHULZE and are published as supplements to the *Neue Bach-Ausgabe;* HANS DAVID and ARTHUR MENDEL (eds.), *The Bach Reader: A Life of Johann Sebastian Bach in Letters and Documents,* rev. ed. (1966); ROBERT L. MARSHALL, *The Compositional Process of J.S. Bach,* 2 vol. (1972), with transcriptions of all surviving musical sketches and drafts included in vol. 2.

Biography and criticism: PHILIPP SPITTA, *Johann Sebastian Bach,* 2 vol. (1873–80; Eng. trans., *Johann Sebastian Bach: His Work and Influence on the Music of Germany, 1685–1750,* 1883–85, reprinted 1951)—Spitta's monumental study is still the standard biography, although no longer valid in many particulars. Further important full-length studies are: ALBERT SCHWEITZER, *J.S. Bach,* 2 vol. (1905; Eng. trans., 1911, reprinted 1966), an influential, if highly subjective and personal interpretation; CHARLES SANFORD TERRY, *Bach: A Biography* (1928), a useful supplement (based on new archival researches) to the biographical portions of Spitta's work; KARL GEIRINGER, *Johann Sebastian Bach: The Culmination of an Era* (1966), the only full-length account of the life and works to make use of the far-reaching results of research in the 1950s by Alfred Dürr and Georg von Dadelsen bearing on the chronology of Bach's works.

On the vocal music: ALFRED DURR, *Die Kantaten von Johann Sebastian Bach,* 2 vol. (1971), a general survey plus individual essays on each cantata by one of the principal editors of the *Neue Bach-Ausgabe;* WERNER NEUMANN, *Handbuch der Kantaten Joh. Seb. Bachs,* 3rd ed. (1966), a handbook of useful factual data and schematic analyses of all the cantatas; *Johann Sebastian Bach. Sämtliche Kantatentexte* (1956), a complete critical edition of the cantata texts; W.G. WHITTAKER, *The Cantatas of Johann Sebastian Bach, Sacred and Secular,* 2 vol. (1959), a stimulating appreciation, but needs to be used with caution.

On the instrumental music: HERMANN KELLER, *Die Orgelwerke Bachs* (1948; Eng. trans., *The Organ Works of Bach,* 1967) and *Die Klavierwerke Bachs* (1950), the historical context of Bach's organ and keyboard works, and individual analyses of the compositions; D.F. TOVEY, *A Companion to "The Art of Fugue"* (1931), an analysis; HANS T. DAVID, *J.S. Bach's Musical Offering: History, Interpretation, and Analysis* (1945).

On performance: ERWIN BODKY, *The Interpretation of Bach's Keyboard Works* (1960), a controversial but stimulating approach; WALTER EMERY, *Bach's Ornaments* (1953), a discussion of the problems and suggested solutions; ARTHUR MENDEL, the preface of his edition of the vocal score of *The Passion According to St. John* (1951).

(W.Em/Ro.Ma.)

Bacon, Francis

Francis Bacon, lawyer, courtier, statesman, philosopher, and master of the English tongue, is remembered popularly for the sharp worldly wisdom of a few dozen essays; by historians for his power as a speaker in Parliament and in some famous trials and as James I's lord chancellor; and more critically, as a man who claimed all knowledge as his province and, after a magisterial survey, proceeded urgently to advocate new ways by which men might establish a legitimate command over nature to the glory of God and the relief of man's estate.

By courtesy of the National Portrait Gallery, London



Francis Bacon, oil painting by an unknown artist. In the National Portrait Gallery, London.

LIFE

Youth and early maturity. Bacon was born January 22, 1561, at York House off the Strand, London, the younger of the two sons of the lord keeper, Sir Nicholas Bacon, by his second marriage. Through his mother, Ann Cooke, a woman famed for her erudition and Protestant zeal, he was related to William Cecil, Lord Burghley, Elizabeth I's principal minister, and was closely connected later with his cousin Robert Cecil, earl of Salisbury, who succeeded William as chief minister of the crown and retained that post under James I. Francis and his brother Anthony were sent to study under John Whitgift—the vice chancellor of Cambridge and later archbishop of Canterbury—at Trinity College in 1573. Their two years in Cambridge, however, were broken by ill health; neither had a strong constitution. Bacon's distaste for what he termed "unfruitful" Aristotelian philosophy began at this time. In 1576 Francis went to Paris with Sir Amias Paulet, the English ambassador to France, and in his household acquired some knowledge of the French court. He was recalled abruptly after the sudden death of his father in 1579 and found himself with a poorer patrimony than he might have expected. From this time until well into the next reign, Bacon was haunted by financial stress; in 1598 he was arrested for debt, although mainly he kept afloat by borrowing on security, a practice that became habitual and dangerous.

Bacon's family

Early legal career and political ambitions. Bacon turned first toward a legal career. In 1576, shortly before he left for Paris, Francis and his brother had been admitted as "ancients" (senior governors) of Gray's Inn, one of the four inns of court that served as institutions for legal education, in London. In 1579 he took up residence at the Inn and in 1582 became a barrister and progressed in time through the stages of reader (lecturer at the Inn), bencher (senior member of the Inn), queen's, and then king's counsel extraordinary to the solicitor general and the attorney general. These positions left their mark in a number of professional publications. The staple of his career was legal, but the law did not satisfy his political and philosophical ambitions.

He occupied himself with the tract "Temporis Partus Maximus" ("The Greatest Part of Time") in 1582; it has not survived. In 1584 he sat as member of Parliament for Melcombe Regis in Dorset and subsequently represented Taunton, Liverpool, the County of Middlesex, Southampton, Ipswich, and the University of Cambridge. In 1589 a "Letter of Advice" to the Queen and *An Advertisement touching the Controversies of the Church of England* indicated his political interests and showed a fair promise of political potential by reason of his level-headedness and disposition to reconcile. Already in 1585 the sequence of Bacon's fine letters applying for support and preferment had begun, with an approach to Sir Francis Walsingham, the secretary of state, followed by a direct, manly letter to his kinsman Burghley. A grant of the reversion (*i.e.*, the right of succession to the office) of the clerkship to the Star Chamber—a law court that handled certain cases reserved from the ordinary courts to the judgment of the king or queen—was cold comfort and brought no advantage until 1608. In 1593 came a bad setback to his political hopes: he took a stand objecting to the government's intensified demand for subsidies to help meet the expenses of the war against Spain. Elizabeth took offense, and Bacon was in disgrace over several critical years when there were chances for legal advancement. He had refused to apologize, and the penalty was heavy.

Letters in search of support and preferment

Relationship with Essex. Meanwhile, sometime before July 1591, both Francis and his brother, Anthony, had become acquainted with Robert Devereux, the young earl of Essex, Burghley's ward, and a prime favourite of the Queen. Anthony's devotion to Essex was simple and disinterested; Francis saw in the Earl the "fittest instrument to do good to the State" and acted accordingly. He offered Essex the friendly advice of an older, wiser, and more subtle man and provided him with speeches for masques—*i.e.*, short dramatic presentations performed by masked actors at social gatherings. Essex did his best to mollify the Queen, and when the office of attorney general fell vacant he enthusiastically supported the claim of Bacon. His advocacy failed, however, and in 1594 the attorney's position went to Sir Edward Coke, a great English lawyer and Bacon's constant rival, who had been supported by Burghley. In the following year Essex and Burghley joined in recommending Bacon for the post of solicitor general; but Elizabeth refused to appoint him, although she did make him one of her learned counsels. Essex sought to make up for these disappointments by pressing upon Bacon a gift of land at Twickenham. Francis, in his letter of acceptance, employing a metaphor from land tenure, made a significant proviso, explaining that his lordship might have only as much as might lawfully be enclosed from ground that he reckoned to be for common use. Essex recommended him again—and again without success—for the post of master of the rolls, the vice chancellor of the chancery court. It seems that this all too insistent advocacy was more of a hindrance than a help.

The dubious and traitorous actions of Essex

Bacon had begun also to feel doubts about his patron's courses of action. By 1598 Essex's failure in the islands voyage, in which he attempted to intercept Spanish treasure ships returning from the New World at the Azores, made him harder to control; and Bacon's efforts to divert his energies to Ireland—where the Irish Catholics were rebelling against the English troops stationed there to help enforce the establishment of the Elizabethan church—were only too successful. In a later account Bacon averred that what he gave Essex was a warning about the desperate Irish situation; from a private letter, however, it appears that he had also argued encouragingly for this venture. When it went wrong and Essex lost his head and returned against orders, Bacon certainly did what he could to sort things out and steady tempers down. He merely managed to offend both sides, and in June 1600 he found himself as the Queen's learned counsel taking part in the informal trial of his patron. He made his position clear in a letter of July 20: his priority was to be *bonus civis* ("good citizen") and next *bonus vir* ("good man"). Essex bore him no ill will and shortly after his release was again on friendly terms with him. In Essex's desperate project of seizing the Queen, forcing her to dismiss his rivals, and rousing the city of London in his support, Bacon had neither share nor knowledge. After Essex's abortive rebellion of February 8, 1601, Bacon viewed Essex as a traitor and drew up the official report that was heavily corrected on submission to royal authority and published as *A Declaration of the Practices and Treasons attempted and committed by Robert, late Earle of Essex.*

There is no means of knowing what these moves may have cost Bacon in terms of ill feeling. Nevertheless, in 1604 he published the *Apologie in certaine imputations concerning the late Earle of Essex* in defense of his action; it is a coherent piece of self-justification over the changed relationship between the two men, but to the imaginations of posterity it does not carry complete conviction. The assessment made by James Spedding, the 19th-century editor of Bacon's works, however, finds no fault in any part of Bacon's conduct toward Essex. It is perhaps the lack of evidence of personal distress that has left so sour a taste in many mouths.

Career in the service of James I. When Elizabeth died in 1603, Bacon's letter-writing ability was directed to finding a place for himself and a use for his talents in James I's services. He pointed to his concern for Irish affairs, the union of the kingdoms, and the pacification of the church as proof that he had much to offer the new king.

Through his cousin Robert Cecil, Bacon was one of the 300 new knights dubbed in 1603. The following year he was confirmed as learned counsel and sat in the first Parliament of the new reign in the debates of its first session. He was also active as one of the commissioners for discussing a union with Scotland. In the autumn of 1605 he published his *Advancement of Learning*, dedicated to the King, and in the following summer he married Alice Barnham, the daughter of a London alderman. Preferment in the royal service, however, still eluded him, and it was not until June 1607 that his petitions and his vigorous though vain efforts to persuade the Commons to accept the King's proposals for union with Scotland were at length rewarded with the post of solicitor general. Even then, his political influence remained negligible, a fact that he came to attribute to the power and jealousy of Cecil, then earl of Salisbury and the King's chief minister.

Political advancement

After Salisbury's death in 1612, Bacon renewed his efforts to gain influence with the King, writing a number of remarkable papers of advice upon affairs of state and, in particular, upon the relations between Crown and Parliament. The King adopted his proposal for removing Coke from his post as chief justice of the common pleas and appointing him to the King's Bench, while appointing Bacon attorney general in 1613. During the next few years Bacon's views about the royal prerogative brought him, as attorney general, increasingly into conflict with Coke, the champion of the common law and of the independence of the judges. It was Bacon who examined Coke when the King ordered the judges to be consulted individually and separately in the case of Edmond Peacham, a clergyman charged with treason as the author of an unpublished treatise justifying rebellion against oppression. It was he who instructed Coke and the other judges not to proceed in the case of commendams (*i.e.*, holding of benefices in the absence of the regular incumbent) until they had spoken to the King. Coke's dismissal in November 1616 for defying this order was quickly followed by Bacon's appointment as lord keeper in March 1617. The following year he was appointed lord chancellor and baron Verulam, and in 1620/21 he was created viscount St. Albans.

The main reason for this progress was his unsparing service in Parliament and the court, together with persistent letters of self-recommendation; according to the traditional account, however, he was also aided by his association with George Villiers, first earl and, later, duke of Buckingham, the King's new favourite. Bacon accepted the status of favourites—he had cultivated Robert

Carr (later earl of Somerset) a little before, during his period of favour—and regarded it as a calling according to the detailed and sound advice offered in a formal letter (published posthumously) to Villiers. It would appear that he became honestly fond of Villiers; many of his letters betray a feeling that seems warmer than timeserving flattery. Villiers, for his part, stood by Bacon to the best of his ability, even during Bacon's troubled later years.

Among Bacon's papers a notebook has survived, the *Commentarius Solutus* ("Loose Commentary"), which is revealing. It is a jotting pad "like a Marchant's wast booke where to enter all maner of remembrance of matter, fourme, business, study, towching my self, service, others, eyther sparsim or in schedules, without any maner of restraint." This book reveals Bacon reminding himself to flatter a possible patron, to study the weaknesses of a rival, to set intelligent noblemen in the Tower of London to work on serviceable experiments. It displays the multiplicity of his concerns: his income and debts, the King's business, his own garden and plans for building, philosophical speculations, his health, including his symptoms and medications, and an admonition to learn to control his breathing and not to interrupt in conversation. Plainly this was a busy man with no time, one would think, to write three dozen plays. Between 1608 and 1620 he prepared at least 12 draftings of his most celebrated work, the *Novum Organum*, and wrote several minor philosophical works; in 1609 he published the remarkable set of interpreted myths *De Sapientia Veterum* ("The Wisdom of the Ancients").

The major occupation of these years must have been the management of James, always with reference, remote or direct, to the royal finances. The King relied on his lord chancellor but did not always follow his advice. Bacon was longer sighted than his contemporaries and seems to have been aware of the imminent crisis of civil war; he dreaded innovation and did all he could, and perhaps more than he should, to safeguard the royal prerogative.

Whether his policies were sound or not, it is evident that he was, as he later said, "no mountebank in the King's services." James Spedding, one of Bacon's biographers, has drawn up a handsome testimonial: James I knew "that for the last 15 years he had been the most laborious, affectionate, zealous, attentive, faithful, and modest of servants, and the most moderately rewarded."

Fall from power. By 1621 Bacon must have seemed impregnable, a favourite not by charm (though he was witty and had a dry sense of humour) but by sheer usefulness and loyalty to his sovereign; lavish in public expenditure (he was once the sole provider of a court masque); dignified in his affluence and liberal in his household; winning the attention of scholars abroad as the author of the *Novum Organum*, published in 1620, and the developer of the *Instauratio Magna* ("Great Instauration"), a comprehensive plan to reorganize the sciences and to restore man to that mastery over nature that he was conceived to have lost by the fall of Adam.

Charges of bribery

But Bacon had his enemies. Two charges of bribery were raised against him before a committee of grievances over which he himself presided. The shock appears to have been twofold, because Bacon, who was casual about the incoming and outgoing of his wealth, was unaware of any vulnerability and was not mindful of the resentment of two men whose cases had gone against them in spite of gifts they had made with intent of bribing the judge. The blow caught him when he was ill, and he pleaded for extra time to meet the charges, explaining that genuine illness, not cowardice, was the reason for his request. Meanwhile, the House of Lords collected another score of complaints. Bacon admitted the receipt of gifts but denied that they had ever affected his judgment; he made notes on cases and sought an audience with the King that was refused. Unable to defend himself by discriminating between the various charges or cross-examining witnesses, he settled for a penitent submission and resigned the seal of his office, hoping that this would suffice. The sentence was harsh, however, and included a fine of £40,000, imprisonment in the Tower of London during the King's pleasure, disablement from holding any state office, and exclusion from Parliament and the verge of court (*i.e.*, the jurisdiction of the Marshalsea Court, which settled disputes in which members of the royal household were party). Bacon commented to Buckingham: "I acknowledge the sentence just, and for reformation's sake fit, *the justest Chancellor that hath been in the five changes since Sir Nicolas Bacon's time.*" The magnanimity and wit of the epigram sets his case against the prevailing standards. To the lord keeper he confessed: "My affliction hath made me understand myself better and not worse; yet loving advice, I know, helps well."

Bacon did not have to stay long in the Tower, but he found the ban that cut him off from access to the library of Charles Cotton, an English man of letters, and from consultation with his physician more galling. He came up against an inimical lord treasurer, and his pension payments were delayed. He lost Buckingham's goodwill for a time and was put to the humiliating practice of roundabout approaches to other nobles and to Count Gondomar, the Spanish ambassador; what remissions came were hard earned by vexations and disappointments. Despite all this his courage held, and the last years of his life were spent in work far more valuable to the world than anything he had accomplished in his high office. Cut off from other services, he offered his literary powers to provide the King with a digest of the laws, a history of Great Britain, and biographies of Tudor monarchs. He prepared memoranda on usury and on the prospects of a war with Spain; he expressed views on educational reforms; he even returned, as if by habit, to draft papers of advice to the King or to Buckingham and composed speeches he was never to deliver. Some of these projects were completed, and they did not exhaust his fertility. He wrote: "If I be left to myself I will graze and bear natural philosophy." Two out of a plan of six separate natural histories were composed—*Historia Ventorum* ("History of the Winds") appeared in 1622 and *Historia Vitae et Mortis* ("History of Life and Death") in the following year. Also in 1623 he published the *De Augmentis Scientiarum*, a Latin translation, with many additions, of the *Advancement of Learning*. He also corresponded with Italian thinkers and urged his works upon them. In 1625 a third and enlarged edition of his *Essayes*, which had first appeared in 1597, was published.

Profuse literary production of his last years

Certainly, adversity discovered in Bacon the virtues of patience, unimpaired intellectual vigour, and a good sense of acceptance. Physical deprivation distressed him for himself and his wife, but what hurt most was the loss of favour; it was not until January 20, 1622/3, that he was admitted to kiss the King's hand; the full pardon that would have cleared his name never came. Finally, in March 1626, driving one day near Highgate (a district to the north of greater London) and deciding on impulse to discover whether snow would delay the process of putrefaction, he stopped his carriage, purchased a hen, and stuffed it with snow. He was seized with a sudden chill, which brought on bronchitis, and died on April 9, 1626.

THOUGHT

Internal conflict in Bacon's works

Even the simplest recitation of relevant facts reveals that a determining principle in Bacon's composition was the constant competition between two strains: he could argue plausibly now for the active and now for the contemplative way of life; what he could not do was commit himself solely to either. Although each avocation is complicated enough to call for separate consideration, it is essential to recognize their ultimate interdependence. In periods of optimism he argued that one could be made to support the other, with money and position ensuring leisure for philosophical speculation; in more realistic periods of gloom, he came to know that the rivalry had been dangerous and damaging. Change of focus from short-term to long-term calculations was embarrassing for him and for his critics. It was one thing to begin in 1610 the *New Atlantis*—a utopian fiction that describes an ideal state in which the principles of Bacon's philosophy are carried out by political machinery and

under state guidance—and another thing to cope with James I's dilemmas. In the *De Sapientia Veterum*, interpreting the riddles of the Sphinx, he observed:

> for so long as the object of meditation and inquiry is merely to know, the understanding . . . finds in the very uncertainty of conclusions and variety of choice a certain pleasure and delight; but when they pass from the Muses to Sphinx, . . . whereby there is a necessity to present action, choice and decision, then they begin to be powerful and cruel; and unless they be cowed, they strangely torment and worry the mind.

Plan for the restoration of philosophy. Whatever may be thought of Bacon's purposes in attending to his political fortunes and patriotic aims, his sincerity as a philosopher cannot be disputed. If to leave works unfinished and huge plans only sketched in is to fail, then he failed in his philosophical endeavours as in his political ones. Such an estimate, however, takes no account of his true vision—the command of nature by way of obeying her.

His philosophical writings and sources of his thought. The most convenient mapping of Bacon's philosophical writings is to be found in the list of contents in the first three volumes of the edition of his works by R.L. Ellis and J. Spedding. Firstly, there are the completed parts of his *Instauratio Magna:* the magnificent prefatory matter, the *Novum Organum* (a presentation of a new method of logic), the "Parasceve ad Historiam Naturalem et Experimentalem" (a short sketch of the requirements of a fundamental natural history), and the *De Augmentis Scientiarum* (a classification of the sciences based on an analysis of the faculties and objects of human knowledge). Secondly, there are writings that were evidently intended at one time or another to be parts of the *Instauratio:* three short treatises, a few fragments, and a collection of observations of phenomena entitled *Sylva Sylvarum*. Lastly, there is a group of scientific and philosophical speculations not, it is assumed, to be included but interesting as exhibiting something of the emergence of the author's ideas and the variety of his styles.

Modern scholars have studied the development of the distinctive elements in his system beginning with the germinal statements in a masque speech praising knowledge; in a short presentation such as this article, however, the outlines must be drawn from his more mature expressions. In general, two things should be borne in mind. Bacon's chaplain-secretary has asserted that he wasted none of his time or mental reserves on what he could learn from other men's books. Modern critics have shown that he derived a great deal from Bernardino Telesio (a 16th-century philosopher who revolted against medieval Aristotelianism and advocated the empirical method), from Juan Huarte, a 16th-century Spanish philosopher and physician, and from medieval and Renaissance encyclopaedic compilations. What he read he used; it became his substance and was often the better for passing through the sieve of his good sense. Secondly, it was his deliberate intention to stake claims. This habit makes his frequent repetitions and modifications easier to understand.

Advancement of Learning

His theory of knowledge. The value that Bacon attached to knowledge is seen in the first book of the *Advancement of Learning*, which is often neglected because of ignorance of the attacks on knowledge at this time that made a defense pertinent. Bacon was not merely beating the air, however; there was religious, social, and skeptical opposition. First, knowledge was defended from the suspicions and perversions that it suffered from the zeal of divines, from the criticisms of politicians who contended that knowledge weakened action, and from the discredits brought upon it by learned men themselves. Bacon had no use for the manners of pedants, but he did believe that the learned can claim magnanimity in that they are aware of the frailty of their persons, the casualty of their fortunes, and the dignity of their souls and vocations. He deplored fantastic, contentious, and "delicate" learning; men should strive for substantive matter before words, and the matter itself must not decay into the tortuous questions of the medieval Scholastics. More serious is the deceit that destroys the essential form of knowledge, for truth of being and truth of knowing are one. He classed as "peccant humours" the affectations of extremes of novelty or antiquity, distrusted the premature reduction of knowledge into "arts," suspected the admiration of human nature to the neglect of nature proper; he foresaw the bias of the specialist and the dangers of a hasty dismissal of doubts. The greatest error of all is the mistake of misplacing the end of man's knowledge, which is to give a true account of the gift of reason.

Bacon had a love of order, and the second book of the *Advancement of Learning* makes this clear. Bacon's "province"—embraced by the phrase "all knowledge"—is divided according to the faculties of memory, imagination, and reason with elaborate but serviceable subdividing. Bacon paid slight attention to divinity and poetry in this survey because they were flourishing; he preferred to concentrate upon areas of defect or deficiency. He addressed the work to the King in the unpropitious year of the Gunpowder Plot—a conspiracy for blowing up James I and the Parliament on November 5, 1605—and produced the enlarged version in 1623.

Although he was sincerely convinced of the wholeness of knowledge, Bacon made no secret of the bias of his interest: the "domain of philosophy and the sciences" is the realm "without which I care not to live." In the proemium, preface, and plan for the *Magna Instauratio* he proposed as a severe and solitary task the radical recasting of the commerce between the mind of man and the nature of things; he looked for the release of inventive powers as a result of the improved ways of exploring nature's subtlety; he searched for principles that will direct action. In his restoration, syllogistic reasoning is to be demoted in favour of a critical induction based on experimental testing. His aim was to subdue and overcome the necessities and miseries of humanity. In this introduction to his great work, the plan for a natural history was outlined; samples of his scientific method at work were promised, together with a garnering of miscellaneous observations; and, finally, the beginning of the philosophy to which these preparations lead was indicated. A sincere dedication to the glory of God graces the undertaking. The substance of his great undertaking could be pieced together from other passages in public and private writings; the doctrine is also delivered chiefly through the mythological stories in the *De Sapientia Veterum*.

Novum Organum

In *Novum Organum* the teaching was cast into aphorisms in which a group of three ideas work like leaven. They are: the need for a new logic (thus the title of the collection); the attempt to discover the "forms" of what he believed to be a limited number—an "alphabet"—of simple natures (*i.e.*, qualities such as heat or whiteness); and the collection of a comprehensive natural history. Taken separately the three are not difficult to grasp; it is their relatedness that has been found puzzling and unsatisfactory. Bacon envisaged knowledge as a pyramid with natural history as its base, physics as the middle, and metaphysics as the vertical point. Intermittently, he attended to the base, contributing what he could. His valuation of this part of the work seems to have increased, and toward the end of his life he gave it prime importance.

Philosophy of science. His standing as a scientist, in the modern sense of the term, is low. He is found to have been out of date and overcautious in astronomical theory; he failed to recognize the Scottish mathematician John Napier's invention of logarithms and, in general, underestimated mathematics; he was ill informed about leverage, the acceleration of bodies, and the circulation of the blood. Spedding sought to find a cause for these deficiencies and viewed Bacon's failure to recognize the work of some of the foremost men of his day as symptomatic of an intellectual imbalance; his mind, so quick to note likenesses, was not equally ready to distinguish differences. He stood in a dangerous isolation, unwilling to handle what he was not fully equipped to understand. Some critics have referred to his habitual inaccuracy.

Bacon, characteristically enough, was ahead of his critics when (gracing the weakness by a metaphor) he admitted, "I can not thridd needles so well."

His so-called inductive method and its defects. Bacon's loudly announced views concerning the need for a new logic have led to a closer association of his name with induction than is credible. Inductive reasoning did not originate with Bacon, nor does the process as developed by the natural sciences since his day derive directly from his formulation; what he urged, however, is not merely derivative or valueless. Setting aside both deduction and simple enumerative induction as means of interpreting nature's ways, he elaborated his own method by which ascent was to be made from particulars to axioms of a middle class of generality, and then by means of the control of negative and graded instances arranged in tables by presentation, exclusion, and rejection to axioms of the broadest degree of generality. The method, which is demonstrated with some elaboration in *Novum Organum* by means of a search for the essential nature of heat, has been criticized as open to two radical objections. In the first place, the laboriousness of his method requires the aid of hypotheses, and Bacon gave no help toward the formulation of such concepts, although, ironically enough, he supplied just such an aid, or *bona notio*, in the investigation of heat. He was so intent on advertising his new instrument as foolproof that he refused to allow enough for the inventive genius of other men or even to recognize the limits of his own share of genius. Secondly, the method breaks down when it is related to the end proposed, that is, the discovery of "forms." Bacon himself had great difficulty in giving an adequate and exact definition of what he meant by a form. A study of the various passages in the *Novum Organum* in which the definition of forms is attempted seems to show that Bacon's forms were not ideas or abstractions but highly general physical properties that are limitations or specific manifestations of some higher genus. Further, it is hinted that these general qualities may be looked upon as the modes of action of simple bodies. Part of the difficulty in understanding Bacon's method is the result of his choice of terms. He knowingly took over the terms metaphysic and form from medieval Scholasticism and modified their connotations; he was also likely to use the terms motion, law, and nature somewhat loosely.

Forms, laws, motion

Contributions to a philosophy of nature. Spedding was of the opinion that the doctrine of the forms is extraneous to Bacon's method, but it has not been so regarded by recent critics. A scrutiny of what Bacon meant by "law" when the term was used as a synonym for form was made by Adolfo Levi, a 20th-century Italian philosopher, by way of the unusual route of working backward from the test case of the inquiry into the form of heat. His interpretation was adopted by A.E. Taylor, an English historian of philosophy, and it carries important implications when the nature (or form) of Bacon's own genius is considered. According to Levi:

> In modern physics the laws are the expression of functional relationships which exist between phenomena, universal and necessary relationships of quantitative nature, while the Baconian forms consist in geometric-mechanic conditions, that is, they are essences contributed by structures and movements not determined by phenomenal relations.

Levi's interpretation hinges on the understanding of *motus* ("movement") in the phrase *lex actus sive motus* as nominative, and not, as formerly taken, as genitive; thus, the form of heat is movement, not the law of movement.

> Bacon, however, did not wish to determine quantitative relations—but to discover the inmost essence of the physical world, that is, to construct what are now called explanatory hypotheses for phenomena such as the mechanical concept of heat and the wave theory of light.

Such an ambition marks him out among scientists as a philosophical genius. When science seeks with such hypotheses to reach the intimate structure of reality, it becomes almost imperceptibly a philosophical intuition. According to Taylor:

> He has received high praise on the ground that, as has been alleged, his method enables us to answer a question which he was not raising, and had been depreciated on the ground that the true method for dealing with this question is not that which he recommends.

Bacon's reputation is thus somewhat tarnished; nevertheless, his contributions to science and philosophy are real. A remark in *Sylva Sylvarum* ("Forest of Forests") averring that "all bodies whatsoever, though they have no sense, yet they have perception" struck Alfred North Whitehead, an influential 20th-century British philosopher, as a genuine intuition. Furthermore, Bacon demonstrated great prudence in attaching so much importance to the discipline leading to the dynamic theory of motion; his promotion of the study of the natural sciences was greatly influential, even if his method was not followed. Thirdly, he is important for the grandeur of his vision and his heartening conviction of the unity of knowledge. Lastly, in the eyes of Taylor, Bacon has "the *temper* of a philosopher"; whatever may have been the failings of his character, it does not seem that excessive self-conceit was among them.

Enduring features of his search for truth

Bacon was humble in the search for truth; he was willing to admit uncertainty and incompleteness and referred his work to the continuing search of other men and ages. Instead of offering exaggerated acclamation or shrill rejection—and he has suffered from both extremes—it is now possible without insult to accord him a more tempered admiration. In moments of exaltation he foolishly claimed "to see an end of the matter," to have the "key" to things; but in more sober hours he knew better. The precious instrument of scientific method, he realized, would be useless without the natural history; the alphabet of simple natures might need revision; and the logic would need aids to reduce the dangers of confusion in its laboured procedure.

The idols. Perhaps the most valuable of these aids is the warning about the idols or false appearances of the mind. The topic arose five times as he worked out the form it should take. These deep fallacies of the mind, or general classes of errors into which the human mind is prone to fall, are arranged in the *Novum Organum* as the four *idola:* (1) The *idola tribus*, "idols of the tribe," are fallacies incident to humanity or the race in general. Of these, the most prominent are the proneness to suppose in nature greater order and regularity than there actually is; the tendency to support a preconceived opinion by affirmative instances, neglecting all negative or opposed cases; and the tendency to generalize from few observations or to give reality to mere abstractions, figments of the mind. Manifold errors also result from the weakness of the senses, which affords scope for mere conjecture; from the influence exercised over the understanding by the will and passions; from the restless desire of the mind to penetrate to the ultimate principles of things. (2) The *idola specus*, "idols of the cave," are errors incident to the peculiar mental or bodily constitution of each individual, because an individual's view of things is based on the state of his mind. Errors of this class are innumerable because there are numberless varieties of disposition; but some very prominent specimens can be indicated. These include the tendency to make all things subservient to or take the colour of some favourite subject; the extreme fondness and reverence for either what is ancient or what is modern; and excess in noting either differences or resemblances among things. (3) The *idola fori*, "idols of the market place," are errors arising from the influence exercised over the mind by mere words. This, according to Bacon, is the most troublesome kind of error and has been especially fatal in philosophy. Words introduce a fallacious mode of looking at things in two ways: first, there are some words that are really merely names for nonexistent things; secondly, there are names hastily and unskillfully abstracted from a few objects and applied recklessly to all that have the faintest analogy with these objects. (4) The *idola theatri*, "idols of the theatre," are fallacious modes of thinking resulting from received systems of philosophy and from erroneous methods of demonstration.

The Essayes. Bacon sought to help mankind not only with "the secrets of science" but also with "the difficulties of living," and he envisaged his new logic as extensible to human affairs. In the *Essayes* he attempted to study the "simple natures" of such things as ambition, dissimulation, revenge, and love. He recorded experiences derived from history and his own observations. He could not precisely set up experiments, but he did examine instances of behaviour and motivation, and upon these he generalized brilliantly. He subtitled the *Essayes* the "inwards of things" or "dispersed meditations" and associated this literary form not with the writings of the French author Michel de Montaigne but with the letters of Seneca. He found that this form of expression gave him scope for disinterested comment; it did not demand that he argue a case or urge a cause or manipulate anyone to make the best of things; it gave him the freedom to show both what men do and what men ought to do. This mixture of realism and the ideal of expediency and morality makes these succinct and lucid pieces not simple reading but full of worldly wisdom for the reader who has learned the author's technique.

PERSONAL CHARACTERISTICS

Bacon provided a formal self-portrait in the proemium to the *Magna Instauratio:*

> For myself, I found that I was fitted for nothing so well as for the study of truth; as having a mind nimble and versatile enough to catch the resemblances of things (which is the chief point) and at the same time steady enough to fix and distinguish their subtler differences; as being gifted by nature with desire to seek, patience to doubt, fondness to meditate, slowness to assert, readiness to reconsider, carefulness to dispose and set in order; and as being a man that neither affects what is new nor admires what is old, and that hates any kind of imposture.

John Aubrey, a 17th-century English antiquary, wrote a vivid portrayal of Bacon in which he noted that Bacon had a "delicate, lively hazel eie. Dr. Harvey told me it was like the eie of a viper." He noted Bacon's sensuousness: he had "musique in the next room where he meditated"; sweet herbs and flowers were always before him as he dined; his men's boots were of Spanish, not neat's, leather, because of the smell. From the few extant familiar letters of Bacon there is some evidence that he had an appreciation of simple friendships, but he did not invite intimacy; he frequently applied to himself the psalm likening himself to a stranger in life's pilgrimage.

Bacon was proud of his father; he had been affected by his mother, who fussed over his diet and late hours; he was attached to his brother. It seems that his marriage with Alice Barnham, though childless, was not unhappy, although toward the end of his life she gave him reason to alter his will. To judge by two essays, "On Love" and "On Marriage," Bacon had no use for romance. In general, everything points to him as having been reserved, wary, self-sufficient, and yet sensitive. The testimony of two men who had the most chance to study his habits and humours—his chaplain and secretary William Rawley and his friend Tobie Matthew, the son of an English prelate—offer tributes of discerning affection. "I never saw in him any trace of a vindictive mind," wrote Tobie Matthew. "It is not his greatness that I admire, but his virtue; it is not the favours I have received from him, . . . but his whole life and character." Rawley knew him as a good master and host, generous in conversation, eager to hear each man's observations, dictatorial at his table, courteous in the courts, a good servant himself and in return good to those who served him.

It has been observed that Bacon's concern was with things before men—*i.e.*, things of the mind—or, at best, with man rather than men. This attitude has given rise to many critical descriptions; the poet Alexander Pope, for example, added the damning "meanest" after the epithets "wisest, brightest." Few have sprung to Bacon's defense.

Intellectual and imaginative life

Although it is difficult to penetrate the private life of Bacon, his intellectual and imaginative life lies wonderfully open. Nicholas Hilliard, the miniaturist, is recorded as saying, "Oh, that I had a canvas to paint his *mind*." Bacon wrote of himself that he was "naturally fitted rather for literature than for anything else." He practiced a functional rhetoric and sought "masculine, plain sense"; nevertheless, his prose was enriched by the unobtrusive use of a variety of images. It is by style that he attracts, holds, and haunts; the reader is enabled to share the movement of his mind and is infected with the emotion that energized his thinking. He was interested in the theory of the arts of communication and expert in their practice, both as a speaker and as a writer. Equally evident is his sense of control; his sentences are elaborately balanced; his presentations are firmly structured.

His writings are pointers to the development of science and philosophy; they are preservers of a range of human experience, promoters of a belief that there was intended a marriage between nature and man's mind. Sir Henry Wotton, as provost of Eton College, recommended that his contemporary Bacon be "read in my domestic college as an ancient author." This heralded the tributes of the Royal Society in the next generation and the continued praise of his literary genius.

MAJOR WORKS

PHILOSOPHICAL WORKS: *The Twoo Bookes of Francis Bacon of the Proficience and Advancement of Learning Divine and Humane* (1605); *Novum Organum* (1620); *Historia Ventorum* (1622); *Historia Vitae et Mortis* (1623); *De Augmentis Scientiarum* (1623); an enlarged edition in Latin of the *Advancement of Learning.*

LITERARY AND HISTORICAL WORKS: *Essayes* (first published as 10 essays in 1597; enlarged to 38 in 1612; and to 58 in 1625); *De Sapientia Veterum* (1609); *The Historie of the Raigne of King Henry the Seventh* (1622); *The Translation of Certaine Psalmes into English Verse* (1625); *Apophthagmes new and old* (1625).

POLITICAL WORKS: *A Declaration of the Practices and Treasons attempted and committed by Robert, late Earle of Essex* (1601); "Certain Considerations touching the better Pacification and Edification of the Church of England" (1604); *Apologie in certaine imputations concerning the late Earle of Essex* (1604).

POLITICAL WORKS: "The Felicity of Queen Elizabeth" (1651; translation of "In Felicem Memoriam Elizabethae," written in 1608).

LEGAL WORKS: "The Elements of the Common Lawes of England" (2 parts, 1630; the second tract is probably not by Bacon); *Cases of Treason* (1641); "The Learned Reading of Sir Francis Bacon upon the Statute of Uses" (1642).

POSTHUMOUSLY PUBLISHED WORKS: (PHILOSOPHICAL): *Sylva Sylvarum* (1627, with the unfinished *New Atlantis*). (LITERARY): *Promus of Formularies and Elegancies* (1883); *The Poems of Francis Bacon,* collected and edited by the Rev. A.B. Grosart (1870).

BIBLIOGRAPHY. R.W. GIBSON, *Francis Bacon: A Bibliography of His Works and of Baconiana to the Year 1750* (1950 and suppl. 1959); DOUGLAS BUSH, *English Literature in the Early Seventeenth Century,* 2nd ed. rev. (1962), contains a very good critique of Bacon's historical, political, and legal writings.

Biographies: JAMES SPEDDING, *An Account of the Life and Times of Francis Bacon,* 2 vol. (1878), an abridgement of his edition of the *Letters and Life;* R.W. CHURCH, *Bacon* (1881); EDWIN ABBOTT, *Francis Bacon: An Account of His Life and Works* (1885); JOHN NICHOL, *Francis Bacon: His Life and Philosophy,* 2 vol. (1888–89), including a study of Bacon's thought; FULTON H. ANDERSON, *Francis Bacon: His Career and His Thought* (1962), the best short life. See also MARY STURT, *Francis Bacon* (1932); and CHARLES W.S. WILLIAMS, *Bacon* (1933), for more personal, selective interpretations, each with some valuable insights.

Critical studies (philosophy): Besides Nichol (above), see THOMAS FOWLER, *Bacon* (1881); GASTON SORTAIS, *La Philosophie moderne depuis Bacon jusqu'à Leibniz,* vol. 2 (1920); ADOLFO LEVI, *Il pensiero di Francesco Bacone* (1925); C.D. BROAD, *The Philosophy of Francis Bacon* (1926), an excellent lecture; A.E. TAYLOR, *Francis Bacon* (1927), an invaluable exposition of Bacon's ideas in the light of modern criticism; FULTON H. ANDERSON, *The Philosophy of Francis Bacon* (1948), a major influential study; BENJAMIN FARRINGTON, *Francis Bacon: Philosopher of Industrial Science* (1949), on his social objectives; PAOLO ROSSI, *Francesco Bacone: dalla magia alla scienza* (1957; Eng. trans. *Francis Bacon: From Magic to Science,* 1968).

Miscellaneous critical studies: CHARLES W. LEMMI, *The Classic Deities in Bacon* (1933), an important study of *De*

Sapientia Veterum; see also the essays by GEOFFREY BULLOUGH and RUDOLPH METZ in *Seventeenth Century Studies Presented to Sir Herbert Grierson* (1938, reprinted 1967); KEITH R. WALLACE, *Francis Bacon on Communication and Rhetoric* (1943); VIRGIL K. WHITAKER, *Francis Bacon's Intellectual Milieu* (1962); BRIAN VICKERS, *Francis Bacon and Renaissance Prose* (1968).

(K.M.L.)

Bacon, Roger

Roger Bacon was an English Franciscan friar of the 13th century who was active as a philosopher, scientist, and educational reformer. Having become familiar with virtually all of the science known in his time, Bacon (as he himself complacently remarked) displayed a prodigious energy and zeal in the pursuit of experimental science; indeed, his studies were talked about everywhere and eventually won him a place in popular literature as a wonder worker. Bacon thus represents a historically precocious expression of the empirical spirit of experimental science, though his actual practice of it seems to have been exaggerated.

Early life

Bacon was born of a wealthy family about 1220. The place of his birth is uncertain; in one tradition it is assigned to Ilchester in Somerset, in another to the parish of Bisley in Gloucester. He was well versed in the classics and enjoyed the advantages of an early training in the quadrivium (geometry, arithmetic, music, and astronomy); and he boasted that he had frequently "heard" and "read" the works of Aristotle. Inasmuch as he later lectured at Paris, it is probable that his master of arts degree was conferred there, presumably not before 1241—a date in keeping with his claim that he saw the Franciscan professor Alexander of Hales (who died in 1245) with his own eyes and that he heard the master scholar William of Auvergne (died 1249) dispute twice in the presence of the whole university.

University and scientific career. In the earlier part of his career, Bacon lectured in the faculty of arts on Aristotelian and pseudo-Aristotelian treatises, displaying no indication, however, of his later preoccupation with science. His Paris lectures, important in enabling scholars to form some idea of the work done by one who was a pioneer in introducing the works of Aristotle into Western Europe, reveal an Aristotelianism strongly marked by Neoplatonist elements stemming from many different sources. The influence of Avicenna on Bacon has been exaggerated.

About 1247 a considerable change took place in Bacon's intellectual development. From that date forward he expended much time and energy and huge sums of money in experimental research, in acquiring "secret" books, in the construction of instruments and of tables, in the training of assistants, and in seeking the friendship of savants —activities that marked a definite departure from the usual routine of the faculty of arts. The change was probably caused by his return to Oxford and the influence there of the great scholar Robert Grosseteste, a leader in introducing Greek learning to the West, and his student Adam de Marisco, as well as that of Thomas Wallensis, the bishop of St. David's. From 1247 to 1257 Bacon devoted himself wholeheartedly to the cultivation of those new branches of learning to which he was introduced at Oxford—languages, optics, and alchemy—and to further studies in astronomy and mathematics. It is true that Bacon was more skeptical of hearsay claims than were his contemporaries, that he suspected rational deductions (holding to the superior dependability of confirming experiences), and that he extolled experimentation so ardently that he has often been viewed as a harbinger of modern science more than 300 years before it came to bloom. Yet research on Bacon suggests that his characterization as an experimenter may be overwrought. His originality lay not so much in any positive contribution to the sum of knowledge as in his insistence on fruitful lines of research and methods of experimental study. As for actual experiments performed, he deferred to a certain Master Peter de Maricourt (Maharn-Curia), a Picard, who alone, he wrote, understood the method of experiment and whom he called *dominus experimentorum.*

Stress on experiment

Bacon, to be sure, did have a sort of laboratory for alchemical experiments and carried out some systematic observations with lenses and mirrors. His studies on the nature of light and on the rainbow are especially noteworthy, and he seems to have planned and interpreted these experiments carefully. But his most notable "experiments" seem never to have been actually performed; they were merely described. He suggested, for example, that a balloon of thin copper sheet be made and filled with "liquid fire"; he felt that it would float in the air as many light objects do in water. He seriously studied the problem of flying in a machine with flapping wings. He was the first person in the West to give exact directions for making gunpowder (1242); and, though he knew that, if confined, it would have great power and might be useful in war, he failed to speculate further. (Its use in guns arose early in the following century.) Bacon described spectacles (which also came soon into use); elucidated the principles of reflection, refraction, and spherical aberration; and proposed mechanically propelled ships and carriages. He used a camera obscura (which projects an image through a pinhole) to observe eclipses of the sun.

Career as a friar. In 1257 another marked change took place in Bacon's life. Because of ill health and his entry into the Order of Friars Minor, Bacon felt (as he wrote) forgotten by everyone and all but buried. His university and literary careers seemed finished. His feverish activity, his amazing credulity, his superstition, and his vocal contempt for those not sharing his interests displeased his superiors in the order and brought him under severe discipline. He decided to appeal to Pope Clement IV, whom he may have known when the latter was in the service of the Capetian kings of France. In a letter (1266) the Pope referred to letters received from Bacon, who had come forward with certain proposals covering the natural world, mathematics, languages, perspective, and astrology. Bacon had argued that a more accurate experimental knowledge of nature would be of great value in confirming the Christian faith, and he felt that his proposals would be of great importance for the welfare of the church and of the universities. The Pope desired to become more fully informed of these projects and commanded Bacon to send him the work. But Bacon had had in mind a vast encyclopaedia of all the known sciences, requiring many collaborators, the organization and administration of which would be coordinated by a papal institute. The work, then, was merely projected when the Pope thought that it already existed. In obedience to the Pope's command, however, Bacon set to work and in a remarkably short time had dispatched the *Opus majus* ("Greater Work"), the *Opus minus* ("Lesser Work"), and the *Opus tertium* ("Third Work"). He had to do this secretly and notwithstanding any command of his superiors to the contrary; and even when the irregularity of his conduct attracted their attention and the terrible weapons of spiritual coercion were brought to bear upon him, he was deterred from explaining his position by the papal command of secrecy. Under the circumstances, his achievement was truly astounding. He reminded the Pope that, like the leaders of the schools with their commentaries and scholarly summaries, he could have covered quires of vellum with "puerilities" and vain speculations. Instead, he aspired to penetrate realms undreamed of in the schools at Paris and to lay bare the secrets of nature by positive study. The *Opus majus* was an effort to persuade the Pope of the urgent necessity and manifold utility of the reforms that he proposed. But the death of Clement in 1268 extinguished Bacon's dreams of gaining for the sciences their rightful place in the curriculum of university studies.

Appeal to the Pope and *Opus majus*

Bacon projected yet another encyclopaedia, of which only fragments were ever published, viz., the *Communia naturalium* ("General Principles of Natural Philosophy") and the *Communia mathematica* ("General Principles of Mathematical Science"), written about 1268. In 1272 there appeared the *Compendium philosophiae* ("Compendium of Philosophy"). In philosophy—and even Bacon's so-called scientific works contain lengthy philosophical digressions—he was the disciple of Aristotle and

not of St. Augustine or the Persian philosopher Avicenna; even though he did incorporate Neoplatonist elements into his philosophy, his thought remains, nevertheless, Aristotelian in its main lines.

Sometime between 1277 and 1279, Bacon was condemned to prison by his fellow Franciscans because of certain "suspected novelties" in his teaching. The condemnation was probably issued because of his bitter attacks on the theologians and scholars of his day, his excessive credulity in alchemy and astrology, and his penchant for millenarianism under the influence of the prophecies of Abbot Joachim of Fiora, a mystical philosopher of history. How long he was imprisoned is unknown. His last work (1292), incomplete as so many others, shows him as aggressive as ever. A traditional source relates that he died in 1292 and was buried in the Franciscan church at Oxford. Exaggerated accounts of his admirable experiments doubtless suggested the title *doctor mirabilis* ("wonderful teacher") by which he became known to posterity.

BIBLIOGRAPHY. ANDREW G. LITTLE (ed.), *Roger Bacon* (1914), comprehensive and critical essays contributed by various eminent scholars on the occasion of the commemoration of the seventh centenary of Bacon's birth, is a collection that still retains its value. Two works that are complementary and contain fresh biographical insights and extensive bibliographies are THEODORE CROWLEY, *Roger Bacon: The Problem of the Soul in His Philosophical Commentaries* (1950), presenting his philosophical positions; and STEWART C. EASTON, *Roger Bacon and His Search for a Universal Science: A Reconsideration of the Life and Work of Roger Bacon in the Light of His Own Stated Purposes* (1952). ERICH HECK, *Roger Bacon: Ein mittelalterlicher Versuch einer historischen und systematischen Religionswissenschaft* (1957), contains a critical survey of previous work and detailed studies of Bacon's approach to the scientific study of religion. A.C. CROMBIE, *Robert Grosseteste and the Origins of Experimental Science, 1100–1700*, pp. 139–162 (1953), is a balanced account of Bacon's contributions to science.

(T.Cr.)

Bacteria

Bacteria are microscopic organisms present in almost all natural environments, often in extremely large numbers: billions in a gram of rich garden soil and millions in one drop of saliva. They constitute the class Schizomycetes, of the division Schizomycophyta. The influence of bacteria in the biosphere is incalculable. Without them, the soil would not be fertile and thus could not sustain plants, on which animals ultimately depend for food. Although some bacteria cause disease, most species are benign, and many are involved in processes of direct benefit to man. For bacterial diseases, see the articles INFECTIOUS DISEASES; DISEASES OF ANIMALS; DISEASES OF PLANTS.

GENERAL FEATURES

Appearance. Bacteria are unicellular micro-organisms, among the smallest living creatures known (only the related rickettsias and the viruses are smaller). One of the periods on this page would cover about 250,000 average-sized bacteria, which are measured in micrometres, or microns (one micron, μ, equals $\frac{1}{1,000}$ millimetre, or $\frac{1}{25,000}$ inch). There are three bacterial cell types on the basis of shape (see Figure 1): spherical (coccus), rodlike (bacillus), and spiral (spirillum). Each type retains its character under standard conditions of laboratory cultivation but may show changes in appearance under different environments. When conditions are favourable, bacteria grow rapidly (some can reproduce every 15 minutes), forming visible colonies on culture plates in the laboratory. The colonies are distinctive in colour, shape, and texture for certain species and in some cases for the different varieties, or strains, within a single species.

Cell shapes

Distribution. Bacteria are ubiquitous, occurring in virtually every conceivable environment: from polar ice to hot springs; from mountain tops to the ocean deeps; from plant and animal bodies to forest soils. Most bacteria are active in environments in which the temperature is above 5° C (about 40° F); some marine and soil types

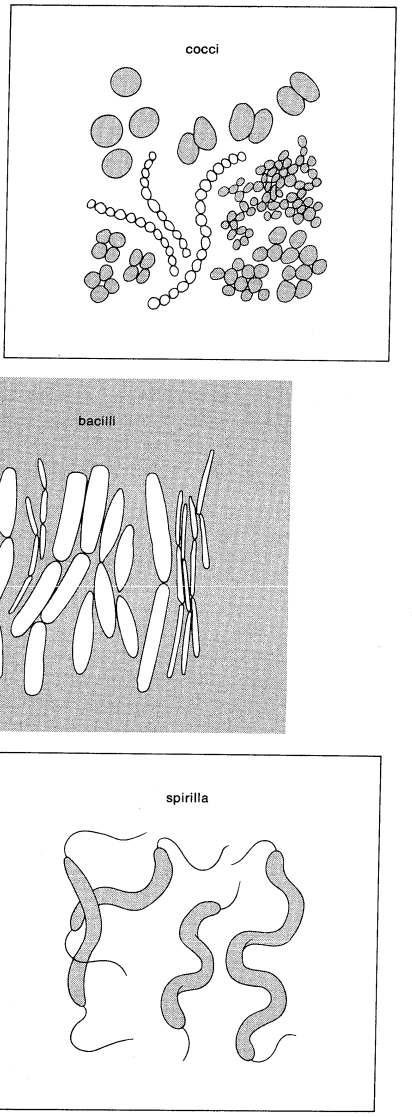


Figure 1: The three major bacterial cell types.
Reprinted from *The Spectrum of Life* by Harold A. Moore and John R. Carlock. Copyright ©1970, Harper & Row, Publishers, Inc.

are exceptional in being active at temperatures near or slightly below 0° C (32° F). The upper limit is around 30° C (86° F) for soil bacteria and 37° C (99° F) for animal parasites; the maximum temperature, above which growth does not occur, is around 70° C (160° F). Beyond these limits bacteria become inactive. Some survive in a dormant (or spore) state, reviving when conditions become more favourable. This capability has allowed bacteria to become perhaps the most widespread organisms on Earth.

Ecological significance

Importance. Bacteria are instrumental in performing numerous critical biochemical transformations of substances in nature, changing them from complex to simple compounds that can be used by plants, man, and other animals. Bacteria in the rumen (largest compartment of the stomach) of the cow digest the cellulose in grasses and animal feed, thus making this matter available as a nutrient for the animal. The organic waste substances in sewage are degraded by bacteria and transformed into compounds that are suitable nutrients for plant growth. In fact, all of the remains of animals and plants are eventually converted to soil through the activities of bacteria and other micro-organisms and thus made available again to growing plants.

It can be assumed—until evidence disproves it—that any naturally occurring substance can be degraded (metabolized) by some species of bacteria. In some instances, these biochemical transformations are judged by man to be beneficial: as when *Streptomyces griseus*

produces the antibiotic streptomycin as a by-product of its metabolism; or when certain strains of bacteria produce cheese and other dairy products by metabolizing constituents of milk; or when certain nitrogen-accumulating bacteria add needed nitrogen to the soil. Other transformations, however, may be detrimental to man's interests, as when *Clostridium botulinum* excretes a toxin responsible for botulism (a type of food poisoning), or when certain strains of bacteria excrete substances that cause disease. Less detrimental are the effects of bacterial deterioration: spoilage of food, corrosion of metals, decay of wood, and other undesirable alterations of substances.

Higher animals, including man, live in constant intimacy with large numbers and a great variety of bacteria. The oral cavities, intestinal tracts, and skin are inhabited by bacteria that, under normal circumstances, create no problems (Figure 2). There are occasions, however, when bacteria break through the normal body barriers and cause an infection. Certain bacteria are more prone to this behaviour than others and are called pathogens, or disease producers (Figure 2). Some pathogens have an affinity for specific parts of the body: meningococcal bacteria infect the meninges, or brain membranes; tubercle bacteria invade the lungs; and diphtheria-causing bacteria establish themselves in the throat. Other bacterial pathogens exhibit less specificity: staphylococcal bacteria, for example, may infect the skin, causing boils or furuncles; the bloodstream, causing septicemia (blood poisoning); or the bones, establishing a condition known as osteomyelitis.

In some cases the biochemical transformations of bacteria are actually exploited industrially to convert raw materials into products that are economically valuable. When the substance is abundant and cheap and the product is valuable and in demand, a bacterial industrial process is likely to be established. Selected examples of products manufactured on an industrial scale through the utilization of bacteria are shown in Table 1. Other industrial processes are developed by man to combat the deterioration of materials by micro-organisms (see FOOD PRESERVATION).

Bacteria as experimental subjects

Bacteria represent a form of life that can be conveniently studied under laboratory conditions. Furthermore, and more importantly, many of the biological processes that take place in the bacterial cell are closely related, if not identical, to processes that take place in higher organisms, including man. Thus, the bacterial cell provides an extremely useful model for the study of intricate biological, physiological, and biochemical processes. Indeed, much of the knowledge that has been acquired since the end of World War II in biochemical genetics, enzymes, and the synthesis of vitamins and their functions has resulted from investigations of bacteria.

NATURAL HISTORY

In view of the widespread occurrence of bacteria, it is not surprising that measures are taken by man to reduce or control their numbers in habitats that are exploited for human welfare. In the following sections, therefore, emphasis is on those particular environments of bacteria that impinge on man's needs.

Growth and reproduction. The life cycle of bacteria involves growth and reproduction, as in other organisms; certain features, however, are unique to micro-organisms.

The term growth, in the microbiological sense, refers to increase in a given population rather than to increase in

A.W. Rakosy—EB Inc.

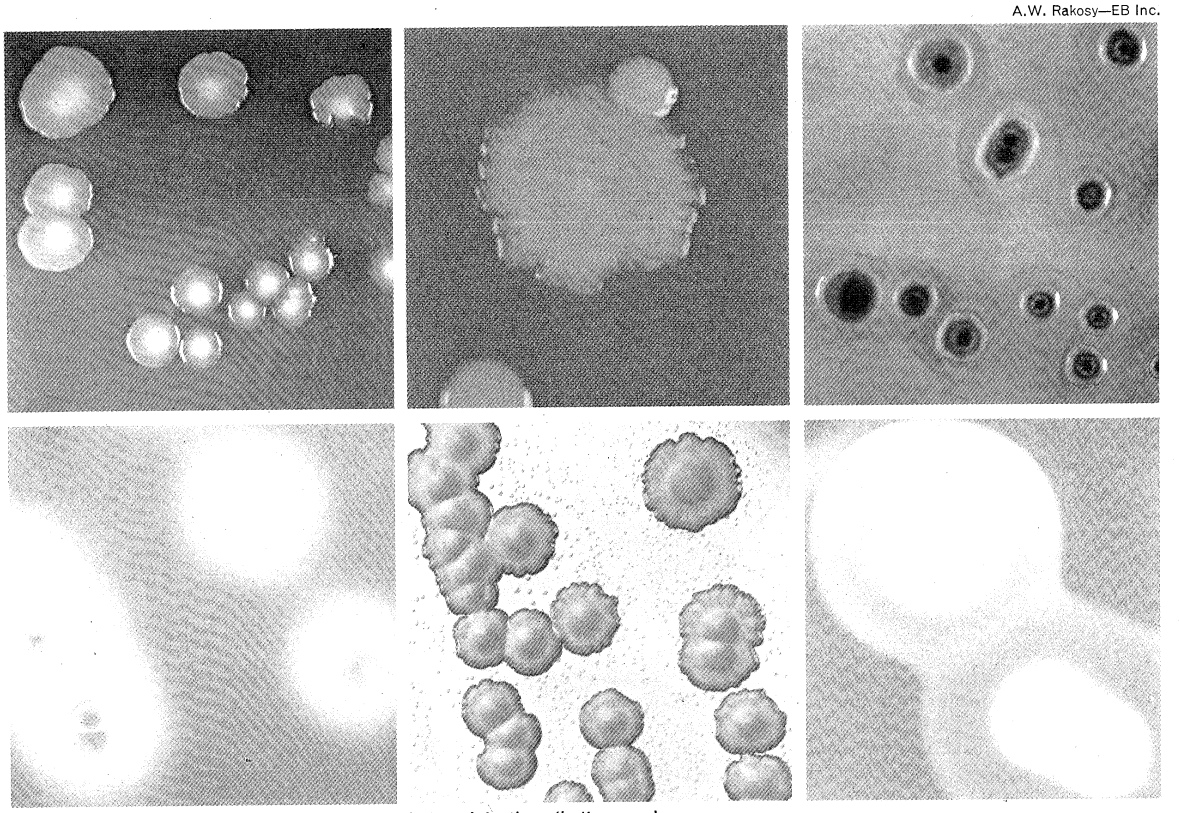


Figure 2: *Typical body flora (top row) and bacteria from infections (bottom row).* (Top left) *Neisseria flava* from the human nasal passage (magnified about 7 ×). (Top centre) Aerobic *Lactobacillus* (large colony) and *Staphylococcus albus* (small colony) from the human vagina (magnified 11 ×). (Top right) *Escherichia coli*, grown on eosin methylene blue dye, from a normal stool (magnified about 7 ×). (Bottom left) Colonies of beta-hemolytic streptococci on blood agar show zones cleared of blood cells surrounding each colony; these bacteria were isolated from the sore throat of a child who later developed rheumatic fever (magnified about 19 ×). (Bottom centre) *Hemophilus influenzae* (small colonies) isolated from a child suffering from spinal meningitis. Position of the *H. influenzae* colonies around larger *Staphylococcus* colonies, which provide a factor necessary for their growth, illustrates the "satellite phenomenon" (magnified about 14 ×). (Bottom right) *Clostridium perfringens*, isolated from infected human gallbladder, as grown under strict anaerobic conditions on blood agar. Zones of complete and partial clearing are characteristic of this species of *Clostridium* (magnified 8.5 ×).

Table 1: Examples of Compounds Produced by Bacteria on an Industrial Scale

product	bacterium	substrate
Lactic acid	*Lactobacillus delbrueckii*	acid-hydrolyzed cornstarch, or whey, plus nutrient and $CaCO_3$
Bacterial amylase	*Bacillus subtilis*	vegetable protein plus sugar for surface cultivation; starch, cereal, grain, and protein for subsurface cultivation
Bacterial protease	*B. subtilis*	protein, carbohydrate, salts
Dextran	*Leuconostoc mesenteroides*	sucrose plus nutrients
Cobalamin (vitamin B_{12})	*Streptomyces olivaceus*	distiller's solubles, dextrose, $CaCO_3$
	Propionibacterium freudenreichii	$CoCl_2$
Vinegar	*Acetobacter* species	alcohol
Streptomycin	*Streptomyces griseus*	hydrolyzed protein and sugar
Monosodium glutamate	*Micrococcus* species	sugar

the size of an individual micro-organism, or bacterium in this context. There are, to be sure, changes in the size of an individual bacterial cell at certain stages of the multiplication process, as is shown below.

Reproduction. Bacteria characteristically reproduce by an asexual process called binary fission, in which one cell divides into two new cells (Figure 3). A single bacterial cell, under optimum physical conditions for growth, performs metabolic functions including synthesis of intracellular substances. The cell elongates, and the cell wall becomes pinched in at the midpoint: finally, a transverse cell wall separates the parent cell into two new cells (daughter cells) that separate, and the process commences again. The reproduction and multiplication process is by geometric progression: one cell forms two, two cells form four, four form eight, eight form 16, 16 form 32, and so forth. The time required for the populations to double—*i.e.*, for one cell to divide in two—is the generation time (*G*); it can be calculated from the following formula:

Generation time

$$G = \frac{t}{n} = \frac{t}{3.3 \log b/B}.$$

In the formula, *B* is the number of bacteria at the start of the investigation; *b* is the number of bacteria at the end of the time period; *t* is the time period; and *n* is the number of generations. The experimental data—that is, the values for *B*, *b*, and *t*—must be obtained during the period of the total growth cycle known as the log phase of growth (described below). The generation times of bacterial species vary over a wide range. *Escherichia coli*, one of the most rapidly growing bacteria, has a generation time of approximately 15 minutes; *Mycobacterium tuberculosis*, a slow-growing bacterium, can have a generation time as long as 16 hours.

Although binary fission is the characteristic and typical mode of reproduction for the true bacteria (order Eubacteriales), among some bacterial species, particularly those of the higher orders of bacteria, other processes of reproduction occur. *Rhodomicrobium vannielii* (order Hyphomicrobiales), for example, exhibits a budding type of reproduction; *Streptomyces* (order Actinomycetales) species produce chains of spores; *Mycoplasma* (order Mycoplasmatales) reproduce by the segmentation of elementary units within a body surrounded by a membrane.

There are also instances of sexual reproduction (conjugation) among bacterial species. It occurs at low frequency among bacteria found in the intestine (enteric, or coliform, bacteria—*Escherichia*, *Shigella*, and *Salmonella*). In this process, conjugal pairs, or mating types, of bacteria make transient physical contact. Conjugal pairs consist of a donor (male) and a recipient (female) cell. In conjugation, a piece of chromosome from the donor cell is transferred into the recipient cell, in which it becomes a part of the recipient's chromosome. This is one way by which genetic material is exchanged in bacteria.

The bacterial growth curve. When bacterial cells are placed in a medium providing all of the nutrients necessary for growth (growth medium), the population increases according to a pattern identified as the bacterial growth curve (Figure 4). The four stages are the lag phase, log phase, stationary phase, and decline phase.

After their introduction into a medium, bacterial cells do not immediately reproduce according to their characteristic generation time; instead, the population remains constant for a period longer than the generation time. During this period individual cells actively metabolize (synthesize new cytoplasmic material) and increase in size. After the cells have undergone a rigorous physiological adjustment to the new medium, they divide.

The first cell division initiates the log phase, during which repeated division occurs at a rate consistent with the generation time. Theoretically, this stage describes a logarithmic progression, the population doubling with each new generation of cells. In theory, also, all cells grow at the same rate and reproduce at the same time. This synchronous growth period eventually terminates as cells phase out of the regular reproduction pattern, and some actually die.

Under optimum conditions, in terms of the nutrients available and the physical environment, the maximum viable population is attained at the end of the log phase, with some vigorous bacterial species achieving a density of 10,000,000,000 cells per millilitre.

For a period of time, the length of which depends on the species, the population is stationary. Among the factors responsible for this levelling off of the population are the following, which generally occur in some combination: production of inhibitory substances, depletion of nutrients, and death of cells.

The stationary phase terminates as the death rate of the population exceeds that of formation of new cells. The population steadily decreases until, eventually, all of the cells die.

Ecology. *Bacteria in water.* Good quality drinking water contains very few bacteria per millilitre and no coliform bacteria. The coliform bacteria, including *Escherichia coli* and *Aerobacter aerogenes*, are found in the intestinal tract of man and other animals. Their presence in water indicates that the water has been polluted with fecal material and hence may contain

The coliform bacteria

From Jack Maniloff, "Electron Microscopy of Small Cells: *Mycoplasma hominis*," *Journal of Bacteriology* (December 1969)

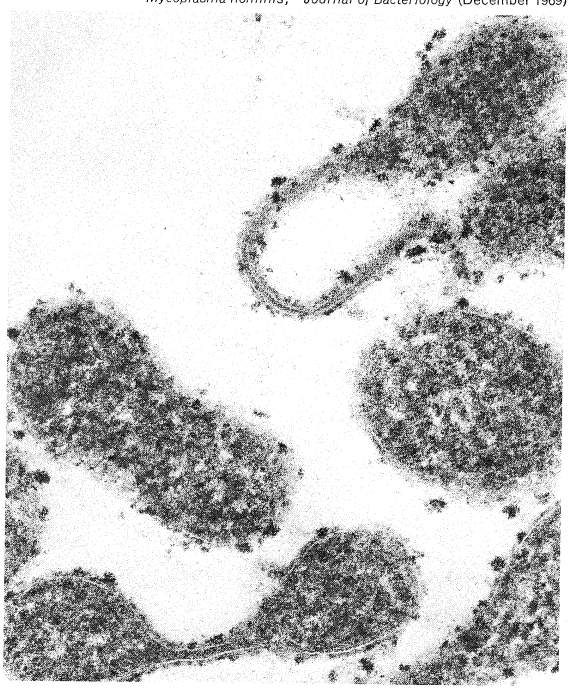


Figure 3: *An electron micrograph of cells of Mycoplasma hominis in several stages of reproduction.*
(Centre left) The elongated cell about to undergo binary fission. (Bottom) The dividing cell connected by a tubule, which, at its thinnest, consists only of the two membranes joined back to back. (Centre right) The coccoid-shaped daughter cell (greatly magnified).

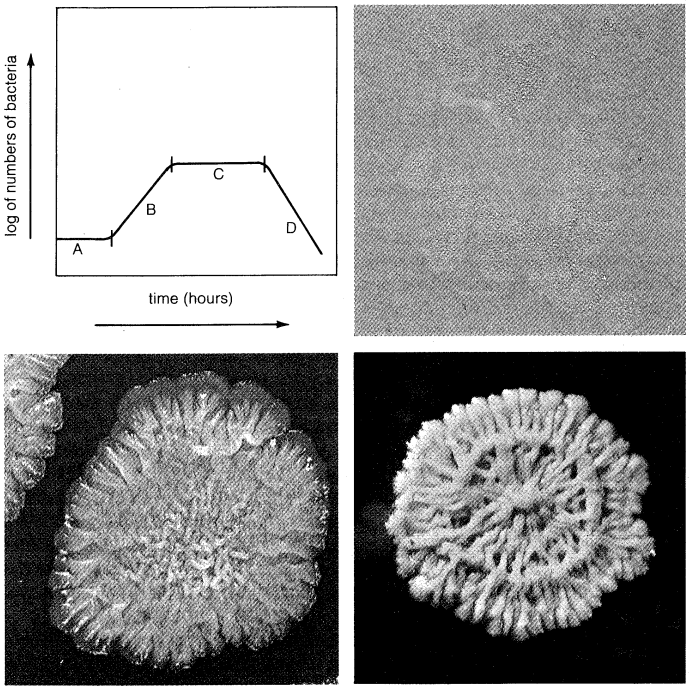


Figure 4: *Bacterial growth.*
(Top left) Generalized bacterial growth curve showing: (A) lag phase, (B) log phase (period of logarithmic or exponential growth), (C) stationary phase, and (D) death, or decline, phase. Sequence of bacterial colony growth in *Bacillus subtilis* grown at 37° C at: (top right) 18–24 hours (magnified about 6 ×), (bottom left) 48 hours (magnified about 9 ×), (bottom right) 96 hours (magnified about 9 ×).
(Top right, bottom left, bottom right) A.W. Rakosy—EB Inc.

pathogens. The usual procedures employed in municipal water-purification plants—settling, filtration, and chlorination—are designed to remove or destroy these and other micro-organisms.

Sewage, defined as the used water supply of a community, contains wastes from domestic and industrial sources. Bacteria in sewage are of importance for two reasons. First, as pollutants: human excrement in sewage may contain pathogenic bacteria, which, if not removed or killed, may enter domestic water sources or supplies (Figure 5). Second, as cleansers: the treatment of sewage, particularly the breakdown (dissimilation) of organic material (*e.g.*, proteins, carbohydrates, and fats) requires the activity of bacteria. Sewage treatment facilities, including residential septic tanks, municipal sludge digesters, activated sludge digesters, and trickling filter and sand filter processes, are all designed to utilize bacteria to break down organic matter in sewage.

The breakdown of organic matter, however, imposes a biochemical oxygen demand (BOD) on the environment into which sewage is dumped (*e.g.*, a body of water). The greater the amount of organic matter, the greater is the amount of oxygen required for its oxidation and, hence, the greater the BOD. This process can be very disruptive to aquatic life in natural streams and lakes. One of the objectives in sewage treatment is to oxidize organic matter as completely as possible and thereby reduce the BOD prior to the discharge of the sewage (effluent) into natural bodies of water. Sewage digestion tanks and aeration devices are designed to exploit the metabolic capacity of bacteria to accomplish this objective (see SEWAGE SYSTEMS).

The microbial population of the sea consists of bacteria, algae, protozoans, and fungi. Bacteria of all physiological and metabolic types inhabit the various regions, extending from the surface layer of the sea to the bottom mud. They are responsible for transformations of both organic and inorganic compounds that serve as nutrients for marine life. The dissimilation of organic compounds, under aerobic (oxygenated) conditions, yields ammonia, carbon dioxide, and sulfate and phosphate salts. These products serve as the nutrients for algae and other planktonic life, which, in turn, synthesize organic compounds that may eventually serve as food for mollusks and fishes.

Bacteria in air. Air contains bacteria and other microorganisms that are suspended and circulated for varying periods of time, depending upon atmospheric conditions and the size of the particles that carry the micro-organisms.

Extremes of bacterial concentration

Generalizations about the microbial component of air can be made only with reference to a particular environment and the circumstances that prevail at a given time: for example, a hospital ward during bed-making time (agitation of bed linens and movements of personnel stir dust that may bear large numbers of bacteria into the air); a city street following a heavy rain (the air is washed and relatively free from bacteria). Air at an altitude of 10,000 feet (3,000 metres) usually has relatively few dust particles and, therefore, considerably fewer bacteria than are common in air at lower altitudes.

Tremendous numbers of bacteria are ejected into the atmosphere by a sneeze or cough. They remain suspended in air on particles referred to as droplet nuclei, which may consist of a single bacterium—in which case the particle remains airborne for a long period of time. When the particles consist of aggregates of bacteria coated with mucus or affixed to other cells, they settle out on surfaces in a short time. Airborne bacteria from the respiratory tract of man are potentially hazardous; they include *Mycobacterium tuberculosis*, which causes tuberculosis; *Neisseria meningitidis*, which causes meningitis; *Streptococcus pyogenes*, which causes strep infections; and *Diplococcus pneumoniae*, which causes pneumonia.

A.W. Rakosy—EB Inc.

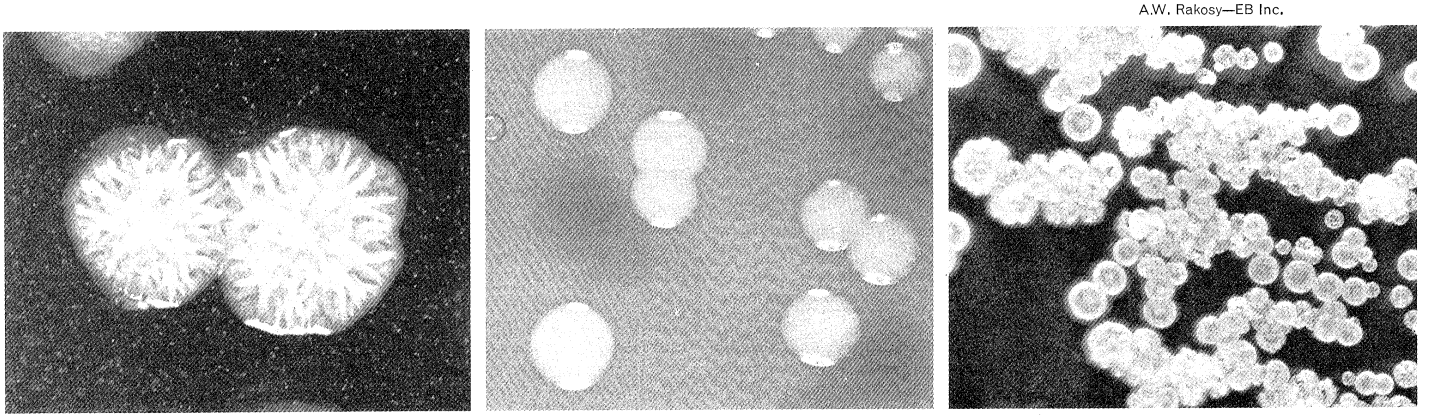


Figure 5: *Colonies of bacteria from water, dust, and soil.*
(Left) Rough colonies of *Pseudomonas pseudomallei* isolated from surface waters in Southeast Asia. These bacteria cause systemic meliodosis, a lethal infection in man. (Centre) Smooth colonies of *Micrococcus luteus*, on blood agar, isolated from dust particles from a clean room. (Right) Rough colonies of *Streptomyces griseus* isolated from soil (magnified 13 ×).

Bacteria occur on plant surfaces and will grow if conditions are favourable. They also invade and infect plant tissues, from which they are dispersed periodically into the atmosphere. Soil, particularly rich garden soil, contains countless bacteria per gram; soil dust dispersed into the air contributes a multitude of bacteria and resistant bodies formed by them (spores).

Airborne bacteria, as already mentioned, are important in the dissemination of diseases of man, of other animals, and of plants. They are also important as the source of contamination of many materials—*e.g.*, pharmaceutical preparations, surgical devices, and foods—thus necessitating precautionary procedures when "bacteria-free" conditions are desired (see below *Control*).

Bacteria in soil. Soil bacteria are extremely active in effecting biochemical changes in soil, which is the repository of the remains of plant and animal life; through microbial attack, materials are eventually transformed into the very substances that characterize soil—humus and minerals.

It is customary to view the chemical changes performed by bacteria in the soil as cyclic processes of the various elements; *e.g.*, the nitrogen cycle, the carbon cycle, and the sulfur cycle. At some stage of each cycle the element exists in its elemental form; *i.e.*, uncombined with any other elements. Certain species of bacteria are capable of converting the element into an inorganic compound that can be utilized as a plant nutrient and thereby transformed into an organic compound. The plant is consumed by animals, and the organic compound is incorporated into animal tissue. Eventually, animal and plant tissues return to the soil, where bacteria decompose them and thereby again release the elements as well as various other products (see BIOSPHERE).

Nitrogen fixers

The nitrogen cycle serves to illustrate the role of bacteria in performing various chemical changes. Nitrogen fixation, one of the key transformations in the nitrogen cycle, is performed by bacteria of several physiological types. Nitrogen-fixing bacteria are capable of transforming raw nitrogen into a form suitable for use by plants. They live in intimate association with a leguminous plant, in which case they are designated symbiotic nitrogen fixers; this type lives within lumps (nodules) on the plant's root system. Species of bacteria that can fix nitrogen in soil that is free from the plant-root system are referred to as nonsymbiotic nitrogen fixers. The best known genera are *Azotobacter* and *Clostridium*, although many others are now known to be capable of nonsymbiotic nitrogen fixation.

Bacteria in food. Milk drawn from a healthy cow contains relatively few bacteria per millilitre, but it is not sterile (free of living bacteria). Furthermore, procedures for handling the milk may add additional bacteria through contamination. Because milk is an excellent medium for the growth of bacteria, an initial small population (inoculum) can increase rapidly if the milk is not properly processed. Bacteria may merely cause spoilage or may present a serious health hazard if they are pathogenic. Bacterial pathogens transmitted through milk may originate in the cow or in man. A cow infected with the tubercle bacillus may transmit this organism, through milk, to man. Brucellosis, or undulant fever, also can be transmitted from cows to man via milk. Alternatively, an infected milk handler can contaminate the milk; outbreaks of typhoid fever, scarlet fever, and diphtheria have been known to occur in this fashion. Proper treatment of milk by the process of pasteurization —either the low-temperature, holding method (145° F [63° C] for 30 minutes) or the high-temperature, short-time method (161° F [72° C] for 15 seconds)—destroys all pathogens.

Bacterial conversion of milk

On the credit side, however, selected species and strains of bacteria convert milk and casein (milk protein) into such desirable products as buttermilk, yogurt, and cheese. Commercial cultured buttermilk is prepared from skim milk that has been inoculated with a starter culture and allowed to incubate until the desired changes occur. The starter culture consists of *Streptococcus lactis* or *Streptococcus cremoris*, together with *Leuconostoc citrovorum* or *Leuconostoc dextranicum*. Yogurt and other fermented milk products are produced in a similar manner but through the activities of different selected cultures of bacteria.

The formation of cheeses is likewise dependent upon the activity of micro-organisms. The curd from which cheese is made is precipitated (made to settle out) from milk by an acid-producing bacterium, such as *Streptococcus lactis*. Following removal of moisture and the addition of salt, the curd is allowed to ripen by the action of selected bacteria. Lactobacilli, streptococci, and propionibacteria are important for the ripening of Swiss cheese; *Brevibacterium linens* is responsible for the flavour of Limburger cheese; and molds (*Penicillium* species) are used in the manufacture of Roquefort and Camembert cheeses.

Bacteria in nondairy foods are as significant as those in milk and dairy products. The variety of bacteria that contaminate foods and the diversity of foods can result in a wide array of types of food spoilage. When allowed to grow in food, certain bacteria can cause food poisoning; they secrete a toxin that, when ingested by humans, can cause either a severe gastrointestinal upset—as in *Staphylococcus aureus* food poisoning—or death—as in botulism (caused by the toxin of *Clostridium botulinum*). It follows that one of the major concerns of food microbiology is the development and assessment of techniques to preserve foods from spoilage and contamination.

Food may be the carrier of pathogenic bacteria and thus be responsible for food-borne infections, among which the more frequently occurring are typhoid fever (*Salmonella typhosa*); salmonellosis (*Salmonella* species other than *S. typhosa*); and shigellosis, or dysentery (*Shigella dysenteriae*).

Notwithstanding the detrimental effects of food contamination, other bacterial populations are responsible for a variety of special foods, produced through bacterial fermentation; these include pickles and other pickled products, sauerkraut, and olives.

Control. Many materials and products, as well as certain environmental areas, require either the reduction or destruction of microbial populations. In a hospital operating room, for example, procedures are observed that reduce the microbial contents of the air to a very low level; on the other hand, the glucose solution used for intravenous injection is processed to be sterile—absolutely free of any form of life.

Degrees of control

Physical and chemical means are available to accomplish sterilization. The method of choice depends upon several considerations, not the least of which is the effect of the sterilizing procedure on the object being sterilized. If the material being sterilized is to be discarded following sterilization, such as used bacteriological media from a laboratory, then there is no problem except effectiveness of the sterilization procedure; however, if the material to be sterilized is a vitamin solution, for example, the procedure must be effective enough to sterilize without affecting the quality of the vitamin product.

Heat. High temperature, applied in a variety of ways, is one of the most effective sterilizing agents. Heat may be applied in the form of incineration, steam under pressure (autoclave), or dry heat (hot-air oven).

Incineration procedures range from the passing of an object through a bunsen burner flame in the laboratory to the burning of infected animal carcasses or contaminated bedding in large furnaces.

Steam under pressure, which is the principle of operation of the autoclave (an elaborate pressure cooker), is perhaps the most widely used sterilization procedure. The autoclave is a standard item of equipment in laboratories, hospitals, industries involved in food processing, and enterprises concerned with sterilization procedures and the manufacture of sterile products.

Dry heat, as in hot-air ovens, also accomplishes sterilization, but, in contrast to steam heat, higher temperatures and longer exposure times are required. Equipment that can be sterilized in an autoclave at 121° C (250° F) within ten to 15 minutes requires 160° C (320° F) for a period of two hours in a hot-air oven. Hot-air steriliza-

tion is used for materials that might be adversely affected by moist heat or that should not be directly exposed to moist heat because of necessary packaging.

The heat afforded by boiling water and pasteurization processes markedly reduces the bacterial flora but does not truly sterilize. Pasteurization, as previously mentioned, is designed to kill only the serious pathogens that might occur in milk; some bacteria and, in particular, bacterial spores are not killed. Similarly, boiling water kills the vegetative (active) bacterial cells but not necessarily the spores.

Radiation. Ultraviolet radiation (in the 2650-angstrom region [one angstrom = 10^{-8} centimetre]) is extremely bactericidal; that is, capable of killing bacteria. When properly used—namely, under conditions that allow direct exposure of organisms to the radiation—the microbial population can be effectively reduced. Ultraviolet rays, however, have a very low order of penetration; a thin film of glass filters out most of the rays.

Antibacterial action of gamma rays

Gamma radiations, which are emitted from radioactive isotopes, have great penetrating power as well as a high lethal effect. This combination of characteristics—high penetration and high bactericidal activity—makes them extremely effective as sterilizing agents. But several technical problems are associated with practical applications of gamma radiations: the development of an adequate supply of radiation sources and the availability of equipment designed to guarantee safety to its operators.

Certain materials cannot be exposed to physical agents without being adversely altered in some manner. Many medicines, including antibiotics, and other biological solutions may be destroyed or inactivated by any of the methods described so far. In such cases filtration is the appropriate method for sterilization. A wide variety of filters are available that are porous enough to allow fluids to pass through but not micro-organisms.

Chemicals. Sterilization can also be accomplished with such chemicals as the gas ethylene oxide. With the appropriate concentration, humidity, and temperature, ethylene oxide is a powerful sterilizing agent. In addition, it has the ability to penetrate considerably through materials. Ethylene oxide is widely used for the sterilization of many materials, including plastic devices, that could not undergo the other procedures.

Many chemicals for the control of micro-organisms are available for application to environmental surfaces or materials; they are not intended to effect sterilization, but they do reduce the microbial population or eliminate certain types of micro-organisms. These chemical agents are variously termed antiseptics, germicides, disinfectants, and sanitizers on the basis of their action (see ANTISEPTIC AND GERMICIDE).

FORM AND FUNCTION

In normal environments bacteria exist not only in large numbers but also in great diversity. In order to understand the characteristics of the individual species comprising these mixed populations, it is necessary to study each species as a pure culture, by isolating bacterial cells from colonies on a specific nutrient referred to as the medium. This isolation (pure culture) from the colony can be maintained as a pure culture by periodic transfer to fresh medium. Alternatively, a pure culture can be lyophilized (dried while frozen and sealed under vacuum) and kept viable (capable of dividing) in this condition for many years. Characterization of a pure culture requires a study of cell morphology, cultural and physiological characteristics, biochemical (or metabolic) processes, antigenic characteristics, and pathogenicity.

Morphological features. *The bacterial cell.* The typical cell of a species of Eubacteriales, or "true" bacteria, is a bacillus approximately one micron in diameter and a few microns in length. Coccoid cells may occur in characteristic arrangements: *e.g., Diplococcus pneumoniae* occurs in pairs; *Staphylococcus aureus* occurs in grapelike clusters; *Streptococcus pyogenes* occurs in chains; *Sarcina lutea* occurs as a cuboidal arrangement of cells; and *Micrococcus tetragenus*, as the name suggests, occurs in tetrads, or groups of four cells.

In contrast to the true bacteria are the "higher" bacteria, whose appearance suggests some primitive or abortive attempt toward cellular differentiation. They are usually much larger than the true bacteria and often bear some resemblance to yeasts, molds, algae, or protozoans.

The typical bacterium

A typical bacterial cell is shown in Figure 6; not all bacterial species possess all of the structures shown. All bacterial cells contain nuclear substance, but it is not organized into a discrete nuclear structure as in higher organisms; there is no nuclear membrane, and mitotic division, the mechanism by which the cells of plants and animals divide, does not occur. The blue-green algae are akin to bacteria in this respect; together they are characterized as procaryotes, in contrast to eucaryotes, the cells of which possess a discrete nuclear structure.

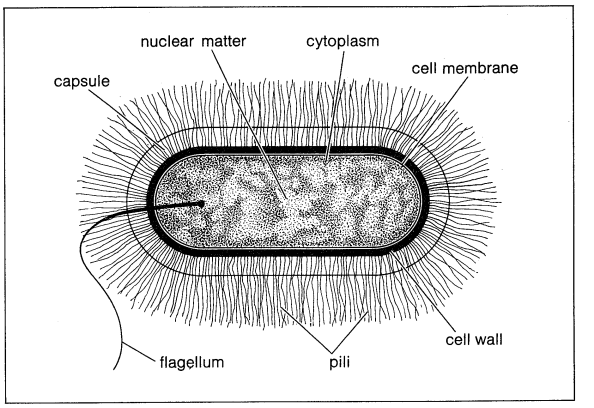


Figure 6: Schematic drawing of structure of a typical bacterial cell of the bacillus type.

All bacterial cells possess a membrane and a cell wall. There are distinct differences among species in terms of the chemical composition of these structures, however. The cells of some bacterial species possess whiplike structures called flagella, the number and arrangement of which are typical for a species. Shorter, rigid appearing, spikelike projections known as pili, or fimbriae, appear on some cells. Certain cells are surrounded with a gelatinous or slimy material, the capsule (Figure 7). Many bacteria can assume a dormant state as a spore; in fact, during sporulation the entire vegetative cell, in essence, is transformed into the spore body.

From W.H. Taylor and E. Juni, "Pathways for Biosynthesis of a Bacterial Capsular Polysaccharide," *Journal of Bacteriology* (May 1961)

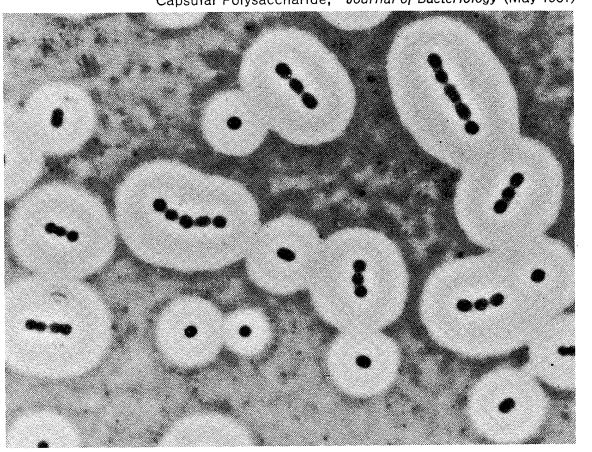


Figure 7: Capsular material surrounding these bacteria (*Acinetobacter calcoaceticus*) is revealed in a suspension of India ink and viewed through the light microscope (magnified about 2500 ×).

Staining techniques are used to demonstrate the various bacterial cell structures. One of the most important staining procedures involves the gram stain (named after its inventor, H.C.J. Gram, a Danish physician). This is a differential stain; *i.e.*, bacteria are said to be gram positive or gram negative depending upon whether they retain the

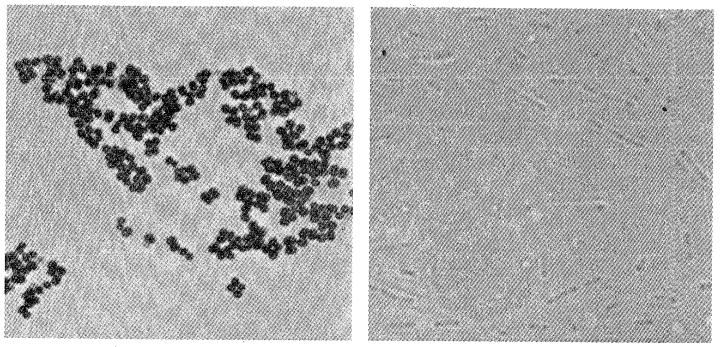


Figure 8: *Bacteria isolated and coloured with gram stain.*
(Left) Gram positive cocci, *Staphylococcus aureus*, from a laboratory culture. (Right) Gram negative bacilli with a capsule, *Klebsiella pneumoniae*, from a pneumonia lung abscess (magnified 1000 ×).
A.W. Rakosy—EB Inc.

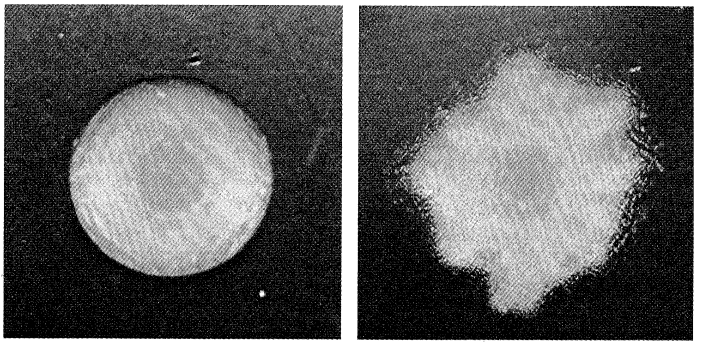


Figure 9: *Colonial morphology.*
Citrobacter freundii displaying two forms: (left) smooth colony, (right) rough colony.
Rhodes Scherer, National Animal Disease Laboratory, Ames, Iowa

purple colour of the original stain (crystal violet) at the end of the procedure, or whether it is washed out and the red colour of the counterstain (safranin) shows (Figure 8).

More refined characterization of the anatomy of a bacterial cell can be accomplished by electron microscopy, in which bacterial cells are sliced into very thin sections and then viewed under very high magnification; electron micrographs reveal a great deal of complex detail: layers in the cell surface, internal structures, as well as connections or continuity between certain structures.

The bacterial colony. The gross appearance of bacterial growth in or on media defines the cultural characteristics of a species. When a specimen containing bacteria is inoculated onto an agar medium (a standard preparation of nutrients with a gel-like consistency), colonies develop, each from an isolated bacterial cell. Such colonies (Figure 9) vary in characteristics depending upon the species. They may be pinpoint in size or several millimetres across; flat or raised and convex; smooth or broken edged; stringy, brittle, or buttery in consistency; coloured internally or excreting substances that colour the surrounding medium; and opaque, transparent, or translucent.

Bacteria cultivated as agar slant cultures (slanted medium in test tubes) exhibit differences in characteristics much like those described for colonial appearance (Figure 10). In liquid media (broth), growth may be confined to the surface as a film (pellicle); uniformly distributed throughout the liquid; or particulate, with a tendency to form a sediment.

Physiological features. *Physical requirements.* On the basis of temperature requirements for growth, bacteria are grouped into three categories: psychrophiles, 0° to 30° C (32° to 86° F); mesophiles, 15° to 45° C (59° to 113° F), with best growth in the range 25° to 40° C (77° to 104° F); and thermophiles, 45° to 60° C (113° to 140° F) and above. Bacteria within each group exhibit specific minimum, maximum, and optimum temperatures for growth; for example, *Pseudomonas delphinii* grows at 1° C and 30° C (34° F and 86° F), with optimum growth occurring at 25° C (77° F), and *Bacillus thermoliquefaciens* grows at 37° C and 70° C (99° F and 158° F), with optimum growth occurring at 60° C (140° F).

Temperature categories of bacteria

Atmospheric oxygen is required by some, but not all, bacteria; others are inhibited by its presence. Bacteria are classified as aerobes when they require oxygen to grow and as anaerobes when they cannot grow in the presence of oxygen; facultative anaerobes do not require oxygen and can grow in its presence.

Some bacteria grow in a wide range of salt concentrations; others, such as marine bacteria, require salt levels of 10 to 15 percent for optimal growth. Most bacteria grow best in an environment near neutrality (neither acidic nor alkaline); some, however, grow under strongly acid conditions (*Thiobacillus thiooxidans*) and others under strongly alkaline conditions (*Nitrobacter* species). Most bacteria, and especially those intimately associated

A.W. Rakosy—EB Inc.

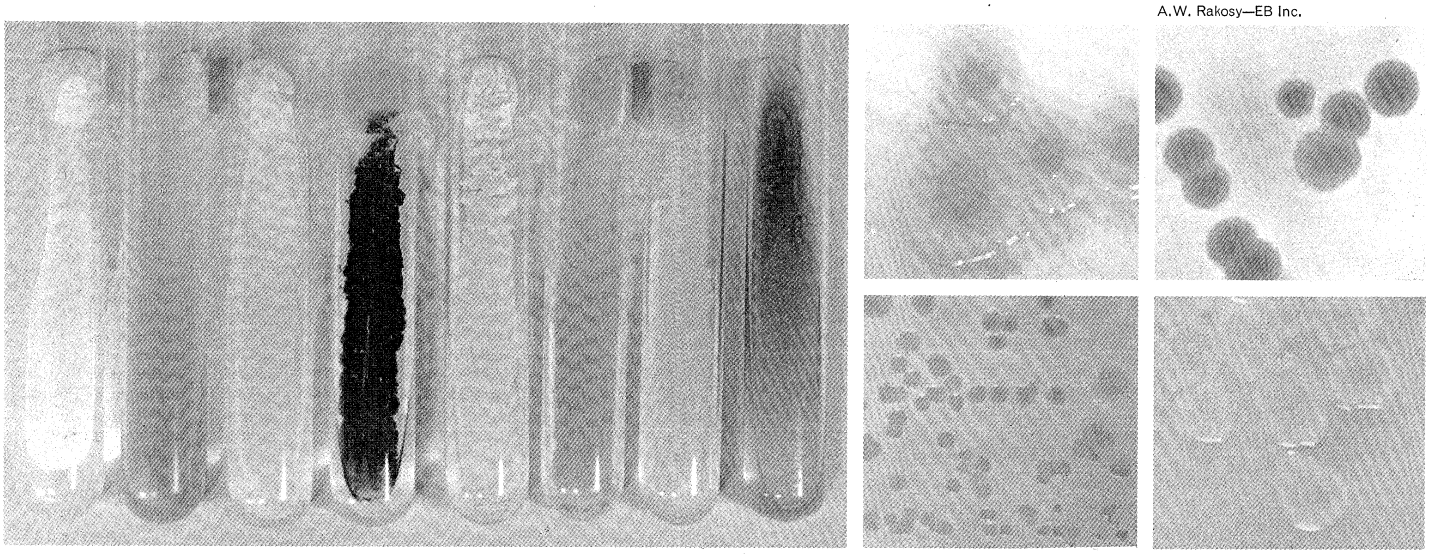


Figure 10: *Chromogenic bacteria.*
(Left) Agar slant cultures, left to right: *Sarcina lutea; Pseudomonas aeruginosa*, from water in a hospital humidifier; *Pseudomonas fluorescens; Chromobacterium violaceum; Sarcina aurantiaca; Serratia marcescens*, from soil; *Staphylococcus albus; Pseudomonas aeruginosa* from the urine of a person with cystitis. Agar colonies: (top centre) *Pseudomonas aeruginosa* isolated from sputum, (top right) *Chromobacterium violaceum*, from soil, (bottom centre) *Serratia marcescens*, from inadequately cleaned eating utensil, (bottom right) *Pseudomonas aeruginosa* from urine.

with man, grow best without light. Photosynthetic bacteria, however, require light.

Nutritional requirements. The establishment of the specific nutritional requirements for various bacterial species required the development of media consisting of known chemical compounds in prescribed amounts, on which the bacteria could be cultivated in the laboratory. In a broad sense, bacteria, on the basis of nutritional requirements, can be divided into two groups: autotrophs and heterotrophs. Autotrophs are capable of growing entirely on inorganic ingredients, with atmospheric carbon dioxide as the exclusive source of carbon. Heterotrophs require an organic form of carbon and, in addition, may require other organic substances, such as amino acids and vitamins.

The autotrophic sulfur-oxidizing bacteria can grow on powdered sulfur, ammonium sulfate, potassium phosphate, calcium chloride, magnesium sulfate, ferric sulfate, water, and carbon dioxide. From these relatively simple chemical substances the autotrophic bacterium can synthesize and organize the vast array of complex organic substances of which it is made.

Heterotrophs exhibit great diversity in their nutritional requirements. *Escherichia coli*, the most frequently studied heterotrophic species, can grow on the ingredients listed above for autotrophic sulfur-oxidizing bacteria, except that carbon from an organic source (such as the sugar glucose) is also essential. *Salmonella typhosa* has a requirement slightly different from that of *E. coli* in that not only glucose but also an amino acid (tryptophan) is required. *Staphylococcus aureus*, another typical heterotroph, requires several amino acids and at least one vitamin, thiamin. *Lactobacillus* species are considerably more fastidious; they require many amino acids, several vitamins, and purine and pyrimidine compounds.

Some bacteria, especially certain pathogens, such as the syphilis bacterium (*Treponema pallidum*), require a complex medium that is provided only by living animal tissue. Such bacteria have so far not been cultivated even in a complex bacteriological medium of peptone (a digest of animal protein) and meat extract (or meat broth), with supplemental substances such as blood, serum, and other animal fluids.

Biochemical activities. Bacteria are capable of carrying out many chemical changes, both in the breakdown of complex substances and in the synthesis of new products. A compound may be degraded to different end products by different bacterial species, just as the same nutrients may be synthesized into different substances by different species.

Decomposition of carbohydrates

Dissimilation. Complex organic carbon compounds, such as pectin, cellulose, and starch, are readily degraded by many bacteria. Pectins are complex carbohydrates that occur in plants. Enzymes produced by species of *Erwinia*, *Bacillus*, and *Clostridium* are capable of converting pectin to galacturonic acid and further to sugar and other products. Cellulose, the major constituent of plant tissue, is transformed to dextrins and then to glucose by species of *Cellulomonas*, *Cytophaga*, *Streptomyces*, and *Clostridium*. Several *Bacillus* and *Clostridium* species produce enzymes that convert starch to sugar.

After the complex carbohydrates have been broken down into their constituent units (*e.g.*, glucose), these units are utilized in a variety of ways—*e.g.*, to produce a variety of other compounds. Bacterial species are characterized in part on the basis of the characteristic products they form from glucose. Table 2 lists several genera of bacteria, together with representative products of their glucose dissimilation.

The process of protein dissimilation (proteolysis) is accomplished by enzymes known as proteinases, which are produced by species of such genera as *Clostridium*, *Bacillus*, *Proteus*, and *Pseudomonas*. The proteinases decompose the protein molecules, hydrolyzing (breakdown involving water) the linkages (peptide bonds) between the amino acids that constitute the molecule. The result is the formation of peptides, chains of amino acids. Enzymes called peptidases then break down the peptides into individual amino acids. In turn, the amino acids may be degraded through several processes; deamination, the removal of the amino group ($-NH_2$) by enzymes called deaminases, the end products being ammonia (NH_3) and a fatty acid; decarboxylation, the removal of the carboxyl group ($-COOH$) by enzymes called decarboxylases, the end products being carbon dioxide and an amine; or amino acid fermentation, a breakdown more extensive than either deamination or decarboxylation, and resulting in more complex and varied end products, depending on the amino acid being degraded.

Table 2: Bacteria Grouped According to Major Products of Glucose Dissimilation

groups (with examples of some genera)	representative products
Lactic acid bacteria	
Streptococcus *Lactobacillus* *Leuconostoc*	lactic acid only or lactic acid plus acetic acid, formic acid, and ethyl alcohol; those species producing only lactic acid are *homofermentative;* those producing lactic acid plus other compounds are *heterofermentative*
Propionic acid bacteria	
Propionibacterium *Veillonella*	propionic acid plus acetic acid and carbon dioxide
Coli–aerogenes–typhoid bacteria	
Escherichia *Aerobacter* *Salmonella*	formic acid, acetic acid, lactic acid, succinic acid, ethyl alcohol, carbon dioxide, hydrogen, 2,3-butylene glycol (produced in various combinations and amounts depending on genus and species)
Acetone, butyl-alcohol bacteria	
Clostridium *Butyribacterium* *Bacillus*	butyric acid, butyl alcohol, acetone, isopropyl alcohol, acetic acid, formic acid, ethyl alcohol, hydrogen, and carbon dioxide (produced in various combinations and amounts depending on species)
Acetic acid bacteria	
Acetobacter	acetic acid, gluconic acid, kojic acid

Source: M.J. Pelczar, Jr., and R.D. Reid, *Microbiology* (1972).

Fats, or lipids, composed of fatty acids and glycerol (*i.e.*, triglycerides), can be degraded by such bacteria as *Pseudomonas*, *Proteus*, *Achromobacter*, *Alcaligenes*, *Bacillus*, *Micrococcus*, and *Clostridium*. The end products of the dissimilation of fats are glycerol and fatty acids, which can be broken down further.

Synthesis. The extensive capacity of a bacterium to synthesize complex molecules is shown when the simple chemical substances upon which an autotroph grows are compared to the complex chemical composition of the bacterial cells that are produced. In addition to the large, complex molecules that constitute cell structure, other substances of significance are elaborated by bacterial cells, including antibiotics, pigments, toxins, and polysaccharides (complex carbohydrates).

Pathogenicity. It was implied above that a pathogen has the potential to produce a disease in a given host species—a plant, an animal, or even another bacterium species. The power of a bacterium to cause a disease, termed virulence, varies within a species. Some strains are highly virulent; *i.e.*, a small number of cells can establish the infection. Other strains have a low degree of virulence and produce infection only when transmitted in massive numbers. Furthermore, some strains of a pathogen may lose their virulence. A quantitative expression of virulence can be determined by performing laboratory experiments to establish the number of bacterial cells required to infect or kill a standard experimental animal. The expression lethal $dose_{50}$ (LD_{50}) refers to the number of bacteria necessary to kill 50 percent of the organisms inoculated.

Virulence

Factors that contribute to the virulence of a bacterial pathogen are not yet completely understood, except in those instances in which the pathogen causes damage by secreting a potent toxin (an exotoxin). Diphtheria bacteria (*Corynebacterium diphtheriae*), for example, establish themselves in the mucous membrane of the upper respiratory tract of man; as they grow, they produce an exotoxin that is absorbed by the mucous membranes and causes the death of cells in the host. It is the exotoxin that does the damage to the host; a strain of *C. diph-*

theriae that does not produce the exotoxin cannot cause diphtheria.

Not all virulent pathogens produce an exotoxin; most, in fact, do not. *Salmonella typhosa*, the typhoid fever bacterium, produces an endotoxin, a complex substance associated with, or bound to, the surface structures of the bacterial cell. Endotoxins damage host tissues and host metabolism in ways not yet clearly understood. Other substances that contribute to the virulence of various pathogens are described in Table 3.

Table 3: Substances Contributing to the Virulence of Pathogenic Bacteria

substance	action
Hyaluronidase	increases permeability of tissue spaces to bacterial cells
Coagulase	increases resistance of bacteria to phagocytosis (engulfment by defense cells, or phagocytes)
Hemolysins	destroy red blood cells
Collagenase	dissolves collagen, a connective tissue protein
Leucocidin	kills white blood cells (specifically leucocytes) and hence decreases phagocytic action
Exotoxins and endotoxins	interfere with normal metabolic processes

Antigenic features. When bacterial cells enter the tissues of an animal, it is likely that they will act as antigens, agents that evoke the production of substances called antibodies. Antibodies are produced by an animal as part of its immunological defense against any foreign substance that can threaten its welfare.

Micro-organisms can be characterized in detail by performing antigenic analyses on them as well as on products (*e.g.*, toxins) of their metabolism.

Antigen–antibody reactions are highly specific and highly sensitive. A bacterial cell consists of many different antigens (*e.g.*, flagellar, capsular, and cell wall antigens). The antigen pattern in one bacterial species differs from that in another. Although it is not uncommon for related species (*e.g.*, *Salmonella typhosa* and *Salmonella paratyphi*) to share certain antigens, each also contains other antigens that are unique.

Variability. The characteristics of one bacterial species are sufficiently definitive and constant to delineate it from other species. This distinctiveness does not mean, however, that each characteristic is evident and manifest by all strains of all species under all conditions.

Environmentally induced variation

Reversible changes. Bacteria, under uniform conditions, will manifest a constancy of characteristics. If the same species of bacteria is placed in different environments—different physical conditions or different chemical composition of medium—the resultant growth may differ. In fact, the morphological and physiological characteristics are not identical at all stages of the growth curve of a bacterium. The morphology of cells from an "old" culture differs from that of "young" cells. In addition, the formation of a capsule is significantly influenced by the composition of the medium; bacteria that, in a nutrient broth, exhibit no capsules may produce large capsules when grown in milk. Some of the enzymes produced by bacteria are produced only when the compound on which the enzyme acts (substrate) is present; they are called adaptive enzymes, in contrast to constitutive enzymes, which are produced irrespective of the substrate.

Changes of the types just described are transient; they reflect what occurs during a stage of growth or in response to a change in the environment. The genetic endowment (genotype) of the organism remains the same, regardless of the different expressions (phenotype) that are seen under different environmental circumstances.

Permanent changes. Daughter cells that contain genetic information different from that of the parent cell constitute a new genotype. Such permanent-type genotypic changes may occur through four different processes: mutation, conjugation, transformation, and transduction.

Mutation involves a sudden alteration of a gene that is inherited by subsequent generations. Some bacterial cells in the process of normal growth undergo mutation; however, the ratio of mutant cells to unchanged cells is very small. The number of mutants in a population can be greatly increased by exposing the bacteria to certain physical agents (*e.g.*, ultraviolet rays, X-rays) or to chemicals (*e.g.*, mustard gas and organic peroxides).

Many different kinds of bacterial mutants have been isolated and characterized; they exhibit alterations in nutrition, drug resistance, pigmentation, and colonial form, among other characteristics.

Conjugation (sexual reproduction), a rarity in bacteria, results in recombinations of genes in the cells that pair (see above *Reproduction*).

Transformation also involves the transfer of genetic information from one cell to another, but pairing does not occur. Deoxyribonucleic acid (DNA) that is released from one cell (the donor) is taken up by another cell (the recipient) and incorporated into the genetic apparatus of the latter.

Changes caused by viruses

Transduction involves the transfer of genetic substance from one cell to another via a bacteriophage, a virus that infects bacteria. The bacteriophages produced in a host cell are released when the cell is destroyed by the infection. Some of the virus particles may be contaminated—*i.e.*, carry fragments of DNA from the host bacterium. When such a contaminated virus particle penetrates another bacterial cell, the bit of DNA from the original host is carried with it and, under certain circumstances, may become incorporated into the DNA of the recipient cell, thus changing the genetic constitution of the latter.

EVOLUTION AND CLASSIFICATION

Origin and relationships. Bacteria have existed from very early periods in the history of life on Earth. They have been detected as fossils in rocks dating from at least Devonian times (as early as 395,000,000 years ago). On the basis of their indistinct nuclear matter, bacteria are assumed to be closely related to the blue-green algae. It has been speculated that a photosynthetic ancestor may have given rise to both the bacteria and the blue-green algae.

Several groups of bacteria show features that suggest relationships to other classes of organisms: these features do not necessarily indicate a close relationship, however. Actinomycetes form branching units and reproductive stages similar to the fungi. Mycoplasmas (formerly called PPLO and L forms) are variable organisms that seem to lack in organization. Myxobacteria resemble, in some ways, the slime molds.

Rickettsias, rodlike or oval unicellular animal parasites, are smaller than bacteria but have a similar internal structure. Speculation is that they may represent highly modified small bacteria that became parasitic in the past and now can live no other way.

Classification. *Distinguishing taxonomic features.* Shape and size of the bacterial cell are aids in classifying bacteria, as is the appearance of the colonies that are formed. Other characteristics, however, assume even greater significance because they are more consistent under different environmental conditions. These include the kinds of foods utilized, the products of metabolism, reactions to specific chemicals, antigenic composition, and degree of tolerance to environmental change.

The ten orders of the class are distinguished primarily on the shape and rigidity of the bacterial cell and on locomotor ability; on the capacity of individual bacteria to aggregate in chains or clusters of special shape; and on special physiological characteristics.

Annotated classification. The following classification, featured in the edition of *Bergey's Manual of Determinative Bacteriology*, is generally accepted.

CLASS SCHIZOMYCETES

Unicellular micro-organisms generally ranging from 1 to 5 microns in size; variable in shape and in nutritional needs; lacking a distinct nucleus; most species without chlorophyll; occur singly or in chains or clusters and form distinctive colonies; about 1,500 species; worldwide distribution.

Order Pseudomonadales

Rigid-walled cells of variable shape, in some species forming chains; photosynthetic pigment present in certain species; cells usually motile by means of a single flagellum. Species in soil and in freshwater and saltwater. Examples of genera: *Vibrio*

comma (cholera bacteria), *Pseudomonas, Nitrosomonas, Thiobacillus.*

Order Chlamydobacteriales
Rigid-walled cells in many-celled filaments (trichomes), frequently ensheathed; occasionally produce motile spores; trichomes often attached to a surface; species in freshwater and marine habitats. *Sphaerotilus natans* common in polluted water.

Order Hyphomicrobiales
Rigid-walled cells often attached to surface by a stalk; reproduction by budding (as in yeasts) rather than by ordinary division. Genera include *Rhodomicrobium* and *Hyphomicrobium.*

Order Eubacteriales
Rigid-walled cells, coccoid or bacilloid, sometimes in chains; motile forms move by means of laterally emergent flagella; not acid-fast (*i.e.*, retaining a bacterial dye when treated with an acidic solution); includes the largest number of genera of concern to man—*e.g., Escherichia, Diplococcus, Staphylococcus, Streptococcus, Bacillus, Lactobacillus.*

Order Caryophanales
Rigid-walled cells in trichomes; motile by means of lateral flagella; very large cells (up to 30μ long and 3μ across); occur in water and decomposing matter as well as in the intestines of arthropods and vertebrates. *Caryophanon latum*, common in cow dung, and *Simonsiella muelleri*, found in the mouths of humans and domestic animals.

Order Actinomycetales
Rigid-walled cells that may grow out in a branching system, resembling mold colonies; includes *Mycobacterium tuberculosis* (tuberculosis bacterium), *Streptomyces.*

Order Beggiatoales
Rigid-walled cells, usually large and often in trichomes that move by gliding motion as do some blue-green algae; genera include *Beggiatoa, Thiothrix.*

Order Myxobacteriales
Flexible-walled cells that creep on surfaces. Stalked fruiting bodies usually develop from a spreading colony, like slime molds. Found in soil, compost, manure, and rotting wood; genera include *Myxococcus, Chondrococcus, Sorangium.*

Order Spirochaetales
Spiral cells that swim by flexion; found in water and in the bodies of vertebrates; genera include *Borrelia, Treponema,* and *Leptospira,* all parasites of man and other animals.

Order Mycoplasmatales
Flexible-walled cells, nonmotile, highly variable in shape at different life stages; includes *Mycoplasma* and forms once known as pleuropneumonia-like organisms (PPLO) and L forms, which are apparently intermediate between true bacteria and rickettsias.

The stability of bacterial taxonomy

Critical appraisal. The classification of bacteria, relatively stable and generally accepted as given above, encompasses ten well-established orders. The question of higher categories, however, has not yet been resolved: are the bacteria very primitive plants or another kind of organism altogether? Some biologists favour a taxonomic system in which the bacteria (class Schizomycetes) are grouped with the rickettsias and viruses (class Microtatobiotes) as the division Schizomycophyta, which, in turn, is grouped with the Phylum Cyanophyta (blue-green algae) as a subkingdom, Monera, of the kingdom Protista (of equal rank with the kingdoms of Plantae and of Animalia). Such a scheme obviates the need for assigning these procaryotic organisms (*i.e.*, without distinct nuclei) to the kingdom Plantae, which is not quite suitable.

For many years the actinomycetes, myxobacteria, and mycoplasmas were listed as bacteria even though they have unusual life cycles and variable structure. It is now fairly well established, however, that they do belong in the class Schizomycetes. The rickettsias, which were occasionally grouped with the bacteria, are now considered an order, Rickettsiales, of the class Microtatobiotes.

BIBLIOGRAPHY. R.S. BREED *et al.* (eds.), *Bergey's Manual of Determinative Bacteriology*, 7th ed. (1957), a reference and sourcebook accepted as standard throughout the world for classification of bacteria and related micro-organisms; J.E. BLAIR, E.H. LENNETTE, and J.P. TRUANT (eds.), *Manual of Clinical Microbiology* (1970), a reference work describing methods and techniques for the isolation and identification of pathogenic bacteria and other disease-producing micro-organisms; C.J. CORUM (ed.), *Developments in Industrial Microbiology*, vol. 10 (1970), a documentation of the Proceedings of the 25th General Meeting of the Society for Industrial Microbiology on Low Level Microbiological Assays, Industrial Microbiology, and World Food Problems, with other contributed papers; H.W. DOELLE, *Bacterial Metabolism* (1969), details of the biochemical reactions and processes of particular micro-organisms; W.C. FRAZIER, *Food Microbiology*, 2nd ed. (1967), a textbook on the activities of micro-organisms important in foods, as well as methods for their detection and control; T.R.G. GRAY and DONALD PARKINSON (eds.), *The Ecology of Soil Bacteria* (1968), reports given at the International Symposium on Soil Bacteriology, University of Toronto; A.E. KRISS *et al.*, *Microbial Population of Oceans and Seas* (1967; orig. pub. in Russian, 1964), a synoptic picture of the distribution and range of marine micro-organisms; C.A. LAWRENCE and S.S. BLOCK (eds.), *Disinfection, Sterilization, and Preservation* (1968), a handbook on the principles and practical aspects of the control of micro-organisms by chemical and physical methods; JOEL MANDELSTAM and K. MCQUILLEN (eds.), *Biochemistry of Bacterial Growth* (1968), a description of the way in which the simple organic and inorganic constituents of the medium are transformed into bacterial cell material; M.J. PELCZAR, JR. and R.D. REID, *Microbiology*, 3rd ed. (1972), a textbook presenting the major areas of study in the field of microbiology; ALAN RHODES and D.L. FLETCHER, *Principles of Industrial Microbiology* (1966), a general survey of industrial microbiological processes; A.J. SALLE, *Fundamental Principles of Bacteriology*, 6th ed. (1967), a textbook concerned with the fundamental concepts of basic and applied microbiology; R.Y. STANIER, M. DOUDOROFF, and E.A. ADELBERG, *The Microbial World*, 3rd ed. (1970), an advanced textbook covering the major characteristics of micro-organisms—*i.e.*, morphology, physiology, and biochemistry; F.S. THATCHER and D.S. CLARK (eds.), *Microorganisms in Foods: Their Significance and Methods of Enumeration* (1968), reports of the meeting of the International Committee on Microbiological Specifications for Foods, including information on occurrence, methods of detection, and technical procedures; A.H. WALTERS and J.J. ELPHICK (eds.), *Biodeterioration of Materials* (1968), a collection of scientific papers given at the First International Biodeterioration Symposium.

(M.J.P.)

Baden-Württemberg

Theodor Heuss, the first president of the Federal Republic of Germany, called his native Baden-Württemberg "the model of German possibilities," and by the early 1970s there were several indications that the possibilities were becoming realized in this very young German state (*Land*). By that time, Baden-Württemberg ranked third in both area and population among the West German states, having grown more than any other in the period following World War II. Its total area, 13,803 square miles (35,750 square kilometres), was larger than that of Belgium or Israel; and its population, in passing the 8,800,000 mark, approximated to that of Cuba or Greece. Included in the figure were over 500,000 foreign workers from the Mediterranean region, a higher number than that for any other German *Land* and, because it reflects the need for workers, a good indication of the high rate of industrialization in the region. Other favourable economic indicators demonstrated that Baden-Württemberg had high export totals, a high proportion of population engaged in industry and trade, few strikes, and a high percentage of homeowners' and building loans. Strategically located in the southwest of the nation, the *Land* is bordered by France in the west and by Switzerland in the south—a factor that has facilitated the influx of foreign workers. Formed under post-World War II occupational rule, and confirmed by a December 1951 referendum, the *Land* consists of three former *Länder:* Württemberg-Baden (in the American zone), Südwürttemberg-Hohenzollern and Südbaden (both in the French zone). The merger took effect in 1952. The capital is at Stuttgart.

The Land. Within the 1,026-mile- (1,651-kilometre-) long border of Baden-Württemberg lies one of the most geographically varied territories of the German Federal Republic, with the forests of the upland regions alternating with fertile highlands, green meadows, lakes, and marshes, giving the landscape a unique character. The geographical boundaries of the *Land* are the waters of the Bodensee (Lake Constance) and the upper Rhine in the south, the widening Rhine Valley in the west, the

River Main in the north, and the River Iller in the east. In addition, the source of the River Danube is at Donaueschingen, a popular excursion point, and the river cuts through the eastern part of the state on the first part of its journey across the European continent. The Danube is the main drainage basin south of the European water divide, which bisects the *Land*.

Source of the Danube

Using criteria from both physical and human geography, it is possible to divide Baden-Württemberg into the following eight natural regions.

The valley of the upper Rhine. Before the Roman conquest of western Europe, the upper Rhine River was one of the main trading arteries on the Continent, and this region also included immense hardwood forests, most of which have fallen prey to floods and the timber industry over the ensuing centuries. The fertile southern part of the upper Rhine Valley now has many vegetable orchards, and the sun-drenched vineyards around Mount Kaiserstuhl produce wine that ranks among the finest of all wines produced in the German Federal Republic.

The Black Forest. Germany's largest continuous forest area, the Black Forest spreads westward to the banks of the Rhine River. Idyllic valleys break its uniformity, and over the years, low-lying portions have filled with water, with many small lakes now contributing to the forest's enchanting, if somewhat foreboding, scenery. The highest point is the Feldberg, 4,898 feet (1,493 metres). The Black Forest edges into the Hotzenwald (Hotzen Forest) in the south, where many lakes and reservoirs feed the numerous power stations. Typical of this area is the so-called *Schwarzwaldhaus*, or Black Forest house, with its roof jutting far beyond its sides and its driveway leading straight up into the hayloft under the roof of the barn. The owners of these small holdings live predominantly from cattle breeding, the timber industry, and tourism.

The Alpenvorland ("alpine foreland"). This deep trough at the edge of the Alps stretches from the formerly volcanic area of the Hegau Mountains in the west to the meadows of the Allgäu in the east. Within its area lies the famous Bodensee and numerous, apparently irregular, rolling hills, with many lakes and marshes, which give the region a distinct appearance. The marshy ground is used for therapeutic baths, hence the number of health spas in this area. Here, too, small holdings predominate, with a solitary main building containing living quarters in the front, barn and hayloft in the back, threshing floor in the middle, and stables lining both sides. The farmers' main income is derived from cattle breeding and dairy products. The Allgäuer cheese is internationally famous.

The Schwäbische Alb (Swabian Alb). Emerging from the flats of the Alpenvorland but sectioned off from it by the Danube Valley, the Schwäbische Alb covers the area between the Black Forest and the Fränkische Alb (Franconian Alb). In the north its mountains fall abruptly into the valley of the Neckar River. Chalk formations and depleted forests make the Schwäbische Alb a barren terrain and Baden-Württemberg's poorest district. The weaving of linen textiles and sheep raising were the main sources of income for the population before the onset of synthetic textiles curtailed the breeding of sheep and forced many farmers to seek additional income in the cities of Heidenheim, Ulm, Reutlingen, or Balingen.

The Neckarland. The fertile Neckarland region belongs among the most densely populated areas in the entire Federal Republic. There is a profusion of vineyards along the Neckar and its many tributaries; other produce grown in the region includes potatoes, sugar beets, and a variety of fruit and vegetables, together with some grain. The many medieval castle ruins have left a distinctive mark on the partly forested landscape, which is also broken by occasional cornfields. Small villages used to line the local highways, but since the end of World War II new high buildings have pushed city and town limits further and further into these surrounding rural districts.

Hohenlohe. The granary of Baden-Württemberg, the Hohenlohic district, lies around the old free city of Schwäbisch Hall and extends all the way to the borders of Bavaria at Rothenburg ob der Tauber. Unlike the custom of the Alb region, where holdings were divided among heirs, the laws of primogeniture (inheritance by the first-born) in this area resulted in a preservation of large estates. A bad effect of this has been that the many young people who do not inherit any land at all have had to find work somewhere else. The numerous, often well preserved, castles in this area are nevertheless ample evidence of the wealth of Hohenlohe in past centuries.

Large estates preserved

The Odenwald (Oden Forest). Often called *Badisch-Sibirien* (the "Siberia of Baden"), the hilly Odenwald region unites Baden-Württemberg with the *Land* of Hessen, in the north. Its location outside the main traffic arteries as well as its raw climate prevented any cultural or economic growth for centuries, and only in the years since 1950 has a developing small industry created extra income possibilities for the local small farmer.

The Kraichgau. Located between the Rhine and Neckar rivers, the fertile Kraichgau district is the site of wheat, corn, tobacco, and fruit culture. The Schwetzinger asparagus of this area is famous far beyond its borders. The castles of Schwetzingen and Bruchsal, reconstructed since World War II, complement the many castles around the cities of Karlsruhe and Mannheim.

The climate of Baden-Württemberg varies greatly among the regions of the *Land*. The upper Rhine Valley is the warmest area, with a yearly mean average of 48°–50° F (9°–10° C) whereas the Alb—the "raw Alb"—is the most inhospitable, with a mean average of about 40°–44° F (4.5°–7° C). Here, and in parts of the Black Forest, there is a yearly average of two months of frost. As a rule, spring comes to the southern part of the upper Rhine Valley before the 20th of April but does not reach the highest regions of the Alb until after the 25th of May. The latter region also has the highest amount of precipitation in the *Land*, because of the westerly winds that drive ocean cloud formations across France to discharge over the slopes of the Black Forest and the Alb. The annual rainfall in the upper Rhine Valley is 26 inches (650 millimetres), compared with 79 inches (2,000 millimetres) on the Feldberg, a favourite ski resort. The average precipitation in the Alb district is 40 inches (a little over 1,000 millimetres), but in the valley of the Neckar River and in the valley of the Tauber River, lying further east, the amount is often less than 24 inches (600 millimetres), and most of this is summer rain.

Climatic variation

A characteristic feature of Baden-Württemberg is the great number of urban settlements; the urban density is two to three times that of northern Germany. By the 1970s, more than 60 of these settlements, many of which had been founded by the Staufers (one of the numerous lesser rulers who governed this area at one point or another in its long history), had populations less than 2,000. Such towns as Ludwigsburg, Rastatt, and Öhringen still retain their typical residential character. The garrison towns, such as Ulm and Münsingen, are of more recent date. Heidelberg, Tübingen, and Freiburg im Breisgau, university centres dating back to the Middle Ages, have been joined in recent years by new universities in Konstanz and Ulm.

The people. The north German regards the people of Baden-Württemberg with some contempt. The nickname *Schwaben*, or even *Spätzle-Schwaben*, is often used. *Spätzle*, a local variety of homemade dumplings, is the favourite staple dish of south Germans. The term *Schwaben* is a misnomer, since most of the native Swabians, descendants of the Suabi, an ancient Germanic tribe, live only in the southeast of the state. The people in the west and southwest of Baden-Württemberg are Alemanni, blood relatives of the French Alsacians and the neighbouring Swiss Alemanni. The influence of the Palatinate population is very strong in the northwest of Baden-Württemberg, whereas the Franconians pushed their way into the centre of the state from the northeast. The linguistic boundary between Franconians and Swabians runs approximately from Baden-Baden in the west, through the Stuttgart area, to Crailsheim in the east.

The geographical boundary between religions has no connection with the origin of the people. Catholics outnumber other denominations in the predominantly Ale-

Religious distribution

mannic Südbaden and Südwürttemberg; Protestants and Evangelists constitute the majority in the more Franconian Nordwürttemberg, and both faiths are more or less equally represented in Nordbaden. Historical developments within the state account for these differences: some ruling houses were Catholic, others, Lutheran Protestant, and each left its mark on the local subjects. In addition to these two main religions, there is a great variety of sects and free churches, especially in Württemberg, most of them a part of the Pietistic movement or of other Protestant origin.

Baden-Württemberg's great post-World War II expansion owed much to the fact that almost a quarter of its population is composed of people who moved to the *Land* as fugitives or displaced persons from the east. Their influx is partially explained by ancestral links between them and the states of Baden and Württemberg in previous centuries. In addition, many simply saw opportunities for a new start in this part of Germany, which had been spared the brunt of wartime destruction. From 1945 to 1950, the rural areas of the state provided the best prospects for housing and employment, but the following years saw a return of the working force to the industrial centres—so much so that many a local farmer's son or daughter got caught up in the ensuing migration from rural areas to the cities. The capital, Stuttgart, witnessed a spectacular growth, gaining as many as 10,000 people a year, and there was severe depopulation of many rural districts. By the early 1970s, only the high rents in the cities apparently kept even more people from moving to the locality in which they worked. Many preferred to build their own home on cheaper ground in small dormitory villages, and to commute instead. Stuttgart alone has more than 100,000 commuters daily, almost one-quarter of the total working force, and a quarter of the entire working force of the *Land* are daily commuters.

The economy. Baden-Württemberg may be regarded as the one West German *Land* in which economic life is dominated by middle class businessmen and small farmers. Although such world famous firms as Daimler–Benz started as small workshops in Stuttgart and Mannheim, there is no heavy industry in the region. On the other hand, Baden-Württemberg is the centre for highly specialized mechanical and textile industries. The lack of valuable mineral and other deposits in Baden-Württemberg forces the population to earn its livelihood by the manufacture, improvement, and finishing of goods. Baden-Württemberg now produces 90 percent of all clocks and watches produced in West Germany, 83 percent of all custom jewelry, 50 percent of all leather goods, 45 percent of all musical instruments, 40 percent of all instruments used in medicine, 39 percent of all food and produce, 37 percent of all cigars, and 30 percent of all hardware.

Lack of resources

The industrial centres are concentrated in the Neckar Valley between Esslingen, Stuttgart, and Heilbronn, and this area accounts for more than half the total production of the *Land*. Other industrial areas are found on the banks of the Rhine near Mannheim—the *Land's* second largest city after Stuttgart—and near Karlsruhe and Ulm. More recently, the border district of the upper Rhine has gained in economic importance; close to the French and Swiss borders and in the centre of the European Economic Community, it is the preferred site for new branch offices, of German, as well as French and Swiss, companies.

At the 1970 census, 56 percent of all those gainfully employed in Baden-Württemberg worked in the production industry, with a shrinking 7 percent in agriculture and forestry, 14 percent in trade and commerce, and 23 percent in other fields of the economy.

Agricultural decline

Agriculture continues to pose problems: at the start of the 1970s there were 157,000 farmers with total holdings of less than 12 acres (about five hectares) of arable land apiece. Their economic survival was seen to depend on their ability to buy or lease additional land. Similarly, the 51,000 farmers with slightly larger holdings had to specialize—in animal breeding, produce, or wine production—if they, too, were not to become bankrupt. Of the state's total of 264,000 farms, only 3,250 were larger than 72 acres (30 hectares) of land. Most of the small farmers were therefore forced to earn their livelihood in industry, returning to their farms in the afternoon, and taking their factory vacations during harvest time.

Many of the farmers of the early 1970s added to their income by converting either their own homes, or other nearby property, to tourist use. The well-known spas of Baden-Baden, Wildbad, and Badenweiler provided additional tourist facilities, while 30 other smaller spas had been enlarged and improved considerably with financial help from the *Land* authorities.

Transportation. Lacking natural resources, and forced to depend mainly on commerce and trade, Baden-Württemberg pays particular attention to its transportation system. As early as 1955 the government prepared a general plan that, by the early 1970s, had been twice improved and adapted to more recent technological developments. The plan called for three express highways (*Autobahn*), traversing the state from north to south, and four more running from west to east. These were to be supported by an extensive system of improved four-lane smaller highways, together with appropriate railway developments. The Rhine and Neckar have been improved as waterways, augmenting this intricate network. By 1971, the Neckar had been canalized as far as Plochingen, and the Rhine could be used for shipping as far as Rheinfelden. In the early 1970s, negotiations were under way with Switzerland with a view to dredging the Rhine up to Waldshut, and possibly to the Bodensee, making the river navigable up to these points. Finally, Baden-Württemberg, near Stuttgart, has one international airport (accounting for over 1,650,000 air passengers in 1970) and some 22 smaller airfields.

Administration and social conditions. Since 1952 the state assembly (Landtag) of Baden-Württemberg has had 120 members, distributed in proportion to the population of the four administrative districts of Nordwürttemberg (capital Stuttgart), Südwürttemberg-Hohenzollern (capital Tübingen), Südbaden (capital Freiburg), and Nordbaden (capital Karlsruhe). Seventy members are elected by direct popular vote, the remainder by proportional representation. The *Land* entered the 1970s as the only one in the Federal Republic still governed by the "grand coalition" of the Union of Christian Democrats (CDU) and the Social Democrats (SPD). The smaller Free Democrats (FDP/DVP) also retained representation, but the Union of Refugees and Disowned (Bund der Heimatvertriebenen und Entrechteten; BHE) and the Communists lost their minor representation of the 1950s in the ensuing decade, which, from 1968 onwards, saw the rise of the right-wing National Democrats. As a rule, the strongest faction in the state assembly chooses the president after the elections; after majority confirmation, he in turn appoints his ministers, choosing within and without the assembly. In the early 1970s, the left-centre Social Democrats, the strongest party in Baden-Württemberg in terms of organization, had over 50,000 members; the Union of Christian Democrats had slightly more than 45,000 members; and the Free Democrats had a little more than 10,000 members.

Political parties

The *Land* is divided into two judicial districts: those of the supreme assize courts of Baden and Württemberg, each including several provincial courts and many local courts. The local courts were extensively reorganized into larger districts in the early 1970s. A peculiarity of Baden-Württemberg is its community courts, in which lay officials may settle civil rights disputes within the village or community. The local notary office is also a unique feature among the German *Länder*.

The Centre for the Clearing of National Socialist Crimes (Zentrale Stelle zur Aufklärung nationalsozialistischer Verbrechen) in Ludwigsburg has gained an international reputation, sifting thousands of documents from foreign archives concerning atrocities committed by Germans under the Third Reich. It has also started court proceedings against former Nazis.

The West German supreme courts are located in Karls-

ruhe: the federal constitutional court (Bundesverfassungsgericht) settles constitutional questions, and the federal court (Bundesgerichtshof) is the highest court of appeal for criminal and civil law in the Federal Republic.

At the end of World War II, the greater part of Baden-Württemberg was occupied by American troops, and United States military headquarters have continued to be located in Heidelberg, with many American garrisons in the cities and towns of north Württemberg and north Baden. The headquarters of the limited French forces are in Baden-Baden, while the commander of the Federal German forces for Baden-Württemberg has his headquarters in Stuttgart.

Baden-Württemberg has more universities than any of the other *Länder* of the Federal Republic. In addition to the old classical universities of Heidelberg, Freiburg, and Tübingen, there are technical universities at Stuttgart and Karlsruhe, an agricultural university in Stuttgart-Hohenheim, and a university in Mannheim specializing in economics. The Ulm University for medicine and natural sciences and the reform University of Konstanz were both founded in the 1960s. There are also many other institutions of higher education. Approximately two-thirds of the three- to six-year-olds in the *Land* are enrolled in *Kindergarten*, which get most of their support from churches and similar organizations, and there is a substantial enrollment in other branches of primary and secondary education, in which structure follows the national pattern.

From the 1950s onward, the *Land* government has been greatly concerned with the social welfare of its citizens. It has produced a hospital plan, a plan for the aged, a plan for youth, and an extensive social report. As a result, medical services were extensively improved during the 1960s, with no less than 45 specialized hospitals having been constructed, with further enlargement and modernization of some 70 existing hospitals.

It is not surprising that living costs, wages, and rents differ greatly in the various parts of the state because of its diverse economic structure, with the rural areas being low in living costs and wages and the cities offering high wages, often with excessively high rents. Generally, the level of earning power in Baden-Württemberg exceeds that of other *Länder* in the German Federal Republic. The overall cost of living index rose from 100 to 158 points between 1950 and 1970.

Cultural life and institutions. Baden-Württemberg is strong in architectural monuments. Gothic churches abound in Ulm and Freiburg, and baroque churches in Weingarten (Kreis Ravensburg), Birnau, Steinhausen, Zwiefalten, and Mannheim, together with the former Kaiserpfalz (Kaiser Palace) in Wimpfen and the castle of Rastatt, are popular sightseeing attractions. The state theatres in Karlsruhe and Stuttgart have an international reputation, particularly marked in the case of the Stuttgart ballet. Of the three provincial and eight city theatres, the Mannheimer Nationaltheater merits special mention: Friedrich Schiller's *Die Räuber* (*The Robbers*) had its world premiere on this stage. The chamber orchestra of Stuttgart (Stuttgarter Kammerorchester) has a growing reputation. Such poets and writers as Friedrich Schiller, Friedrich Hölderlin, and Hermann Hesse, together with the great philosophers Georg Friedrich Wilhelm Hegel and Martin Heidegger, are among the *Land*'s most famous sons. The sculptor Otto Dix made an important contribution to the German expressionist movement.

Literary heritage

The two radio broadcast stations in the *Land,* the Süddeutsche Rundfunk in Stuttgart and the Südwestfunk in Baden-Baden, have well-known popular orchestras, and each broadcasts three different program services. In addition to five or more important regional newspapers, the *Stuttgarter Zeitung* is of national significance.

The Baden-Württembergian is particularly likely to be a member of a club or society, and membership in such bódies is far above the average of the other *Länder* in the Federal Republic of Germany. Singing, sports, and gardening clubs abound throughout the *Land,* which is also a leader in the number of local historical and archaeological societies. The Schwäbische Albverein (Swabian Alb Walking Club) is the largest in the whole Federal Republic. Like the Schwarzwaldverein (Black Forest Walking Club), it concerns itself mainly with wildlife preservation.

The numerous adult education clubs and the many university extension courses in the *Land* testify to the continuing importance of education tradition in the region. Since 1970 all branches of public adult education have been brought together and are now a separate unit, the Fortbildungswerk. The annual class attendance of over 1,000,000 persons, many of them under 25, is an indication of the degree of public interest.

From 1952 to 1970, the total economic output of the *Land* increased by 500 percent. During the same time jobs in industry increased by 60 percent and, at the start of the 1970s, were confidently expected to continue to increase at a rate exceeding that of any other German *Land.* As a direct result, Baden-Württemberg is able to set aside more than a quarter of its annual total tax budget for the extension of its educational system, while paying more than 500,000,000 (US $137,000,000; $1 = DM 3.66) marks annually to the poorer states of the Federal Republic, under a national plan for the equalization of finances. Next to the city states of Berlin, Hamburg, and Bremen, Baden-Württemberg is the state with the highest proportion of students in all age groups. In the early 1970s, this was combined with a high number of car owners, and the second lowest rate of unemployment per 1,000 inhabitants. A government report of this period was not necessarily unjustified in concluding with a quotation from Theodor Heuss, "I believe that the daring prophecy that a self-reliant state could evolve there has been fulfilled."

BIBLIOGRAPHY. EBERHARD KONSTANZER, *Die Entstehung des Landes Baden-Württemberg* (1969), is a history of southwest Germany, with emphasis on the period from 1945 to the present; earlier history is treated in ERNST MULLER, *Kleine Geschichte Württembergs* (1949). See also H. GARDINER BARNUM, *Market Centers and Hinterlands in Baden-Württemberg* (1966), for commerce; and W.M. SCHEDE, *Baden-Württemberg: A Panorama in Color* (1965), for a description of the area.

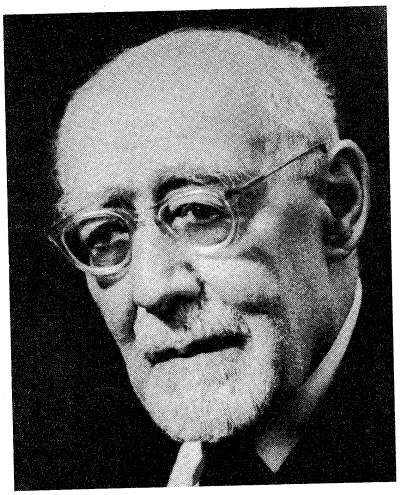
(Er.R.)

Baeck, Leo

Leo Baeck, German Jewish Reform rabbi and theologian, was the spiritual leader and symbol of German Jewry during the period when the Nazis virtually exterminated the European Jews. Baeck was born in Lissa, Posen (then in Prussia), on May 23, 1873. He studied for the rabbinate in Breslau and Berlin, received his Ph.D. in philosophy at the University of Berlin in 1895, and was ordained in 1897 by the progressive Hochschule in Berlin. He immediately displayed his courage and personal independence of thought by being one of the two rabbis within the German Rabbinical Association who refused to condemn the Zionist leader Theodor Herzl (1860–1904) and the First Zionist Congress then meeting in Basel.

Baeck first served as rabbi in Oppeln Silesia (1897–1907), then Düsseldorf (1907–12), and finally Berlin (1912–42). In 1901, Baeck challenged the Protestant theologian and church historian Adolf von Harnack (1851–1930), whose lectures on *The Essence of Christianity* then presented essential original Christianity as a liberal faith that appeared at a unique moment of history and was unrelated to the Jewish tradition. Striving to show the originality of Jesus' teachings, Harnack denigrated the Pharisees and the Judaism they represented and committed lapses of scholarship singled out by the young Baeck. Baeck's own masterpiece, *The Essence of Judaism* (1905), established him as the leading liberal Jewish theologian. In contrast to Harnack, Baeck stressed the dynamic nature of religion, the ongoing development that is man's response to the categorical "Ought," the Divine Imperative. The influence of the German-Jewish philosopher Hermann Cohen (1842–1918) and Neo-Kantianism (German philosophical movement, 1870–1920) is visible, but behind it stands the ethical rigorism of traditional rabbinic thought. The next edition of this

Jewish theologian



Baeck, *c.* 1948
Picture in the archives of the Leo Baeck Institute, New York

work (1922), greatly expanded, moved on toward Baeck's "religion of polarity" with its dialectical movement between the "mystery" of the Divine presence in life and the "commandment" of the ethical imperative that comes to man in his encounter with God. Baeck expressed the twofoldness of religious experience in the concept of *toladot* "generations," the chain of generations that is Jewish history and that made the Jewish people the vehicle of a continuous revelation that became that people's mission. "A light to the nations," it had to teach the revelation by living these teachings. Judaism was seen as the supreme expression of morality, a universal message expressed through the particular existence of Israel.

The dialogue between Christianity and Judaism was brought to greater clarity and intensity by Baeck's refusal to use evasions in his criticism. Traditional Jews disliked Baeck's early (1901) claim that Jesus was a profoundly Jewish figure and his view in *The Gospel as a Document of Jewish Religious History* (1938) that the Gospels belonged with the contemporary works of rabbinical literature. Christians, on the other hand, felt challenged by his definition of Judaism as the "classic" rational faith confronting a "romantic" Christianity of emotion, in his essay "Romantic Religion" (1922). The American philosopher Walter Kaufmann views this work as Baeck's greatest achievement next to the *Essence of Judaism*. Yet one cannot ignore Baeck's final work, written in the concentration camp, *This People Israel: The Meaning of Jewish Existence* (1955), which moves from the essence of an "ism" to the concrete existence of a people and creates an approach to Jewish life that must be set alongside the thought of the great 20th-century Jewish religious philosophers Martin Buber (1878–1965) and Franz Rosenzweig (1886–1929). Its full implications emerge only when the work is placed into the life of the author.

Baeck's life was his work, revealing his concept of polarity: an army chaplain in World War I, he became a pacifist; a non-Zionist, he became head of the German Keren Hayesod ("Foundation Fund" for Palestine land purchases). Baeck was the president of the German B'nai B'rith ("Sons of the Convenant," the main Jewish fraternal and service organization); the chairman of the Rabbinical Association he had once defied; and taught Midrash (interpretative rabbinical literature) and homiletics at the Berlin Lehranstalt. He was called away from this to preside over the end of the 1,000-year-old German Jewish community.

Leader of German Jewry

In 1933, German Jewry's organizations united in the Reichsvertretung der Juden in Deutschland (National Agency of Jews in Germany) under Leo Baeck and Otto Hirsch (1885–1941), the jurist and community leader who was killed in the Mauthausen concentration camp. Under constant attack, this group took charge of Jewish life in Germany. Millions of dollars were spent annually in clearly defined fields; emigration, economic help, charity, education, and culture. Meanwhile, at the conference table with the Nazis, Baeck and the others battled for time so that lives could be saved. Later critics have felt that all resources should have been focussed on emigration; but the extermination camps were inconceivable to the German Jewish community of the 1930s. It planned to survive Hitler behind ghetto and prison walls—a tragic error of judgment, but scarcely avoidable. Negotiating with Nazis always carried dangers of corruption but Baeck was untouched by this. As late as 1939, he brought a trainload of children to England—and returned to Germany. In public and private, his life was a pattern of moral resistance that, after five arrests, brought Baeck to the Theresienstadt (Terezín) concentration camp.

Theresienstadt was a "model" camp, sometimes shown to outsiders. Its inmates were killed by neglect or illness or sent on to the extermination camps. Out of the 140,000 Jews sent to Theresienstadt less than 9,000 survived. The Nazis confused the death of a Rabbi Beck of Moravia with Leo Baeck; the latter became Number 187,894 and, incredibly, survived. Baeck set up classes inside the camp: over 700 persons would press into a small barracks to listen to lectures on Plato and Kant. This, too, was a way of resistance. There were also Christian inmates whom Baeck served as pastor. Once more, the miasma of evil surrounded him but could not touch him. Critics have said that he was too aloof, or too liberal, but the only criticism to be taken seriously deals with Baeck's decision not to pass on rumours that the "resettlement" trains led to the death camps. The eminent Protestant theologian Paul Tillich (1886–1965), who admired Baeck, asserted that "Baeck should have spoken out . . . the full existential truth must always be made available." Baeck, however, thought the helpless victims should not be deprived of the hope keeping many alive.

On May 8, 1945, the day before Baeck was to be executed, the Russians liberated Theresienstadt, and Baeck stopped the inmates from killing the guards. He survived for a number of years, settling in England, and becoming a British subject; he taught and lectured in Britain and the United States, including a term at Hebrew Union College in Cincinnati. His final writings, notably *Individuum Ineffabile* (1948) and *This People Israel*, continued to express hope in man and the human situation as the area of the revelation. He died in London, on November 2, 1956. In his life, Baeck summarized the greatness and perhaps also some of the flaws of German Jewry, which placed all of its hopes and commitments in western European civilization. In his teachings Baeck gave perhaps the clearest systematic exposition of liberal Jewish religious thought in the 20th century.

BIBLIOGRAPHY. Important works of Leo Baeck include: *Das Wesen des Judentums* (1905, 6th ed. 1932; Eng. trans., *The Essence of Judaism*, 1948), a classic text of modern Judaism; *The Pharisees and Other Essays*, with an introduction by KRISTER STENDAHL (Eng. trans. 1966), a text covering basic historical questions concerning Jewish life at the time of the emergence of Christianity; *Judaism and Christianity*, with an introduction by WALTER KAUFMANN (Eng. trans. 1961), a collection of the more polemical writings of Baeck clearly defining Judaism's disagreements with Christianity; and *Dieses Volk; Jüdische Existenz* (1955; Eng. trans. by A.H. FRIEDLANDER, *This People Israel*, 1966), Baeck's final work, partly written in a concentration camp, covering 3,000 years of Jewish history. The only full-length biographical study published to date is A.H. FRIEDLANDER, *Leo Baeck: Teacher of Theresienstadt* (1968), containing an exposition of Baeck's teachings and an extensive bibliography.

(A.H.F.)

Baer, Karl Ernst von

The 19th-century scientist Karl Ernst von Baer led an extraordinarily varied life; he had a universal mind that ranged both wide and deep. By his comparative and descriptive studies of animal development, he established in modern biology the idea that the life of an animal is a historical event, proceeding from simple to complex, from general to special. He showed that mammals develop from eggs, and that the same organs in different animals develop from the same basic tissue

layers. He gave meaning, by his observations and generalizations, to the concept of epigenesis. This systematization of embryology was responsible for the fact that this new science supplemented comparative anatomy as the dominant biological discipline. He was also a pioneer in geography, ethnology, and physical anthropology.

By courtesy of the Hunt Botanical Library, Carnegie-Mellon University, Pittsburgh



Baer, lithograph by Rudolph Hoffmann, 1859, after a photograph.

Born on his family's estate in Piep, Estonia, on February 29, 1792, Baer, one of ten children, spent his childhood with an uncle and aunt before he returned at the age of seven to his own family. His parents, of Prussian descent, were first cousins. After private tutoring Baer spent three years at a school for members of the nobility. In 1810 he entered a provincial Estonian university at Dorpat to study medicine, receiving his medical degree in 1814. Dissatisfied with his medical training, he studied in Germany and Austria from 1814 to 1817. The crucial year of his education was the academic year 1815–16, when his studies in comparative anatomy at the University of Würzburg with Ignaz Döllinger introduced him to a new world that included embryology.

In 1817 Baer began his teaching in Königsberg, where he remained until 1834. In 1820 he married Auguste von Medem of Königsberg, by whom he had six children. Although Döllinger had suggested that Baer begin a study of chick development, he was unable to meet the expense of purchasing the eggs and paying an attendant to watch the incubators. This work was done instead by Baer's more affluent friend Christian Pander, who in 1817 described the early development of the chick in terms of what are now known as the primary germ layers—that is, ectoderm, mesoderm, and endoderm.

Major works

From 1819 to 1834 Baer devoted most of his time to embryology, extending Pander's concept of germ-layer formation to all vertebrates. In so doing Baer laid the foundation for comparative embryology. He made many important technical discoveries. In 1827 he described his discovery of the mammalian ovum (egg) in his *De Ovi Mammalium et Hominis Genesi* ("On the Mammalian Egg and the Origin of Man"), thereby establishing that mammals including man develop from eggs. He opposed the popular idea that embryos of one species pass through stages comparable to adults of other species. Instead, he emphasized that embryos of one species could resemble embryos, but not adults of another, and that the younger the embryo the greater the resemblance. This was in line with his epigenetic idea—basic to embryology ever since—that development proceeds from simple to complex, from homogeneous to heterogeneous. One of the most important books in embryology is his *Über Entwickelungsgeschichte der Thiere* (vol. 1, 1828; vol. 2, 1837; "On the Development of Animals"), in which he surveyed all existing knowledge on vertebrate development and from which he derived his far-reaching conclusions. He identified the neural folds as precursors of the nervous system, discovered the notochord, described the five primary brain vesicles, and studied the functions of the extra-embryonic membranes. This pioneering work established embryology as a distinct subject of research, at least in its descriptive aspects. He marked out the main lines of descriptive and comparative study that had to be accomplished before the modern approach—the causal analysis of development—could emerge.

In 1834 Baer moved to St. Petersburg, Russia, where he became a full member of the Academy of Sciences; he had been a corresponding member since 1826. His first duties were as librarian of the foreign division, but he eventually served the Academy in a variety of administrative positions. He retired from active membership in 1862 but continued to work as an honorary member until 1867. After moving to Russia, Baer abandoned embryology. Particularly interested in the Russian north, he became a courageous explorer there; he was the first naturalist to collect specimens from Novaya Zemlya, then uninhabited. During his extensive travels throughout Russia, he developed a great scientific and practical interest in its fisheries. He made significant discoveries in geography, including one concerning the nature of the forces responsible for the configuration of riverbanks in Russia.

Baer's travels also increased his long-standing interest in ethnography. He contributed to the Academy at St. Petersburg by establishing an extensive skull collection. As a result of his interest in skull measurements, he called a meeting of craniologists in Germany in 1861, which led to the establishment of the German Anthropological Society and the journal *Archiv für Anthropologie*. He was responsible also for the founding of the Russian Geographical Society and the Russian Entomological Society, of which he was the first president.

Assessment

In his early days as an embryologist Baer had begun to consider possible relationships, in terms of kinship, between animals. In 1859, the year that Darwin's *Origin of Species* appeared, Baer published a work on human skulls suggesting that stocks now distinct might have originated from one form; the ideas of the two men were formulated completely independently. Baer, however, was no strong adherent to the doctrine of transformation (the pre-Darwinian term for evolution). Although he believed that some very similar animals, such as goats and antelopes, might be related, he was vehemently against the concept expressed in the *Origin of Species* that all living creatures might have evolved from one or a few common ancestors.

In his philosophical writings—and all his embryological writings were philosophical to some degree—Baer saw nature as a whole, even though not in terms of modern evolutionary theory. He viewed the development of organisms and of the cosmos in the same light, and his all-encompassing view of the universe brought together what might otherwise have seemed diverging threads in his thought.

Following his retirement in 1867, Baer went to Estonia; he died in Dorpat (now Tartu) on November 28, 1876.

BIBLIOGRAPHY. BAER'S autobiography is *Nachrichten über Leben und Schriften*, 2nd ed. (1886). Two biographies in German (there are none in English), both entitled *Karl Ernst von Baer*, are by L. STIEDA, 2nd ed. (1886); and by B.E. RAIKOV (1968), both with extensive bibliographies. BAER'S *Entwickelungsgeschichte der Thiere, Beobachtung und Reflexion*, 2 vol. (1828–37), embodies his major discoveries and concepts in embryology; three volumes containing 700 letters by Baer are in course of publication; one has appeared by T.A. LUKINA (1970)—the letters are in Russian and German, the notes in Russian only; see also JANE OPPENHEIMER, "Baer, Karl Ernst von," in the *Dictionary of Scientific Biography*, vol. 1 (1970), for an analysis of his scientific contributions.

(J.M.O.)

Baffin Bay

A thumb-shaped extension of the North Atlantic Ocean, Baffin Bay is really a sea lying between the west coast of Greenland and the islands of the north Canadian Arctic

Archipelago. It extends northwest from the Davis Strait, which covers the Baffin-Greenland shelf (a submarine feature) at the latitude of the Arctic Circle, to latitude 80° north, where the narrow Nares Strait leads directly into the Arctic Basin. Jones and Lancaster sounds provide alternative links to the Arctic Basin through the Canadian Arctic Archipelago. Baffin Bay covers an area of 266,000 square miles (689,000 square kilometres)—slightly larger than Burma—and is about 900 miles (1,450 kilometres) long. A pit at its centre, the Baffin Hollow, plunges to a depth of 7,000 feet (2,100 metres), and the bay, although little exploited by man because of its hostile environment, is of considerable interest to geologists studying the evolution of the North American continent.

Exploration and scientific investigation

Baffin Bay also played an important role in the exploration of North America by Europeans. It was discovered in May 1615 by Robert Bylot, an English sea captain, but his supposedly mutinous tendencies prevented his name from being given to the entity, and the honour went instead to his lieutenant, William Baffin. Even the latter's discoveries came to be doubted until the later explorations of Captain (later Sir) John Ross, in 1818. The first scientific investigations since Bylot's map of the shores were conducted in 1928 by a Danish and also by an American expedition, followed by another, more extensive, survey in the 1930s. Patrol vessels, now aided by aircraft, have long investigated ice distribution in the region, and after World War II a Canadian expedition undertook complex investigations. (For related information see ARCTIC ISLANDS; GREENLAND; NORTHWEST TERRITORIES; and NORTHWEST PASSAGE. See also NORTH AMERICA; WESTERN ARCTIC CULTURES.)

The physical environment. Baffin Bay's oval floor is fringed by the submarine shelves of Greenland and Canada and by ledges at the mouths of sounds. Apart from the central pit, depths range from 800 feet in the north to 2,300 feet in the south. The sediments are mostly terrigenous (composed from earth materials, rather than from animal shell deposits) and include gray-brown homogeneous silts, pebbles, and boulders. Gravel lies everywhere.

The climate is severe, especially in winter, when northeast winds blow off Baffin Island (in the south) and in the bay's northern sector. Northwesterly and southwesterly winds predominate in summer. Easterlies blow off the Greenland coast, and storms are frequent, notably in the winter. January temperatures average −4° F (−20° C) in the south and −18° F (−28° C) farther north. The absolute minimum recorded is an icy −46° F (−43° C) at Jakobshavn, on the Greenland coast, but, paradoxically, there are occasional winter thaws occasioned by the warm, dry, foehn winds that sweep down from the valleys containing Greenland's glaciers. In July the temperature on the shores averages 45° F (7° C), with some snow. Overall, the annual precipitation off Greenland amounts to four to ten inches (100 to 250 millimetres), reaching twice this amount off Baffin Island.

The currents in Baffin Bay flow in a counterclockwise direction, as the West Greenland Current brings in about 35,000,000 cubic feet (990,000 cubic metres) of warm water across the threshold of Davis Strait. At the same time, about twice as much cooled water empties out into the Atlantic along the Baffin Island coast, as Arctic waters push in through the northern sounds. North of 72° N a zone of intensive counterclockwise currents forms an impediment to the West Greenland Current, which consequently hugs the Greenland coast (herding numerous icebergs northward in the process) from Disko Island to Thule (Greenland), at the Bay's northeast tip, where the current turns left into the cold flow.

Icebergs are dense even in August: the ice cover is formed from Arctic pack ice entering through the northern sounds, from local sea ice, and from icebergs that have broken off adjacent glaciers. By late October icefields reach Hudson Strait (between Baffin Island and the Quebec mainland), a region where coastal ice has already been thickening, mostly near Greenland, where prevailing easterly winds make for sheltered conditions. The centre of Baffin Bay is covered with compounded ice in winter, but in the north there is actually a permanent ice-free area (the "northern water") that may be related to the warming effect of the West Greenland Current. Iceberg concentration in the bay is similarly affected. By origin, about 70 percent of the icebergs that drift down to the Newfoundland area have been found to come from glaciers pushing into Baffin Bay between 69° N and 71°40′ N, whereas some 20 percent are formed still further north, in the Melville Bay area, just south of Thule. The largest icebergs are thrust out into the unimpeded waters of the "northern water" area and can be 230 feet long above the surface, reaching down to 1,300 feet below.

The "northern water" icebergs

Sea-water temperature and salinity follow naturally from the ice distribution, currents, and the submarine and other topographic influences already described. The salinity of Arctic waters flowing into Baffin Bay ranges from 30.0 to 32.7 parts per thousand (‰) and their temperature warms up to 41° F (5° C) on the surface in summer, cooling in winter to 29° F (−2° C). The layers 1,300 to 2,000 feet deep reach 34° F (1° C) and a salinity of 34.5 ‰. Below 3,300 feet in the central regions, the water—probably Atlantic in origin—reaches 31° F (−0.5° C), and has a salinity of 34.4 ‰.

Tides are an important and interesting feature: near Baffin Island and the shores of Greenland, the tidal range is about 13 feet, reaching as much as 30 feet where the water is forced through narrow passages. The tidal rate varies between 0.6 and 2.3 miles per hour and the direction of the tides varies by as much as 180°. This phenomenon produces unequal pressure on the fields of floating ice and results in the churning together and crushing of fresh, old, and pack ice. The tidal currents also mix the water layers, enriching the upper strata with nutrient salts and the lower ones with dissolved oxygen.

The biological environment. The dissolution of salts in the water and the warming effect of the currents moving in from the south encourage the development of life-forms from the smallest creatures of the sea to whales. The numerous single-cell algae nourish small invertebrates, notably euphausiids (an order of small, shrimp-like crustaceans), and these in turn are food for larger invertebrates, fish, birds, and mammals. Baffin Bay contains Arctic flounder, four-horned sculpin (a spiny, large-headed, broad-mouthed fish), polar cod, and capelin (a small fish of the smelt family). Migrant fish from Atlantic waters include cod, haddock, herring, halibut, and grenadier (a tapering-bodied, soft-finned fish). Wildlife also includes ringed seals, bearded seals, harp seals, and —in the north—walrus, dolphins, and whales (including killer whales). The coasts are homes of gulls, ducks, geese, eiders, snowy owls, snow buntings (a type of finch), ravens, gerfalcons, linnets, and sea eagles.

Plant cover, too, is remarkably varied, with about 400 types represented. Shrubs include birch, willow, and alder, and also halophytic plants (*e.g.*, those adapted to salty soils), as well as lyme (or tussock) grass, mosses, and lichens. These provide food for rodents and the splendid caribou of the area. Polar bears and Arctic foxes also abound. Large-scale fishing remains undeveloped because of the perils of the heavy ice cover, but local residents—who are mainly Eskimos—carry on some fishing and hunting, often with traditional methods.

BIBLIOGRAPHY. FARLEY MOWAT (ed.), *Ordeal by Ice* (1960), is a collection of original sketches of British expeditions carried out through the 16th–19th centuries, including a history of the discovery of Baffin Bay. See also W.B. BAILEY, "Oceanographic Features of the Canadian Archipelago," *J. Fish. Res. Bd. Can.*, 14:731–769 (1957).

(M.M.A.)

Bagehot, Walter

Walter Bagehot has been described as Victorian England's most versatile genius; essentially he was one of the most influential British journalists of the mid-Victorian years as editor of *The Economist* from 1860 until his early death in 1877. But in addition, he wrote a series of literary essays that have been continually republished

throughout the past century, a book on British politics that remains a widely read classic, and one of the earliest sociological studies to apply the concept of evolution to societies themselves; in addition, he made an important contribution to the theory of central banking. "Had I command of the culture of men," wrote U.S. president Woodrow Wilson, "I should wish to raise up for the instruction and stimulation of my nation more than one sane, sagacious, critic of men and affairs like Walter Bagehot." "Those who have the good fortune to know him still remember him as perhaps the most original mind of his generation," wrote Lord Bryce, British ambassador in Washington and the author of *The American Commonwealth*.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Bagehot, mezzotint by Norman Hirst (born 1862), after a photograph.

Bagehot was born at Langport, Somerset, on February 3, 1826. His father's family had been general merchants there for several generations, whereas his mother—who was a great beauty but was ten years older than his father and had had a tragic first marriage—was a sister of Vincent Stuckey, the head of the largest bank in the west of England. It was the opinion of his relations that his acute political sense derived from his father, whereas the sparkle and originality of his mind came from his mother, even though she became partly insane as she grew older.

Education

Bagehot had the severe schooling of an early Victorian. As a child he went to Langport Grammar School, which had a famous headmaster who had been a friend of the poet Wordsworth; at 13 he was sent to Bristol College, one of the best schools in Great Britain. There he received a grounding in philosophy, mathematics, literature, the classics, and the new natural sciences, of an intensity that no English child today would be thought capable of assimilating.

The obvious university to choose was University College, London, because his father was a Unitarian, and Oxford and Cambridge in those days were dogmatically Anglican. He was a "lanky youth, rather thin and long in the legs with a countenance of remarkable vivacity and characterised by the large eyes that were always noticeable," wrote Sir Edward Fry, one of his friends at Bristol. He had a rather sardonic manner that did not endear him to all of his contemporaries, but he did make a number of lasting friends at University College, notably Richard Holt Hutton, who was for the latter part of the century the distinguished editor of *The Spectator;* William Roscoe, the grandson of the famous historian of the Medicis; Arthur Hugh Clough, the poet; and, of an older generation, Henry Crabb Robinson, who had been the friend of Goethe, Schiller, and Coleridge, and *The Times* correspondent in the Napoleonic Wars. In 1846 Bagehot took his bachelor's degree with first-class honours, despite bad health, and in 1848 his master's degree with the university's gold medal in moral and intellectual philosophy.

For three years after graduation, he studied at the bar but he never liked it, and it was chance that took him into literature. He happened to be in Paris at the end of 1851, when Louis Napoleon's coup d'état took place, and he wrote a series of articles in the leading Unitarian weekly journal of the day describing the coup at first-hand and defending Napoleon. These articles caused much controversy because the coup was widely disapproved of in England. But they convinced Bagehot that he could write, and he settled down to work in his uncle Stuckey's bank, writing in the next six or seven years a series of literary essays on Milton, Shakespeare, Gibbon, Sir Walter Scott, Pierre-Jean de Béranger, together with studies of leading political figures of the past century—Henry St. John Bolingbroke, William Pitt, Sir Robert Peel, and others—that are still widely quoted.

His entry into professional journalism was also accidental. In his role as a banker, he had written various economic articles that had attracted the attention of James Wilson, the man who had founded *The Economist* in 1843 and who was then an influential member of Parliament and financial secretary to the treasury in Lord Palmerston's government. Wilson asked Bagehot to stay, and he immediately fell in love with Eliza, the eldest of Wilson's six daughters. They were married in April 1858, but they had no children, and it is doubtful if Eliza's rather cold personality really suited the warmth and vigour of her husband's.

Work on *The Economist*

He went back to manage the Bristol branch of Stuckey's bank. But a year later Wilson was asked to go to India to reorganize the finances of the Indian government, and he died in Calcutta in 1860, leaving Bagehot in charge of *The Economist*. For 17 years he wrote the main article and improved and expanded the statistical and financial sections that have made it the leading business journal and one of the leading political journals of the world for more than 100 years. More than that, he humanized its political approach with a greater emphasis on social problems. As the American political scientist Walt Rostow has commented, "*The Economist* was not simply the hard bitten advocate of the mid Victorian capitalist."

Bagehot described himself as a conservative Liberal or "between size in politics." Unlike many Liberals, he had grown up in the deep countryside, and he had a strong feeling for the social problems that rapid industrialization and urbanization were creating in Britain. He was also an acute observer of international affairs, with an instinctive affection for France and an equal distrust of Otto von Bismarck's Germany. His early years at *The Economist* coincided with the American Civil War, on whose development he wrote nearly 20 articles; instinctively, he was a Confederate like many of his British contemporaries, but his reason made him a supporter of Abraham Lincoln, of whom he wrote on the day the news of his assassination reached England:

> We do not know in history such an example of the growth of a ruler in wisdom as was exhibited by Mr. Lincoln. Power and responsibility visibly widened his mind and elevated his character. Difficulties, instead of irritating him as they do most men, only increased his reliance on patience; opposition, instead of ulcerating, only made him more tolerant and determined.

In 1867 he published *The English Constitution*, which was an attempt to look behind the facade of the British system of government—crown, Lords, and Commons—in order to see how it really operated and where true power lay. He was one of the first to observe the overriding power of the Cabinet in a party that commanded an effective majority in the House of Commons. He cultivated a number of close political friendships, notably with William Ewart Gladstone, who became the first Liberal prime minister in 1868; with Lord Carnarvon among the Conservatives (the author of the British North America Act, the constitution of Canada); and with William Edward Forster (the author of the first public education act in Britain).

Bagehot never succeeded, however, in entering politics himself. He tried at Manchester, at Bridgwater near his Somerset home (a district that had a notorious reputation

for corruption), and in 1867 for London University. But he was a poor speaker and failed each time.

In 1872 Bagehot published *Physics and Politics*, which was an attempt to apply the new discoveries in anthropology to the development of societies and nations themselves. It is largely forgotten by reason of the vigour acquired by sociological investigation in the 20th century, largely under the stimulus of Karl Marx and Max Weber. But one of its central points, the process of unconscious imitation as a molding force in the development of nations—what Bagehot called "the cake of custom"—had a considerable influence on such philosophical sociologists as William James and Graham Wallas.

All this time, Bagehot and his wife were living in London and he was editing a weekly of growing influence. In his 40s, he became increasingly frail, and such energy as he had was concentrated on professional economic studies. In 1873 he published *Lombard Street*, which, though really a tract arguing for a larger central reserve in the hands of the Bank of England, in fact contains the germ of the modern theory of central banking and exchange control. He was working on a major series of economic studies when pneumonia struck him down on March 24, 1877, at the age of 51. The economist John Maynard Keynes, two generations later, paid tribute to his insight into business psychology.

But the greatest tribute to Bagehot's lively style, humanity, and insight is that his books have been read, republished, and subjected to a continuous stream of critical essays ever since his death. He once made fun of Thomas Macaulay for seeking posthumous fame but has, nevertheless, received a good measure of it himself.

BIBLIOGRAPHY

Biographies: EMILIE I. BARRINGTON (ed.), *The Love-Letters of Walter Bagehot and Eliza Wilson, Written from 10 November, 1857 to 23 April, 1858* (1933); ALASTAIR BUCHAN, *The Spare Chancellor: The Life of Walter Bagehot* (1959), a short, critical biography dealing with all aspects of Bagehot's life, work, and thought; NORMAN ST. JOHN-STEVAS, *Walter Bagehot* (1959), a selection of Bagehot's political studies with a biographical introduction and a useful bibliography.

Collected works: EMILIE I. BARRINGTON (ed.), *The Works and Life of Walter Bagehot*, 10 vol. (1915), series containing his books, most of his essays, and over 50 of his *Economist* articles; *The Collected Works of Walter Bagehot*, ed. by NORMAN ST. JOHN-STEVAS (1966–), a more comprehensive edition containing considerably more journalistic material. Vol. 1 and 2, *The Literary Essays*, with an introduction by SIR WILLIAM HALEY, and vol. 3 and 4, *The Historical Essays*, with an introduction by JACQUES BARZUN, have already been published.

Critical works: JOHN MAYNARD KEYNES, "The Works of Bagehot," *Economic Journal*, 25:369–375 (1915), an estimate of Bagehot as an economic writer by the greatest economist of his day; SIR HERBERT E. READ, "Bagehot," in *The Sense of Glory: Essays in Criticism* (1929), a sensitive critique of Bagehot as a litterateur; WOODROW WILSON, "A Wit and a Seer," *Atlantic Monthly*, 82:527–540 (1898), one of the earliest works drawing attention to Bagehot's gifts and versatility by one of his greatest American admirers; GEORGE M. YOUNG, "The Greatest Victorian," in *Today and Yesterday: Collected Essays and Addresses* (1948), an excellent brief portrait.

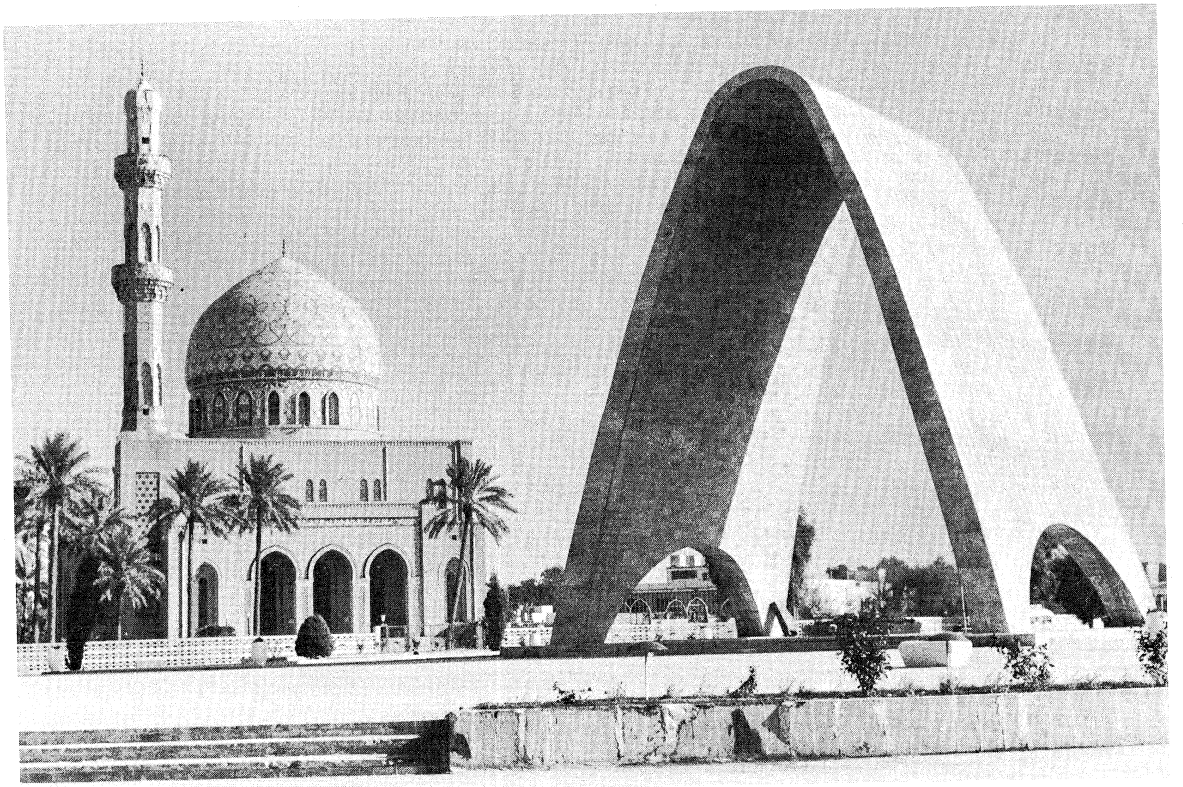
(A.F.B.)

Baghdad

Baghdad, the foremost city of ancient Mesopotamia, is the largest city and the capital of modern Iraq and of the *muḥāfaẓah* (province) of Baghdad. It is situated on the Tigris River in north central Iraq about 350 miles (560 kilometres) northwest of the Persian Gulf. Since the overthrow of the monarchy in 1958, the capital city has been the scene of political turmoil as the country has tried to establish itself as a modern, Socialist, Pan-Arab state. The population of the city early in the 1970s was over 2,100,000.

History. Many capital cities have stood in the vicinity of Baghdad: Agade, Babylon, and Burj ʿAqarqūf, the Kassite capital, lay to the west; Seleucia and Ctesiphon lay about 20 miles to the south. In AD 750 the caliphate was established in Iraq with the foundation of the ʿAbbāsid dynasty by Abū al-ʿAbbās as-Saffāḥ, whose brother and successor, Abū Jaʿfar, known as al-Manṣūr, determined in 762 to build himself a new capital on the site of a Sāsānian village, Baghdad. The Round City of Manṣūr, called Madīnat as-Salām (City of Peace), stood on the west bank of the Tigris, but no traces of its buildings remain, and its exact site is unknown. It was 3,000 yards (2,700 metres) in diameter, with three concentric walls each pierced by four gates, through which passed highways radiating from the caliphal palace at the centre of

Hubertus Kanus—Bavaria-Verlag



Mosque and arch to the Unknown Soldier in Baghdad, Iraq.

the city to the four quarters of the empire. Bazaars were not permitted within the city, and a merchants' quarter, al-Karkh, developed outside the Basra Gate. From the northeast gate the Khorāsān road crossed the Tigris on a pontoon bridge, and beyond this on the east bank lay the palace of al-Manṣūr's heir apparent, al-Mahdī, around which grew up the three suburbs of ar-Ruṣāfah, ash-Shammāsīyah, and al-Mukharrim, the earliest forerunners of the modern city. Baghdad reached the zenith of its prosperity in the 8th and early 9th centuries under Hārūn ar-Rashīd, son of al-Mahdī, who accumulated in his capital the riches and learning of the known world. Many tales reflecting the glory of this period occur in *The Thousand and One Nights*. Hārūn's death was followed by civil war between his two sons, during which the Round City was severely damaged. It was never fully restored; from 836 to 892 the caliphs abandoned Baghdad, where their unruly Turkish bodyguard had become unpopular, for a new capital at Sāmarrā' and after their return at the end of the 9th century they resided in Mukharrim on the east bank. In 1095 a new wall was built around east Baghdad, which survived with many restorations until the 19th century, when it was largely dismantled. One of the two surviving gates, the Talisman Gate, was blown up by the retreating Turks in 1917; the second, Bāb al-Wastanī, or the Middle Gate, still stands and was restored as an arms museum.

Baghdad at its zenith

Despite the declining power of the later caliphs, Baghdad remained a great centre of trade and culture as long as the agricultural prosperity of Iraq continued. Its real downfall came when Hülegü, the Mongol, overran Mesopotamia and sacked Baghdad in 1258, killing the caliph al-Musta'ṣim and, it is said, 800,000 of the inhabitants. The overthrow of the 'Abbāsid government removed the authority that had secured the maintenance of the irrigation system and the protection of the cultivated land against nomadic tribesmen. With its political importance gone, its population decimated, and the economic basis of its life destroyed, Baghdad never again rose above the status of a provincial city until the emergence of Iraq as an independent state after World War I. It remained subject to the Il-Khanid dynasty, successors of Hülegü, until 1340, when it achieved local independence under the Jalāyirids, who made some attempt to restore its fortunes. In 1401 the city was sacked by the last of the Mongol invaders, Timur (Tamerlane), and in 1410 fell under the sway of the Kara Koyunlu (Black Sheep) Turkmens, who gave way in 1469 to the Ak Koyunlu (White Sheep) dynasty. They in turn were expelled by the Persians under Shāh Esmā'īl I in 1508, and in 1534 Baghdad was taken for the first time by the Ottoman Turks under Süleyman I the Magnificent. It was recaptured by the Persian Shāh 'Abbās I in 1623 but was recovered for Turkey by Murad IV in 1638; its frequent changes of allegiance were due in part to its geographical position between the two great empires and in part to the division of its inhabitants between the Shī'ah and the Sunnī sects of Islām, the former favouring the Persian and the latter favouring the Turkish rule. During the 17th century a rapid succession of Turkish governors did little to restore the fortunes of Baghdad. The beginning of the 18th century, however, saw the appointment of two successive governors, father and son, Ḥasan and Aḥmad Pasha, who reformed the administration by introducing the slave hierarchy of Circassian civil and military officials known as the Mamlūks and raised the prestige of their capital until it controlled Mesopotamia from Basra to Mardin, owing only a formal allegiance to Istanbul. They were succeeded by a line of Mamlūk governors until the abolition of the Mamlūk system by Sultan Mahmud II in 1831. During the Mamlūk period foreign, notably British, influence became more marked in Baghdad. A British residency was established in the city in 1798, and in 1802 the resident was granted consular rank; under C.J. Rich (1808–21) and H.C. Rawlinson (1843–55) the prestige of the resident was second only to that of the governor. Baghdad was captured by British forces in 1917 and became the capital of the independent kingdom of Iraq in 1921. The city was the scene of an abortive revolt, inspired by anti-British feeling, in 1942 and of a successful army-led coup d'etat in 1958, when the monarchy was overthrown. Another Baghdad coup d'etat in 1963 brought the Ba'th Party to power, which it held briefly and regained in 1968.

Alternate Turkish and Persian control

The contemporary city. *The city site and layout.* The city's situation on the Tigris River is at the river's nearest point to the Euphrates (25 miles) and 20 miles above the Tigris' confluence with its largest tributary, the Diyālā. The site is a level plain at an altitude of 112 feet (34 metres) above sea level. The city was originally built on the west bank of the river, but for more than 1,000 years the greater part has been on the east bank. There is, however, a large and growing suburb on the west side, comprising the quarters of al-Karkh, Karradet Mariam, and al-Manṣūr. The east-bank city is linked with these quarters by bridges, one of which carries a railway. A bridge also leads to Kāẓimayn (Kadjimain), north of al-Karkh, a considerable settlement that has grown up around the tombs of Mūsā al-Kāẓim and Muḥammad al-Jawād, the seventh and ninth of the 12 *imāms* recognized as the true successors of the Prophet by the majority of Shī'ite Muslims. Although the distinction between the city and the metropolitan area appears imprecise, the area of the former is about 14 square miles and that of the latter, 75 square miles.

Climate. Summers (May–October) are dry and intensely hot. Average daily temperatures in July and August are about 80° F (27° C) before sunrise and 105° to 110° F (41° to 43° C) and up to 122° F (50° C) at midday, but they drop below 76° F (24° C) at night. Winters are cooler, with daily average temperatures about 55° F (13° C). The prevailing dry, northwesterly winds, known as *shamāls*, bring some relief from the heat but also produce dust storms, especially in July. Rainfall averages about five inches (130 millimetres) per year.

Traditional neighbourhoods. The crowded houses and shops of the old city, which were not for the most part of great antiquity or interest, and the labyrinthine lanes and alleys are being replaced by wide streets and new office blocks, stores, and hotels so that the appearance of the old city was gradually changing in the 1970s. Copper, silver, cloth, and other bazaars remain, as do many old buildings, sidewalk cafes, and characteristic minarets and mosques. On the west side of the river is the parliament building, as well as the foreign embassies at the capital. Such principal suburbs as al-Karkh, Karradet Mariam, and al-Manṣūr have their own bazaars and shops, and residential garden suburbs have grown up on both sides of the river. The completion of the Tigris barrage at Sāmarrā' in 1956 relieved Baghdad of the danger of disastrous floods, which caused severe damage on many occasions in history. Flood control made possible an outward expansion of the city, and its new city plan is circular in shape, encompassing both sides of the river.

Changing appearance and expansion

Transportation. Transportation within the city is provided by buses and taxis. Motor traffic has increased sharply since about 1950. The city, which grew up at the intersection of the ancient trade routes from Persia and the Far East down the Diyālā Valley to Baghdad and then up the Euphrates Valley to Syria and on to the Mediterranean or up the Tigris Valley to Armenia and the Black Sea, is still the communications centre of the region. In the 19th century a number of plans were put forward to improve the communications of Baghdad. A survey of the Euphrates and the Lower Tigris was made in 1836, and in 1860 a regular steamer service was established on the Tigris between Baghdad and Basra. By the mid-20th century, river traffic had declined in importance, although in the 1970s there was still some barge traffic on the Tigris between Baghdad and Basra.

The intersection of ancient trade routes

The three major railway lines of the state-owned and operated system all meet at Baghdad. A line runs north to Mosul; not until 1940 was the capital finally linked with Europe by rail when the Baghdad Railway across Syria to Istanbul was completed. One line connects Baghdad with Basra, and another, from Irbīl through Kirkūk, connects it with the northeast provinces.

A regular motor service between Baghdad and Damascus was established in 1923, and in the second half of the 20th century the city was served by highways, most of them following the traditional routes, connecting it to all of Iraq's major cities. Air-conditioned motor coaches transport many of the pilgrims from Iran. The Cairo–Baghdad airmail service was inaugurated in 1921 and was followed by passenger service in 1929. The city has two civilian airports. Its international airport was dedicated in 1970.

Demography. The population of the city in 1970 was over 2,100,000, compared with 1,600,000 in 1965. The rapid growth of the city has resulted in a shortage of housing for recent arrivals, in spite of low-cost public housing projects. The overwhelming majority of the population are Arab. Minor elements include Kurds, Lurs, Afghans, Armenians, Iranians, Sabaeans, and others. The principal language is Arabic, the official language of Iraq. The principal religion is Islām, with the Muslims divided between the two major sects of Sunnī and Shīʿah. The mosques and tombs at Kāẓimayn are Shīʿite shrines, but Baghdad also has a large population of Sunnites. There are a sizable number of Christians. Most of the former Jewish community have migrated to other countries.

Housing and architecture. As in other rapidly growing cities, the influx of population has resulted in housing shortages, and the construction of housing projects has not kept up with housing needs. Such large-scale construction projects as the straightening and widening of streets and the erection of large, modern office and commercial buildings have not only reduced the number of housing units in some areas but have also attracted large numbers of workers seeking jobs. Fortunately, flood control is permitting the city to expand.

The architecture of the city ranges from ancient oriental bazaars and alleys to modern steel, glass, and stone hospitals and hotels. Examples of 13th-century architecture surviving from the ʿAbbāsid period include the ʿAbbāsid palace and the Mustanṣirīyah, a large law college built and endowed by the caliph al-Mustanṣir in 1232; both were restored as museums. Another notable group of buildings surviving from the 14th century includes a mosque and college built in 1358 by Mirjān ibn ʿAbd Allāh, a Jalāyirid governor of Baghdad, together with a fine vaulted *khān* (caravansary, or inn) that is now the Museum of Arab Antiquities. The 100 or more mosques and minarets, including the spectacular gold-domed mosque at Kāẓimayn, completed in the 19th century, add to the architectural interest of the city. Modern structures include the royal mausoleum of King Fayṣal I, founder of the monarchy, the White Palace (government guesthouse), city hall, the ministry of defense building, the Royal Bilat (now the Republican Palace), the buildings of al-Hikma University at Zaʿfarānīyah, and the New Iraq Museum.

Economic life. After World War II Iraq's oil revenues considerably increased, and the new wealth of the country is reflected in the capital, the merchants of which are concerned with the import and distribution and, to a lesser but growing extent, with the manufacture of capital and consumer goods. Most Iraqi industries are in Baghdad and include leather, silks, cotton stuffs, bricks, cement, tobacco products, and the distilling of the alcoholic beverage arrack (from dates and grapes). There are also railway workshops and a steel mill. The country's financial services also are concentrated in the capital, including the Central Bank of Iraq, which has the sole right of issuing currency.

Concentration of industries

Government and services. The city has a military governor and a mayor (*amīn ʿāl-ʿāṣimah*, "guardian of the capital"); the province also has a governor. Baghdad is also the site of the national government and the headquarters of its principal organs and agencies.

Educational facilities were rapidly expanded after the revolution and in Baghdad more nearly achieved the aim of universal compulsory education. Higher education is available free at the University of Baghdad (founded 1958). Also in Baghdad are al-Hikma University of Baghdad (1956), al-Mustansiriya University (1963), four colleges, including the Higher Technical Institute (1960), and the Institute of Fine Arts (1936).

The capital also has been the first to benefit from new housing programs for the poor and from the expansion of health and welfare services and facilities, including new hospitals and clinics. The city's electric-power plants and water-supply facilities are state owned and operated.

Cultural life. As may be inferred from the city's history and its many mosques and minarets, much of its cultural life has centred on its religious sects. The tombs of the two *imams* and the adjacent mosque at Kāẓimayn are among the most important places of Shīʿah pilgrimage.

The Kāẓimayn tombs and mosque

Libraries include al-Awqāf (Library of Waqfs; founded 1929), with collections of Arabic history and literature, and Central Library of the University of Baghdad (1960), which is the depository for Iraqi publications and serves as Iraq's International Exchange Centre. Other cultural institutions include the ʿAbbāsid Palace Museum (1935), the Arms Museum (1940), the Costumes and Ethnographic Museum (1941), The Iraq Museum (1923, reformed 1966), the Iraq Natural History Museum (1946), the Museum of Arab Antiquities (1937), and the National Museum of Modern Art (1963). There are also a number of learned societies in the arts and sciences.

The media. Baghdad had half a dozen or more major daily papers in the 1970s, az-Zaman being generally regarded as the most important. In addition there are political weeklies and periodical technical journals and government publications on education and other subjects. English being the most widely known foreign language, British and U.S. newspapers and periodicals also are read in the capital.

Radio and television have been nationally owned since 1967. Radio Baghdad is Iraq's only station and broadcasts to the entire country over several frequencies and in several languages. Baghdad's television station, the first in any Arabian country, went into operation in 1956.

BIBLIOGRAPHY. Up-to-date literature on Baghdad is sparse, although there are a number of works covering the city's history, such as GUY LE STRANGE, *Baghdad During the Abbasid Caliphate: From Contemporary Arabic and Persian Sources* (1900, reprinted 1924); STEPHEN H. LONGRIGG, *Four Centuries of Modern Iraq* (1925) and *Iraq, 1900 to 1950* (1953); NABIA ABBOTT, *Two Queens of Baghdad, Mother and Wife of Harūn al-Rashīd* (1946); and TAHA BAQIR, *Baghdad* (1959). The early history of modern transportation is covered in MAYBELLE REBECCA CHAPMAN, *Great Britain and the Baghdad Railway, 1888–1914* (1948). ROBERT M. ADAMS, *Land Behind Baghdad* (1965), is a technical study concerning the changing patterns of irrigation, agriculture, and urban settlement. An intimate account of Baghdad itself is given by FREYA STARK in *Baghdad Sketches* (1937).

(E.E.D.M.O.)

Bahā'ī Faith

Bahā'ī faith is a religion founded by Mīrzā Husayn ʿAli (1817–1892; known as Bahā' Allāh, or Bahaullah, Glory of God). The word Bahā'ī derives from *bahā* ("glory, splendour") and signifies a follower of Bahā' Allāh. The religion stemmed from the Bābī faith—founded in 1844 by Mīrzāʿ Alī Moḥammad of Shīrāz, known as the Bāb—which emphasized the forthcoming appearance of "Him Whom God Shall Make Manifest," a new prophet or messenger of God. The Bābī faith in turn had sprung from Shīʿah Islām which believed in the forthcoming return of the 12th *imām* (successor of Muḥammad), who would renew religion and guide the faithful. This messianic view was the basis of the teachings of the Shaykhī sect, so named after Shaykh Aḥmad al-Ahsā'ī. Shaykh Aḥmad and his successor, Sayyid Kāẓim Rashtī, abandoned traditional literalism and gave allegorical interpretations to doctrines such as resurrection, the Last Judgment, and the return of the 12th *imām*. They and their followers (known as Shaykhīs) came to expect the appearance of the Qā'im ("He Who Arises," the 12th *imām*) in the immediate future.

On May 22, 1844, in Shīrāz, Persia, a young descendant of Muḥammad, Sayyid ʿAlī Muḥammad, proclaimed to a learned Shaykhī divine, Mullā Ḥusayn al-Bushrū'ī, that

he was the expected Qā'im, whereupon Mullā Ḥusayn became the first disciple of Sayyid ʿAlī Muḥammad, who assumed the title of the Bāb ("gate," or channel of grace from someone still veiled from the sight of men).

Soon the teachings of the Bāb, the principal of which was the tidings of the coming of "Him Whom God Shall Make Manifest," spread all over Persia, provoking strong opposition on the part of the clergy and the government. The Bāb was arrested and, after several years of incarceration, condemned to death. In 1850 he was brought to Tabriz, where he was suspended by ropes against a wall in a public square. A regiment of several hundred soldiers fired a volley. When the smoke cleared, the large crowd that had gathered at the place of execution saw ropes cut by bullets but the Bāb had disappeared. He was found unhurt in an adjacent building, calmly conversing with a disciple. The execution was repeated, this time effectively. There followed large-scale persecutions of the Bābīs in which ultimately more than 20,000 people lost their lives.

History and extent. Bahā' Allāh, who had been an early disciple of the Bāb, was arrested in connection with an unsuccessful attempt on the life of the *shāh* of Persia, Nāṣer od-Dīn, made in August 1852 by two Bābīs intent upon avenging their master. Though Bahā' Allāh had not known of the plot, he was thrown into the Black Pit, a notorious jail in Teheran, where he became aware of his mission as a messenger of God. He was released in January 1853 and exiled to Baghdad. There Bahā' Allāh's leadership revived the Bābī community, and an alarmed Persian government urged the Ottoman government to move both Bahā' Allāh and the growing number of his followers farther away from Persia's borders. Before being transferred to Constantinople, Bahā' Allāh spent 12 days in a garden on the outskirts of Baghdad, where in April 1863 he declared to a small number of Bābīs that he was the messenger of God whose advent had been prophesied by the Bāb. From Constantinople, where Bahā' Allāh spent some four months, he was transferred to Adrianople. There he made a public proclamation of his mission in letters ("tablets") addressed to the rulers of Persia, Turkey, Russia, Prussia, Austria, and Britain, as well as to the pope and to the Christian and Muslim clergy collectively.

Role of Bahā' Allāh

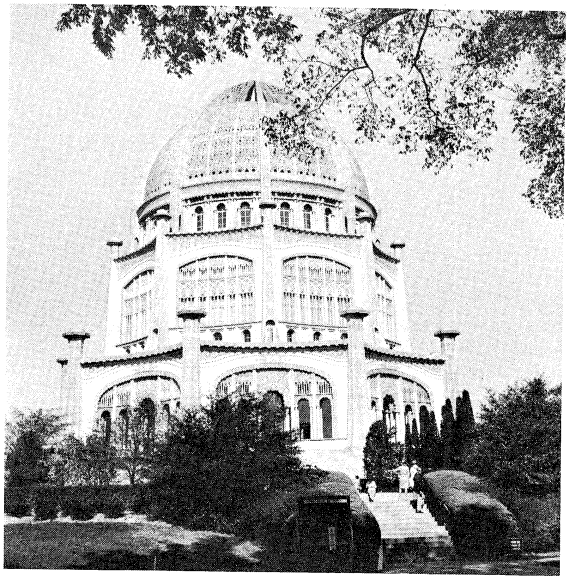
An overwhelming majority of the Bābīs acknowledged Bahā' Allāh's claim and thenceforth became known as Bahā'īs. A small minority followed Bahā' Allāh's half brother, Mīrzā Yaḥyā Ṣobḥ-e Azal, creating a temporary breach within the ranks of the Bābīs. Embittered by his failure to win more than a handful of adherents, Mīrzā Yaḥyā, assisted by his supporters, provoked the Turkish government into exiling Bahā' Allāh to ʿAkkā (ʿAkko, Acre), Palestine. He became, however, a victim of his own intrigues and was himself exiled to Cyprus.

For almost two years, Bahā' Allāh, his family, and a number of disciples were confined in army barracks converted into a jail. One of his sons and several companions died. When the severity of the incarceration abated, Bahā' Allāh was permitted to reside within the walls of ʿAkkā and later in a mansion near the town. Before his life ended in 1892, Bahā' Allāh saw his religion spread beyond Persia and the Ottoman Empire to the Caucasus, Turkistan, India, Burma, Egypt, and the Sudan.

Bahā' Allāh appointed his eldest son, ʿAbd ol-Bahā' ("Servant of the Glory," 1844–1921), as the leader of the Bahā'ī community and the authorized interpreter of his teachings. ʿAbd ol-Bahā' not only administered the affairs of the movement from Palestine but also actively engaged in spreading the faith, travelling in Africa, Europe, and America from 1910 to 1913. ʿAbd ol-Bahā' appointed his eldest grandson, Shoghi Effendi Rabbānī (1896–1957), as his successor, Guardian of the Cause, and authorized interpreter of the teachings of Bahā' Allāh, thus assuring the continued unity of the believers.

Expansion of the Bahā'ī movement

During ʿAbd ol-Bahā''s ministry, Bahā'ī groups were established in North Africa, the Far East, Australia, and the United States. Since then the movement has spread to virtually every country in the world, with particularly large and vigorous communities in Africa, Iran, India, the United States, and certain areas of Southeast Asia and the Pacific. No official membership statistics for the entire Bahā'ī community are available. In 1971, however, Bahā'īs resided in more than 50,000 localities throughout the world, with more than 100 national spiritual assemblies (national governing bodies) and no fewer than 6,000 local spiritual assemblies. A current plan of worldwide expansion envisages the formation of 120 national spiritual assemblies and 13,833 local spiritual assemblies by April 1973 (600 of these in the United States). Bahā'ī literature has been translated into more than 400 languages. By 1970 more than 300 African tribes, some 100 American Indian tribes, and nearly 100 tribes and peoples of the Indian subcontinent and the Pacific Ocean were represented in the Bahā'ī community. In the 1960s and early 1970s the Bahā'ī faith was undergoing a period of extremely rapid expansion.



Bahā'ī House of Worship, Wilmette, Illinois.
By courtesy of the Chicago Convention and Tourism Bureau

Sacred literature. Bahā'ī sacred literature consists of the total corpus of the writings of Bahā' Allāh and their interpretation and amplification in the writings of ʿAbd ol-Bahā' and Shoghi Effendi. Bahā' Allāh's literary legacy of more than 100 works includes *al-Kitāb al-Aqdas* ("The Most Holy Book"), the repository of his laws; the *Ketāb-e Īqān* (*The Book of Certitude*), an exposition of essential teachings on the nature of God and religion; *The Hidden Words*, a collection of brief utterances aimed at the edification of men's "souls and the rectification of their conduct"; *The Seven Valleys*, a mystic treatise that "describes the seven stages which the soul of the seeker must needs traverse ere it can attain the object of its existence"; "Epistle to the Son of the Wolf," his last major work; as well as innumerable prayers, meditations, exhortations, and epistles. The Bahā'īs believe that the writings of Bahā' Allāh are inspired and constitute God's revelation for this age.

Religious and social tenets. Bahā' Allāh teaches that God is unknowable and "beyond every human attribute, such as corporeal existence, ascent and descent, egress and regress." "No tie of direct intercourse can possibly bind Him to His creatures . . . . No sign can indicate His presence or His absence . . . ." Human inability to grasp the divine essence does not lead to agnosticism, since God has chosen to reveal himself through his messengers, among them Abraham, Moses, Zoroaster, Buddha, Jesus, Muḥammad, and the Bāb, who "are one and all the Exponents on earth of Him Who is the central Orb of the universe . . . ." The messengers, or, in Bahā'ī terminology, "manifestations," are viewed as occupying two "stations," or occurring in two aspects. The first "is the station of pure abstraction and essential unity," in which one may speak of the oneness of the messengers of God because they are all manifestations of his will and exponents of his word. This does not constitute syncretism,

since "the other station is the station of distinction . . . . In this respect, each manifestation of God hath a distinct individuality, a definitely prescribed mission . . . ." Thus, while the essence of all religions is one, each has specific features that correspond to the needs of a given time and place and to the level of civilization in which a manifestation appears. Since religious truth is considered relative and revelation progressive and continuing, the Bahā'īs maintain that other manifestations will appear in the future, though not, according to Bahā' Allāh, before the expiration of a full thousand years from his own revelation.

In Bahā'ī teachings, God is, and has always been, the Creator. Therefore, there was never a time when the cosmos did not exist. Man was created through God's love: "Veiled in My immemorial being and in the ancient eternity of My essence, I knew my love for thee: therefore I created thee." The purpose of man's existence as taught by Bahā' Allāh is to know and to worship God and "to carry forward an ever-advancing civilization . . . ." Man, whom Bahā' Allāh calls "the noblest and most perfect of all created things," is endowed with an immortal soul, which, after separation from the body, enters a new form of existence. Heaven and hell are symbolic of the soul's relationship to God. Nearness to God results in good deeds and gives infinite joy, while remoteness from him leads to evil and suffering. To fulfill his high purpose, man must recognize the messenger of God within whose dispensation he lives and "observe every ordinance of him who is the desire of the world. These twin duties are inseparable. Neither is acceptable without the other."

Goals of Bahā'ī faith

Civilization, Bahā' Allāh teaches, has evolved to the point where unity of mankind has become the paramount necessity. The Bahā'ī faith, in the words of Shoghi Effendi,

> proclaims the necessity and the inevitability of the unification of mankind, asserts that it is gradually approaching, and claims that nothing short of the transmuting spirit of God, working through His chosen Mouthpiece in this day, can ultimately succeed in bringing it about. It, moreover, enjoins upon its followers the primary duty of an unfettered search after truth, condemns all manner of prejudice and superstition, declares the purpose of religion to be the promotion of amity and concord, proclaims its essential harmony with science, and recognizes it as the foremost agency for the pacification and the orderly progress of human society. It unequivocally maintains the principle of equal rights, opportunities and privileges for men and women, insists on compulsory education, eliminates extremes of poverty and wealth, abolishes the institution of priesthood, prohibits slavery, asceticism, mendicancy, and monasticism, prescribes monogamy, discourages divorces, emphasizes the necessity of strict obedience to one's government, extols any work performed in the spirit of service to the level of worship, urges either the creation or the selection of an auxiliary international language, and delineates the outlines of those institutions that must establish and perpetuate the general peace of mankind.

Practices. Membership in the Bahā'ī community is open to all who profess faith in Bahā' Allāh and accept his teachings. There are no initiation ceremonies, no sacraments, and no clergy. Every Bahā'ī, however, is under the spiritual obligation to pray daily; to fast 19 days a year, going without food or drink from sunrise to sunset; to abstain totally from narcotics, alcohol, or any substances that affect the mind; to practice monogamy; to obtain the consent of parents to marriage; and to attend the Nineteen Day Feast on the first day of each month of the Bahā'ī calendar. The Nineteen Day Feast, originally instituted by the Bāb, brings together the Bahā'īs of a given locality for prayer, the reading of scriptures, the discussion of community activities, and for the enjoyment of one another's company. The feasts are designed to ensure universal participation in the affairs of the community and the cultivation of the spirit of brotherhood and fellowship. Eventually, Bahā'īs in every locality plan to erect a house of worship around which will be grouped such institutions as a home for the aged, an orphanage, a school, and a hospital. In the early 1970s, houses of worship existed in Wilmette, Illinois; Frankfurt am Main, West Germany; Kampala, Uganda; Sydney Australia; and one was being built in Panama. In the temples there is no preaching; services consist of recitation of the scriptures of all religions.

Houses of worship

The Bahā'īs use a calendar established by the Bāb and confirmed by Bahā' Allāh, in which the year is divided into 19 months of 19 days each, with the addition of four intercalary days (5 in leap years). The year begins on the first day of spring, March 21, which is a holy day. Other holy days on which work is suspended are the days commemorating the declaration of Bahā' Allāh's mission (April 21, April 29, and May 2), the declaration of the mission of the Bāb (May 23), the birth of Bahā' Allāh (November 12), the birth of the Bāb (October 20), the passing of Bahā' Allāh (May 29), and the martyrdom of the Bāb (July 9).

Organization and administration. The Bahā'ī community is governed according to general principles proclaimed by Bahā' Allāh and through institutions created by him that were elaborated and expanded by 'Abd ol-Bahā'. These principles and institutions constitute the Bahā'ī administration order, which the followers of the faith believe to be a blueprint of a future world order. The governance of the Bahā'ī community begins on the local level with the election of a local spiritual assembly. The electoral process excludes parties or factions, nominations, and campaigning for office. The local spiritual assembly has jurisdiction over all local affairs of the Bahā'ī community. On the national scale, each year Bahā'īs elect delegates to a national convention that elects a national spiritual assembly with jurisdiction over the entire country. All national spiritual assemblies of the world periodically constitute themselves an international convention and elect the supreme governing body known as the Universal House of Justice. In accordance with Bahā' Allāh's writings, the Universal House of Justice functions as the supreme administrative, legislative, and judicial body of the Bahā'ī commonwealth. It applies the laws promulgated by Bahā' Allāh and legislates on matters not covered in the sacred texts. The seat of the Universal House of Justice is in Haifa, Israel, in the immediate vicinity of the shrines of the Bāb and 'Abd ol-Bahā', and near the shrine of Bahā' Allāh at Bahjī near 'Akkā.

Universal House of Justice

There also exist in the Bahā'ī faith appointive institutions, such as the Hands of the Cause of God and the continental counsellors. The former were created by Bahā' Allāh and later assigned by 'Abd ol-Bahā' the functions of propagating the faith and protecting the community. The Hands of the Cause appointed by Shoghi Effendi in his lifetime now serve under the direction of the Universal House of Justice. The continental counsellors perform the same functions as the Hands of the Cause but are appointed by the Universal House of Justice. Assisting the counsellors in advising, inspiring, and encouraging Bahā'ī institutions and individuals are auxiliary boards appointed by the counsellors and serving under their direction.

BIBLIOGRAPHY. The classic introduction to the Bahā'ī faith, giving a general view of its history and teachings, is J.E. ESSLEMONT, *Bahá'u'lláh and the New Era*, 3rd rev. ed. (1970). GEORGE TOWNSHEND, *The Promise of All Ages*, rev. ed. (1948, reprinted 1957), approaches the Bahā'ī faith from a background of Christianity. The history of the Bahā'ī faith has been studied by many scholars, but the most detailed and poetic account is *The Dawn-Breakers* by MUHAMMAD-I-ZARANDI, surnamed Nabíl, trans. and ed. by SHOGHI EFFENDI (1932, reprinted 1970; 2nd ed., 1953); the latter's *God Passes By* (1944), recounts to the end of the first Bahā'ī century. The most important source for the study of the Bahā'ī faith is the writings of Bahā' Allāh and their interpretation and application by 'Abd ol-Bahā and Shoghi Effendi. Several of Bahā' Allāh's major works are available in excellent English translations. The *Kitáb-i-íqán* (1950) is indispensable for understanding Bahā'ī views of God, progressive revelation, and the nature of religion. *The Hidden Words*, rev. ed. (1954, reprinted 1970), and *The Seven Valleys, and the Four Valleys*, rev. ed. (1952, reprinted 1968), deal with man's spiritual life and the states of the soul. *Gleanings from the Writings of Bahá'u'lláh* (1951) is a representative selection. ABD OL-BAHA'S *Some Answered Questions*, rev. ed. (1964), is a record of table talks on various religious themes. *The Secret*

of Divine Civilization (1957) uses the problem of modernization and development to set forth the spiritual prerequisites of true progress and civilization. SHOGHI EFFENDI's writings include *The World Order of Bahá'u'lláh* (1955), an exposition of principles for the establishment of universal peace and world civilization; and *The Promised Day Is Come* (1961), an examination of the effects of manifestation upon the modern world.

(F.Ka.)

Bahamas

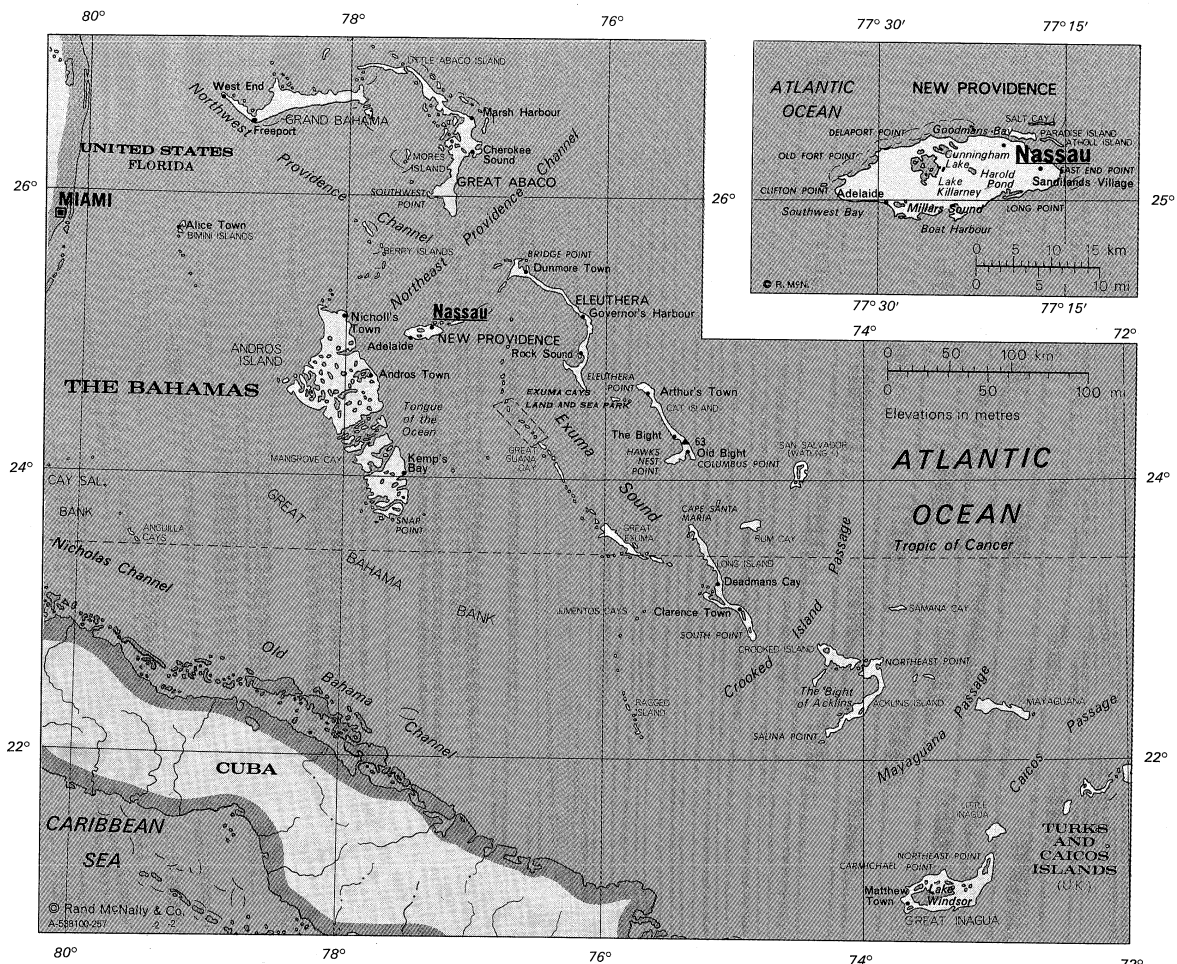
The island group of the Bahamas (Spanish *bajamar*, "shallow water"), a former British colony, since 1973 an independent nation within the Commonwealth of Nations, occupies an irregular submarine tableland that rises out of the Atlantic depths and is separated from nearby lands to the south and west by deepwater channels. Lying to the north of Cuba and Hispaniola, the archipelago comprises nearly 700 islands and cays (small islands), only 22 of which are occupied, and almost 2,400 low, barren rock formations. It stretches for 760 miles (1,220 kilometres) southeasterly from Grand Bahama Island, which lies about 60 miles off the southeastern coast of Florida, to Great Inagua Island, some 50 miles from the eastern tip of Cuba. The total land area is 5,382 square miles (13,939 square kilometres).

Strategic location

The strategic position of the Bahamas, which lie at the geographic centre of the New World landmass, commanding the gateway to the Gulf of Mexico, the Caribbean Sea, and the entire Central American region, has given the history of the islands a unique and often striking character. It was here that Christopher Columbus made his original landfall in the Americas. The subsequent fate of the peaceful original inhabitants remains one of the more tragic episodes in the development of the entire region, while the early attempts at European-dominated settlement were marked by intense national rivalries, interspersed with long periods of lawlessness and piracy. As a result, the society and culture that has evolved in the Bahamas is a distinctive blend of European and African heritages, the latter a legacy of the slave trade. The islands, lacking natural resources other than their magnificent climate and dazzling beaches, have become heavily dependent on the income generated by the extensive tourist facilities that have been developed, often as a result of the injection of foreign capital. The continued popularity of the islands, largely with North American tourists, has maintained a relatively high standard of living among the population, 85 percent of whom are black, but difficulties associated with the transfer of political control to this majority group caused some economic problems from the late 1960s onward.

The capital city, Nassau, is situated on the small (80 square miles [207 square kilometres]) but important New Providence Island. Other islands, known collectively as the Out Islands, include Grand Bahama, which contains the major settlements of Freeport and West End; Andros (2,300 square miles [5,957 square kilometres]), the largest island; and Eleuthera, site of one of the early attempts at colonization. In spite of the concentration of the population in urban centres devoted to tourism, the traditional pattern of small farming and fishing prevails in many villages, notably in the southeastern islands.

The natural environment. *Physical features and climate.* Extensive areas of flat land, generally a few feet in elevation, are the dominant topographic features of the major islands; Andros, for example, has no ground point higher than 71 feet (22 metres). A number of islands fronting the Atlantic have a range or series of ranges of hills on their northeastern side and parallel to the longer axis of the island. These are formed of sand washed ashore and blown inland by the trade winds. The newer hills adjacent to the seashore are normally sand dunes. Solidity increases toward the interior, where the particles become cemented to form Bahama lime-



THE BAHAMAS

stone. Eleuthera and Long Island have the greatest number of hills exceeding 100 feet, and the highest point, 206 feet (63 metres) is on Cat Island. Beneath the soil, the islands are composed of the skeletal remains of coral and other marine organisms.

Temperature and rainfall

The Bahamian climate, mild throughout the year, is one of the great attractions of the area. The average temperature varies from 70° F (21° C) during the winter to 81° F (27° C) during the summer, and extremes very seldom fall below 60° F (16° C) or rise above 90° F (32° C). The average annual rainfall is about 46 inches, occurring mostly during the summer months. Prevailing winds, coming from the northeast in winter and southeast in summer, lend a cooling influence to a generally humid atmosphere. Hurricanes present something of a threat during the period from mid-July to mid-November and have occasionally caused great destruction.

Vegetation and animal life. Extensive and beautiful forests of Caribbean pine are found on Grand Bahama, Great Abaco, Andros, and New Providence. Hardwood forests, known locally as "coppices," also occur on some of the islands. Elsewhere, the woody vegetation consists mostly of shrubs and low trees. Animal life is dominated by frogs, lizards, and snakes, all nonpoisonous, and several species of bats are found in caves along the more rocky coasts. Larger animals include the agouti, a rodent; the raccoon; the iguana; and the elegant flamingo, the national bird. All of these have been much reduced in numbers and in distribution: the barkless dogs reputed to have existed at the time of the original inhabitants may indicate an early link with Africa. In addition, several animals—notably sheep, horses, and other livestock—have been introduced from Europe. The surrounding waters abound with fish and other edible marine animals.

History. It is widely held that on October 12, 1492, Christopher Columbus first stepped ashore in the New World on San Salvador (Watling's Island). He explored this and nearby islands and then sailed to Cuba and Hispaniola. Within 25 years of discovery, the entire indigenous people—Arawak Indians who called themselves Lucayans—had been carried off by the Spaniards to slavery and death in Hispaniola. The islands remained depopulated for more than a century.

In 1648 a band of about 70 English Puritans known as Eleutheran Adventurers, under the governorship of Capt. William Sayle, established a colony on Eleuthera. They had high hopes of establishing a flourishing plantation and an advanced form of government, but unproductive soil, internal discord, and Spanish interference reduced their ambitions to a desperate struggle for survival. Meanwhile, settlers from Bermuda and elsewhere drifted into New Providence. By 1670, when the Bahama Islands were granted to the proprietors of the Carolina colony, New Providence had passed Eleuthera in population and commercial importance, and it became the seat of government. The proprietors took little interest in their colony, however, and it soon became a haven for pirates, whose depredations against Spanish ships provoked frequent and savage retaliatory raids.

Defeat of the pirates

By 1718 it became obvious that proprietary government was a failure, and Capt. Woodes Rogers was sent out as first royal governor to exterminate the pirates and establish conditions attractive to law-abiding settlers. His nearly bloodless conquest of about 1,000 of the brigands set the colony on a course of peaceful progress. During the American Revolution a great many Loyalists fled with their slaves to the islands, doubling the white population and trebling the black. The cotton plantations that they developed yielded well for a few years, but exhausted soil, insect pests, and, finally, abolition of slavery in 1834 led to their ultimate collapse.

Considerable wealth poured into the islands as the result of blockade-running during the American Civil War and the handling of liquor during the Prohibition era of the 1920s in the United States. This kind of activity made no lasting contribution to the islands, however, nor did it establish any firm economic base. Before and after these periods, many attempts were made to grow pineapples, citrus fruits, tobacco, tomatoes, and sisal for export, but despite initial promise, all failed. Through this, however, a majority of Bahamians took to the sea, where searching for wrecks, sponging, and fishing offered excitement and occasional good fortune. After World War II, strenuous efforts to establish tourism as the basis of the economy were strikingly successful, transforming the economic and social structure of the islands.

Politically, Bahamians have had considerable control over their affairs since Captain Rogers gathered the islands' first assembly, in 1729. Constitutional advances in 1964 and 1969 brought the country to the verge of complete self-government. Party politics emerged in 1953, when the Progressive Liberal Party (PLP) was formed by blacks to oppose the group in power, who in 1958 responded with a party of their own, the white-controlled United Bahamian Party (UBP). As the political battle progressed, the PLP raised the cry for majority rule, and many of the acrimonious characteristics of a racial contest were introduced. The climax came after the general elections of 1967, when the PLP was able to form a government with a slight majority. Elections the following year gave that party 29 of the 38 Assembly seats. While there is no doubt that the UBP was defeated mainly on the racial issue, the accusation that that party had introduced criminal-controlled gambling and that some of its members had profited thereby had a telling effect.

In general the PLP advocated stricter government control of the economy, increased Bahamian ownership of business enterprises, and the replacement of foreign workers by Bahamians. Although the move toward self-government received bipartisan support, the UPB advocated that total independence should come later than 1973, the target year of the PLP government. In 1969 the name of Commonwealth of the Bahama Islands was adopted, and on July 10, 1973, on independence, the official form became the Commonwealth of the Bahamas.

The people and the conditions of life. In 1970 there were almost 170,000 permanent inhabitants of the Bahamas, 27,000 of them foreign-born. English is the only language native to Bahamians. In recent years, however, there has been an influx of Haitian labourers who speak French or its creole dialect among themselves. About 15 percent of the population are descendants of English pioneer settlers and Loyalist refugees. The rest are mainly of African descent, many of them with varying amounts of Caucasian blood. The Baptists, with a membership of 49,000, are the largest religious denomination. Episcopalians and Roman Catholics are next in order, with approximately 38,000 each. The Methodist and Evangelical churches claim 12,000 and 10,000, respectively.

Economy. Except for some small, regional pockets of unemployment, Bahamians enjoy full employment and a relatively high standard of living. In 1968 the gross national product was about $227,000,000 and the per capita income $1,260. Economic growth has been stimulated almost entirely by private enterprise, largely by American capital. Traditionally the government has played a passive role, confining its activities to the maintenance of orderly business and social activities. Blacks have every field of opportunity open to them and have made great strides toward economic parity with whites.

Standard of living

Tourism continues to dominate the economy. It is responsible directly and indirectly for more than 70 percent of the gross national product and the employment of almost two-thirds of the manpower. Gambling casinos in several areas add to the attraction of the Bahamas for tourists, 1,300,000 of whom visited the islands in 1970. During the 1960s there was a widespread move away from fishing and farming villages to the centres of tourist activity, with over 17,000 people changing their residence. Of these, 35 percent went to New Providence and 38 percent to Grand Bahama, where Freeport was under development as a deepwater harbour and a tax-exempt tourist, industrial, and commercial area.

Banking and business hold a distant second place in the economy, followed by government and construction. Other industries include a petrochemical refinery and a cement plant at Freeport, which has become second only to Nassau in tourist and commercial activity. Salt is pro-

Bahamas, Area and Population

	area*		population†	
	sq mi	sq km	1963 census	1970 census
Island and groups‡				
Abaco, Great and Little, Moore's Island and cays	649	1,681	6,500	6,500
Acklins Island	192	497	1,200	900
Andros Island	2,300	5,957	7,500	8,800
Berry Islands	12	31	300	400
Biminis (North and South), Cay Lobos, and Cay Sal	11	28	1,700	1,500
Cat Island	150	388	3,100	2,700
Crooked Island	84	218	800	700
Eleuthera, Harbour Island, and Spanish Wells	200	518	9,100	9,500
Exuma, Great and Little, and cays	112	290	3,400	3,800
Grand Bahama	530	1,373	8,200	25,900
Inagua, Great and Little	599	1,551	1,200	1,100
Long Cay	9	23	20	30
Long Island	230	596	4,200	3,900
Mayaguana	110	285	700	600
New Providence	80	207	80,900	101,500
Ragged Island and cays	14	36	400	200
San Salvador and Rum Cay	90	233	1,000	900
Total Bahamas	5,382	13,939	130,200§	168,800§

*Includes certain areas of inland water; total includes land area not otherwise accounted for. †Rounded. ‡The Bahamas have no first-order administrative subdivisions. However, they are divided into 38 electoral districts, corresponding to portions of islands, islands, or groups of islands. §Figures do not add to total given because of rounding.
Source: Official government figures.

duced by solar evaporation at Inagua and Long Island and various forms of food production continue at numerous places for domestic consumption.

Since there is no direct taxation, customs duties produce the major share of revenue, in 1970 accounting for about three-fifths of the government's income.

Transportation. There are modern paved roads in Nassau and Freeport and their environs. Elsewhere, roads are constructed of crushed limestone and sometimes surfaced with asphalt, but increasingly they are capable of accommodating motor vehicles. A fleet of small motor vessels carries passengers, freight, and mail weekly between Nassau and the Out Islands. The deep-water harbours of Nassau and Freeport are dredged to depths of 37 feet and 30 feet, respectively. Nassau was visited by about 2,500 foreign ships in 1970, both passenger and freight. The airplane has become of increasing importance to the Bahamian economy. Throughout the islands there are 46 airports, with varying accommodations and facilities. Most of these serve only interinsular aircraft, but international airports are located at Nassau, Freeport, and West End.

Government and services. The government is patterned after that of Great Britain. The governor general since 1973 is appointed by the English monarch, and in turn, he appoints a prime minister who must be a member of the House of Assembly and must be able to command a majority of Assembly votes. The House of Assembly is composed of 38 elected members. The 16-member Senate, which has severely restricted powers, is appointed by the governor, the majority of the members on the advice of the prime minister. The life of the Senate is, like the Assembly's, normally five years, but if the prime minister is unable to control the Assembly effectively or if he considers it expedient, both bodies are dissolved and reconstituted.

Bicameral legislature

Bahamians are relatively free of malnutrition and debilitating diseases, and medical problems are largely those involving common infections. Alcoholism is causing increasing concern as the chief contributor of mental illness, and care for the aged is a mounting social problem. Severe housing congestion and hazardous sanitary conditions exist only in some black areas of New Providence, where housing is substandard and sewage is disposed of in backyard cesspits that tend to overflow. There is little illiteracy, for schooling is compulsory from 5 to 14 years of age. Public secondary and technical schools have been greatly expanded in recent years, but there are no universities.

Culture and communications. There is little indigenous culture. Outstanding among traditional group activities is the "Junkanoo" parade on Boxing Day and New Year's Day. The main thoroughfare is given over to hundreds of gaily bedecked celebrants who, with clanging cow bells and beating drums, march and dance to a "Goombay" rhythm of African origin. Island folklore includes stories of a three-toed, human-faced creature called the "Chickcharney" and the workings of Obeah, a local brand of witchcraft. In Nassau about a dozen amateur choral, dramatic, and dancing groups provide fine entertainment with much local flavour. A national trust is concerned with the preservation of wildlife, and a historical society promotes local history.

The oldest extant Bahamian newspaper is the daily *Nassau Guardian*, founded in 1844. The daily *Nassau Tribune* (1903) has the largest circulation in the islands. The only other daily is the *Freeport News*. Political organs are published periodically by various parties. The Bahamas government owns and operates two radio stations with complementary programs. The Out Islands are connected with Nassau through more than 100 radiotelephone and wireless-telegraph stations. Contact is maintained through a 180-voice circuit between Nassau and Miami, Florida, and thence with all parts of the world.

Prospects. Efforts to diversify the Bahamian economy, the country's most pressing need, are presently unpromising, since investors in industry are reluctant to establish plants where shortages of both raw materials and manpower exist. It would seem that the tourist-based economy, which lifted the people out of economic stagnation, must be relied upon for the foreseeable future. But the vulnerability of tourism to world economic conditions was demonstrated by the decrease in visitors and revenue in 1970. This points up also the close tie of the islands to the U.S.

Need to diversify

BIBLIOGRAPHY. The DEPARTMENT OF STATISTICS (NASSAU), *Statistical Abstract* (annual) gives data on population and vital statistics, tourism, trade, aviation and shipping, health, education, agriculture, and fisheries. See also the government publications: *Bahamas* (1946–63) and *Bahama Islands* (1964–67), biennial reports. Other recommended works include: G.B. SHATTUCK (ed.), *The Bahama Islands* (1905), a well-illustrated, indexed volume; N.L. BRITTON and C.F. MILLSPAUGH, *The Bahama Flora* (1920, reprinted 1962); C.B. CORY, *The Birds of the Bahama Islands* (1880); MICHAEL CRATON, *A History of the Bahamas* (1968), a well-researched description, with bibliography and index; A. DEANS PEGGS, *A Short History of the Bahamas*, 3rd ed. (1959), brief but scholarly; CLAPP AND MAYNE, INC., *A General Diagnosis of the Economy of the Bahama Islands* (1969), an exhaustive study; and CHECCHI AND COMPANY, *A Plan for Managing the Growth of Tourism in the Commonwealth of the Bahama Islands* (1969), a detailed analysis.

(E.P.A.)

Bahia

A state of eastern Brazil, Bahia is bounded northwest by Piauí, north by Pernambuco, northeast by Alagoas and Sergipe, east by the Atlantic Ocean, southeast by Espírito Santo, south by Minas Gerais, and west by Goiás. With an area of 216,613 square miles (561,026 square kilometres), it had a population of 7,583,140 at the 1970 census. The capital, Salvador, a port commanding an inlet of the Atlantic Ocean, was once commonly known also as Bahia ("Bay"), whence the state derives its name. (For an associated physical feature, see SAO FRANCISCO RIVER.)

History. On All Saints' Day, November 1, 1501, Portuguese explorers entered the bay on which Salvador now stands: they therefore named it Bahia de Toros os Santos, or All Saints' Bay. The subsequent occupation of the vicinity by the Portuguese led, in 1549, to the merging of four captaincies under the first governor general of Brazil, Tomé de Sousa who in the same year

Early exploration

founded Salvador as the seat of his government. Seized by the Dutch in 1624, the city was recovered by the Portuguese the following year.

When the Empire of Brazil was proclaimed in 1822, Bahia was still controlled by forces loyal to Portugal; but on July 2, 1823, Brazilian troops occupied Salvador, and Bahia became a province of the empire. In 1889, under the republic, Bahia became a state of the Brazilian Federation.

The colonization of the territory had begun in the Recôncavo—that is, in the coastal region—which grew sugarcane and tobacco for export and other crops for the settlers' food. In the semi-arid interior, cattle raising was considerably stimulated in the 18th century, when the discovery of gold and gems in the Diamantina Highland attracted more settlers. The 19th century saw a revival of agriculture: it was the golden age for sugarcane; coffee was also grown on a large scale; cotton production went up; and the forests of the south were turned into profitable plantations of cacao. Rubber plantations were developed at the beginning of the 20th century.

Physical geography. The Diamantina Highland and its northern extension, the Serra do Tombador, run longitudinally across Bahia from the borders of Minas Gerais and constitute the line of greatest elevation, the Diamantina reaching its maximum altitude in Pico das Almas, which is 6,068 feet (1,850 metres) in height. From this dorsal ridge depend the Western Plateau and the Eastern, which vary in altitude between about 650 and 2,600 feet and are characterized by the presence of inselbergs (isolated eminences left by erosion). The Eastern Plateau ends with the heights overlooking the coastal plain.

Major rivers

The major river is the São Francisco, which rises in Minas Gerais and flows north across western Bahia before turning eastward in a great curve to form the frontier between Bahia and Pernambuco and between Bahia and Alagoas, on its long way down to the Atlantic. Its most important waterfall is the Paulo Afonso, on the short Bahia–Alagoas border. The São Francisco's principal tributaries in Bahia are the Rãs, the Santo Onofre, and the Paramirim, on the right bank, and the Carinhanha, the Corrente, and the Grande, on the left.

The second most important river is the Paraguaçu, which has its sources in the Diamantina Highland and flows eastward into All Saint's Bay. Its basin is a cattle-raising zone, with small areas of agriculture, in an otherwise arid region. The river floods periodically.

Other rivers flowing eastward into the Atlantic are the Itapicuru and the Contas, north and south of the Paraguaçu, respectively. The Pardo and the Jequitinhonha both flow across southern Bahia from Minas Gerais.

Climate, vegetation, and animal life. Along Bahia's coastline there are areas with an annual rainfall exceeding 55 inches (1,400 millimetres), as well as sandy stretches on which the Brazilian coconut and the *mangabeira* rubber tree (*Hancornia speciosa*) flourish, while the mud of the estuaries favours mangroves. In the broad-leaved tropical forest, where the annual rainfall exceeds 60 inches (1,500 millimetres) and where the coldest month has an average temperature of 64° F (18° C), the sandy clay soil supports an evergreen vegetation. In the zone of transitional forest, where a dry season interrupts the prevalent humidity, the soils are shallower, and there is a deciduous vegetation of shrubs, smaller plants, and numerous lianas or vines. On some of the tablelands, where sandy soil under a surface crust fails to retain much water, all plant life must be of a kind adaptable to changing conditions. In the extensive

The *caatinga*

caatinga, or zone of drought (nearly 60 percent of the state's territory lies within what Brazil's geographers call "the Polygon of Droughts"), the rainy season is irregular, and the annual rainfall never exceeds 24 inches (600 millimetres). Here the landscape is generally wide open and bare, unprotected against erosion, and plants such as cactus predominate. On more sheltered expanses of the higher hinterland, at altitudes between about 2,600 and 2,900 feet, where summers of rain and long seasons of drought alternate, areas of savanna occur on soils of ferrous, clayey, or sandy types.

Peccary, tapir, and the two-toed sloth inhabit the forests. In open country the giant armadillo, the scarlet ibis, and the king vulture can be seen.

Population. According to Brazilian estimates, the population is composed as follows: whites, 30 percent; blacks, 19 percent; mulattos, 51 percent.

Population density varies considerably. More than 50 percent of the state's total population is concentrated in the littoral, and especially in the Recôncavo, which represents only 20 percent of the total area. In some *municípios* (departments) of the Recôncavo there are 520 people per square mile; and the towns are even more densely populated. Salvador, the capital, has a metropolitan area covering six *municípios*. Other major cities of the state are Ilhéus, Itabuna, Feira de Santana, and Vitória da Conquista. The arid interior, on the other hand, is sparsely populated and has few towns. Of the total population, 59 percent live in rural areas, and 41 percent are urban.

The language of the people is Portuguese, influenced to some extent by African idiom and slightly by Indian languages. Roman Catholicism is professed by 93 percent of the population, Protestantism by 2 percent, Spiritualism and other beliefs by 5 percent. Many people practice candomblé, a sort of voodoo, but declare themselves Catholics.

Administration and social conditions. The state is ruled by a governor, elected for a four-year term. The 336 *municípios* are administered by *prefeitos* (mayors) and are subdivided into *distritos* (districts).

The standard of living is low. Hygiene is defective, even in urban areas, and illiteracy is widespread, despite efforts to improve medical services and sanitation and to expand primary and secondary schooling. Salvador has two universities, the Federal (5,500 students) and the Católica do Salvador (4,300 students).

Economy. Bahia's mineral resources comprise petroleum (85 percent of Brazil's output), natural gas, lead, copper, chrome, tin, barite, manganese, magnesite, titanium, hematite, quartz, kaolin, marble, asbestos, and amethyst. There is also a hydroelectric potential: the São Francisco River has been harnessed by the Paulo Afonso Dam at its major waterfall (with a capacity of 1,000,000 kilowatts), but the Paraguaçu, the Contas, and the Jequitinhonha remain unexploited.

The most important crops are cacao, tobacco, vegetable oils, piassava, and sisal. Timber is also obtained from the forests. Cattle, which number 8,500,000 head, yield leather and skins.

Heavy industry is represented by the Landulfo Alves petroleum refinery; and by cement works and ironworks. Salvador, Feira de Santana, and Aratu are industrial centres. Energy is mostly hydroelectric, especially from the Paulo Afonso project.

The state has 63,556 miles of road and 1,550 of railway, mostly serving the Recôncavo.

Cultural life. Bahia's colonial past has left a cultural legacy as rich as that of any state of eastern Brazil—and is represented not only by churches, manorial houses, and fortresses but also by numerous works of art. Popular tradition is maintained in many festival customs.

Prominent among cultural institutions are the Academia de Letras da Bahia, the Instituto Geográfico e Histórico da Bahia, and the Instituto de Musica da Bahia.

BIBLIOGRAPHY

Physical geography: JEAN TRICART *et al.*, *Estudos de Geografia da Bahia* (1958), ranges from hydrographic basins to the study of consumer commerce.

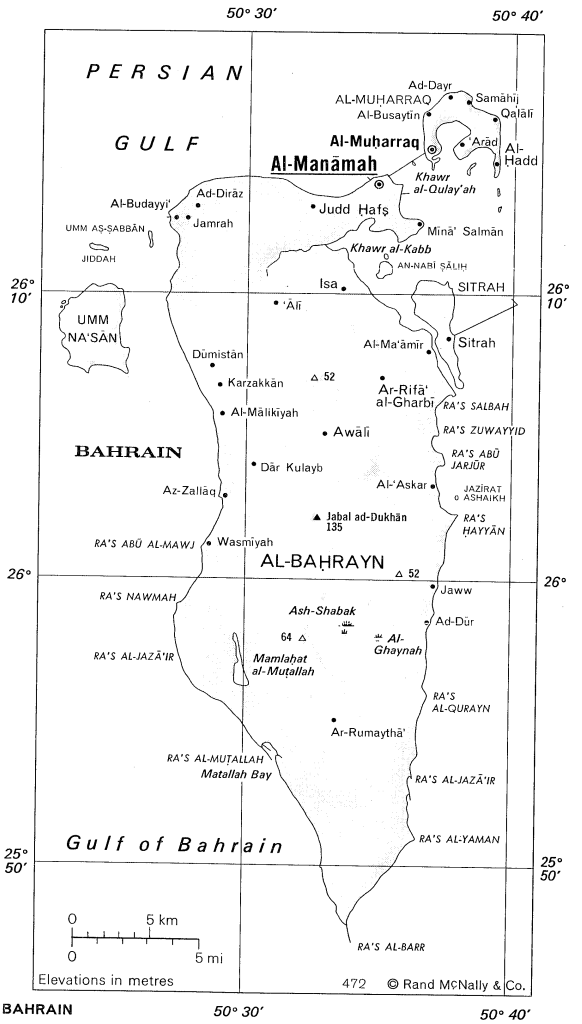
History: THALES DE AZEVEDO, *O Povoamento da Cidade do Salvador*, 3rd ed. (1969), a study of the population of Salvador; LUIZ HENRIQUE DIAS TAVARES, *História da Baía* (1959); MILTON SANTOS, *O Centro da Cidade do Salvador* (1959), an account of the evolution of Salvador and the formation of its sphere of influence; JORGE AMADO, *Terras do sem fim*, 4th ed. (1946; Eng. trans., *The Violent Land*, 1965).

Economics: CENTRO INDUSTRIAL DE ARATU, *Plano Diretor* (1968), a voluminous study of the industrial area around Salvador.

(A.D.E.)

Bahrain

Bahrain, a small Arab emirate in the Persian (Arabian) Gulf, consists of an archipelago of Bahrain Island—extending some 30 miles from north to south and 10 miles from east to west—and some 30 smaller islands. It is situated in a bay on the southern coast of the Persian Gulf between the northeastern Hasa coast of Saudi Arabia to the west and the Qatar peninsula to the east. Its Arabic name, al-Baḥrayn, means "two seas." The total land area is about 256 square miles (662 square kilometres), and its population in the early 1970s was over 200,000.



Bahrain is an independent Arab state which until 1971 had a special treaty relationship with the United Kingdom, which was responsible for the conduct of its foreign affairs and defense. In 1971 Bahrain declared itself independent and became a member of the United Nations. Though oil was discovered in Bahrain in 1932—it was the first discovery on the Arabian side of the Gulf—relatively little oil is now produced, and the emirate's commercial importance is due to its central position in the Gulf. The chief town, port, and seat of administration is al-Manāmah in the northeast of Bahrain Island. The ruler, Shaykh 'Isā ibn Sulmān Āl Khalīfah, resides at ar-Rifa'al-Gharbī, about six miles south of al-Manāmah. (For an associated physical feature see PERSIAN GULF.)

Ancient settlements

History. Bahrain has been inhabited from prehistoric times, and several thousand burial mounds in the north of the main island probably date from the Sumerian period of the 3rd millennium BC. The archipelago was mentioned by Persian, Greek, and Roman geographers and historians. It has been Arab and Muslim since the Muslim conquest of the 7th century AD, though it was occupied and ruled by the Portuguese from 1521 to 1602 and by the Persians from 1602 to 1783. Since 1783 it has been ruled by sheikhs of the Āl Khalīfah family, which originated in the Hasa province of Arabia. Several times during the 19th century the British intervened in the government of the territory to suppress war and piracy and to prevent the establishment of Egyptian, Persian, German, or Russian influence. The first Bahrain–British treaty was signed in 1820; and Bahrain's British-protected status dates from 1861, with the completion of a treaty by the terms of which the sheikh bound himself to refrain from "the prosecution of war, piracy, or slavery." Until 1970 the government of Iran periodically advanced claims to sovereignty over Bahrain, but these were repudiated.

The environment. *Relief.* The state consists of two separate groups of islands. The 217-square-mile island of Bahrain is surrounded by smaller islands. Two of these—al-Muḥarraq and Sitrah—are joined to it by causeway; other islands in the group are an-Nabī Ṣāliḥ Umm aṣ-Ṣabbān, Umm Na'sān, and Jiddah. The second group consists of the Howar Islands, which are situated near the coast of Qatar about 12 miles southeast of Bahrain Island; small and rocky, they are inhabited only by a few fishermen and quarry workers.

The archipelago

While all of the small islands in both groups are rocky and low-lying, rising only a few feet above sea level, the main island is more varied in appearance. Its arched layers of rock consist of limestone, sandstone, and marl (loose clay, sand, or silt containing calcium carbonate); they were formed during the Cretaceous and Tertiary periods (from 7,000,000 to 136,000,000 years ago). The central region rises to 450 feet above sea level at the hill of Jabal ad-Dukhān and is rocky and barren. The southern and western lowlands consist of a sandy plain with some salt marshes. The northern and northwestern areas afford a striking contrast; they form a narrow belt of date and vegetable gardens irrigated from prolific springs and wells that tap artesian water. This water percolates through water-bearing rocks beneath the seabed to rise under pressure in Bahrain. Off the northern coasts of the islands, freshwater also rises in submarine springs on the seabed and has for centuries been captured in skin bags by divers. Since the drilling of artesian wells, however, these submarine springs have been little used.

Climate. Humidity is high throughout the year, and monthly mean temperatures from May until October exceed 85° F (29° C). Winters are cooler and more pleasant, with mean temperatures from December to March below 70° F (21° C). Rainfall, almost entirely confined to the winter months, averages only three inches per year. The predominant wind is the damp, northwesterly *shamal*, but the *gaws*, a hot, dry wind from the south, occasionally brings sand and dust.

Vegetation and animal life. Some 200 different species of desert plants grow in the bare arid portions of the islands, while the irrigated and cultivated areas support fruit trees, fodder crops, and vegetables. The variety of animals is limited by the desert conditions. The gazelle and the hare are not yet extinct, and lizards and jerboas (desert rats) are common, while the mongoose—probably imported from India—is found in the irrigated areas. There is also a surprising variety of birds native to Europe, Asia, and Africa.

Human settlement. In the irrigated areas of northern Bahrain Island, groves of date palms and orchards of banana, citrus, mango, and pomegranate trees create a sense of great fertility. Vegetables and lucerne (alfalfa) are grown under irrigation in the shade of the trees, and cattle graze there also. The Arab villages consist mainly of substantial houses of local stone or of concrete with flat roofs. Some of the temporary settlements of the fishermen and the very poor are constructed of *barasti* (sticks or canes from the date palms). There is little permanent settlement in the southern half of Bahrain Island.

Principal towns

There are two major towns—al-Manāmah and al-Muḥarraq. Al-Muḥarraq still presents the aspect of a traditional Arab town with narrow, winding streets, and dense settlement. Al-Manāmah, with its port of Mīnā' Salmān, blends features of the East and West. It contains several

large hotels, the principal government offices, schools, Western-style shops, and an Arab suq, or bazaar. Another town, Awālī, has tree-lined avenues; most of its houses are single-story bungalows inhabited by expatriate employees of the Bahrain Petroleum Company (Bapco). The town of Isa, a government venture to provide modern housing for 15,000 people and designed by British planners, was inaugurated in 1968.

The people. The population, about 180,000 in 1965, increased to over 200,000 by 1971. Some 82.5 percent of the people are native-born Bahrainis; another 5 percent are Omanis, while the remainder, in lesser proportions, are Indians, Pakistanis, Iranians, and British, in approximately that order. A substantial portion of the foreign community are employed by Bapco, and many Iranians, Indians, and Pakistanis are engaged in commerce. Some of the British work for the Bahrain government. The Muslim population is almost equally divided between the Sunnī and Shīʿah sects; the ruling family and many of the wealthier and more influential people are Sunnī. In the early 1970s the population of al-Manāmah was 89,000 and that of al-Muḥarraq town was 38,000; the two towns between them thus contained more than half of the total population. Arabic is the official language, but English is widely understood.

Bahrain, Area and Population

	area		population	
	sq mi	sq km	1965 census	1971 census*
Geographic divisions†				
Manama Area	...	...	80,000	90,000
Manama town	...	...	79,000	89,000
Jazeera	...	...	600	600
Muharraq Island	...	...	46,000	49,000
Muharraq town	...	...	34,000	38,000
Hidd town	...	...	5,000	5,000
Muharraq villages	...	...	7,000	7,000
Jiddhafs Area	...	...	15,000	20,000
Jiddhaf town	...	...	8,000	11,000
Jiddhaf villages	...	...	7,000	8,000
Northern Area	...	...		
Northern area villages	...	...	9,000	11,000
Western Area	...	...		
Western area villages	...	...	7,000	9,000
Central Area	...	...	5,000	15,000
Isa town	...	...	...	8,000
Central area villages	...	...	5,000	7,000
Sitra Area	...	...	9,000	11,000
Sitra town	...	...	5,000	7,000
Sitra area villages	...	...	4,000	5,000
Riffa Area	...	...	12,000	13,000
Riffa town	...	...	9,000	11,000
Awali town	...	...	2,000	1,000
Other areas	...	...	600	900
Other islands	...	...	...	100
Total Bahrain	256‡	662‡	182,000§	217,000§

*Provisional. †Geographic divisions are for statistical purposes only; they are not administrative. ‡Total area includes numerous small, uninhabited islands and dependencies of Bahrain. §Figures do not add to total given because of rounding.
Source: Official government figures.

Oil refining

The economy. *Oil.* Although oil has been produced since 1934, annual production is small and amounts to only about 3,800,000 tons from the oil field at Jabal ad-Dukhān. A refinery at Sitrah, however, is a major installation that handles 12,000,000 tons a year. Most of the crude oil is pumped by undersea pipeline from the Arabian American Oil Company (Aramco) fields in Saudi Arabia. Bahrain's role as an exporter of refined petroleum products is, therefore, more important than its role as a producer. An oil concession, which expires in 2024, is held by Bapco, in which two United States companies hold equal shares. Concessions for the marine areas to the north and west of Bahrain are held in equal shares by Aramco and another United States company. About 70 percent of the state's revenue comes from oil royalties and the remainder from customs dues, local taxes, and rent for land and property. Al-Manāmah is a free port for goods in transit.

Industries. The traditional industries of Bahrain were building dhows (Arab sailboats), fishing, pearling, and the manufacture of reed mats. These now exist only on a small scale, partly because of decreased demand and partly because of the availability of more remunerative employment. A modern slipway at Mīnā' Salmān handles the repair of ships up to 1,000 tons, and there are manufacturing plants producing building materials, soft drinks, and other consumer goods. The Bahrain Development Bureau, which encourages the development of manufacturing industries, arranged for the establishment of a major aluminum smeltery on the east coast. Production will commence in 1971 and is expected to reach 90,000 tons a year by 1972. Prawns are canned for export by arrangement between a Bahrain company and British fish distributors. Local fishing is also important.

Agriculture. Agriculture contributes significantly to the local food supply, of which dates, other fresh fruits, and vegetables are the main items. Livestock is unimportant, but camels and horses are bred for racing.

Management of the economy. Apart from oil and aluminum smelting, most industry is in private hands. The Bahrain government has a 27.5 percent interest in the international consortium Aluminium Bahrain that controls the aluminum smeltery. Many foreign firms that trade in the Gulf have their head offices in Bahrain.

Transportation. There is an excellent paved road system on Bahrain Island, and buses and taxis serve the principal towns and settlements. The international airport on al-Muḥarraq Island has flights to most countries in the Middle East and is one of the busiest airports in the Gulf. It is used by major airlines and a local aviation service. There are regular steamer services from Bahrain to other Gulf ports, as well as to Karachi and Bombay. Most general cargo is handled at Mīnā' Salmān; petroleum products are loaded at the Sitrah jetty.

Air services

Administration and social conditions. Bahrain is to a large extent a welfare state. Medical care and education are free; there are some 40,000 children in primary schools and more than 10,000 students in intermediate and high schools. Housing and transport are partially subsidized by the state.

The four urban and two rural municipalities are administered by councils, half of whose members are elected by male and female rate payers (local government taxpayers) and the remainder appointed by the government.

The ruler is assisted by a Council of Administration, which consists of the heads of various government departments, several of whom are members of the royal family. In the early 1970s the constitution was in a state of transition from that of a traditional Arabian patriarchal system to one of representative democracy.

Cultural life. In spite of its recent rapid economic development, Bahrain remains, in many respects, essentially Arab in its culture and life-style. The state radio station broadcasts only in Arabic, and television transmissions are received from Saudi Arabia. The traditional sports of falconry, gazelle and hare coursing, and horse and camel racing are still practiced by the wealthier Bahrainis.

Several weekly and daily papers are published in Arabic and a small number in English. The British Council in al-Manāmah is the main foreign cultural institution.

Prospects for the future. Bahrain's economy has developed rapidly in recent years, in part because of the fact that the state was able to exploit its early start in the oil industry and in part because of the development of trade in the entire Persian Gulf. Security and stable government have helped to promote this development. If that political stability both in Bahrain and in other Persian Gulf states survives the withdrawal of British forces, there is hope that the economic future of Bahrain will be bright.

BIBLIOGRAPHY. The most detailed general account is J.H.D. BELGRAVE, *Welcome to Bahrain*, 5th ed. (1965), a guidebook with a good account of the geography, history, and customs of Bahrain, together with a bibliography of works in Arabic, English, and French. Another detailed work is A. FAROUGHY, *The Bahrain Islands* (1951). The geog-

raphy of the islands is briefly described in both G. DALYELL, "The Bahrain Islands," *Scot. Geog. Mag.*, 57:58–61 (1941); and GREAT BRITAIN, NAVAL INTELLIGENCE DIVISION, *Iraq and the Persian Gulf* (1944). SIR RUPERT HAY, *The Persian Gulf States* (1959), is a good general account of the sheikhdoms of the Gulf, with an analysis of their political status and of British interests in the Gulf; J.B. KELLEY, "The Persian Claim to Bahrain," *International Affairs*, 33:51–70 (1957), and *Britain and the Persian Gulf* (1968), deal with the long-standing Iranian claim to Bahrain. One of the few articles dealing with social problems arising from economic development, as well as with the long-standing Sunnī–Shī'ah division among the population is F. QUBAIN, "Social Classes and Tensions in Bahrain," *Middle East Journal*, 9:269–280 (1955). A. VILLIER, "The Arab Dhow Trade," *Middle East Journal*, 2:399–416 (1948), and *Sons of Sinbad* (1940), deal with the old maritime trade of Bahrain and the Gulf.

(C.G.S.)

Baikal, Lake

Lake Baikal (Baykalskoye Ozero in Russian, Bajkalskoje Ozero in the transliteration system of the Soviet Akademiya Nauk) is located in the southern part of eastern Siberia within the Buryat Autonomous Soviet Socialist Republic and the Irkutsk (Irkutskaya) *oblast* of the Russian Soviet Federated Socialist Republic. It is the deepest continental body of water on earth, having a maximum depth of 5,314 feet. Its area is 12,200 square miles (31,500 square kilometres), with a length of 395 miles (636 kilometres) and an average width of 30 miles. It contains about one-fifth of the fresh water on the earth's surface and four-fifths of that in the U.S.S.R., or 5,500 cubic miles. Into Lake Baikal flow 336 rivers and streams, of which the largest are the Selenga, Barguzin, Upper (Verkhnyaya) Angara, Turka, and Snezhnaya.

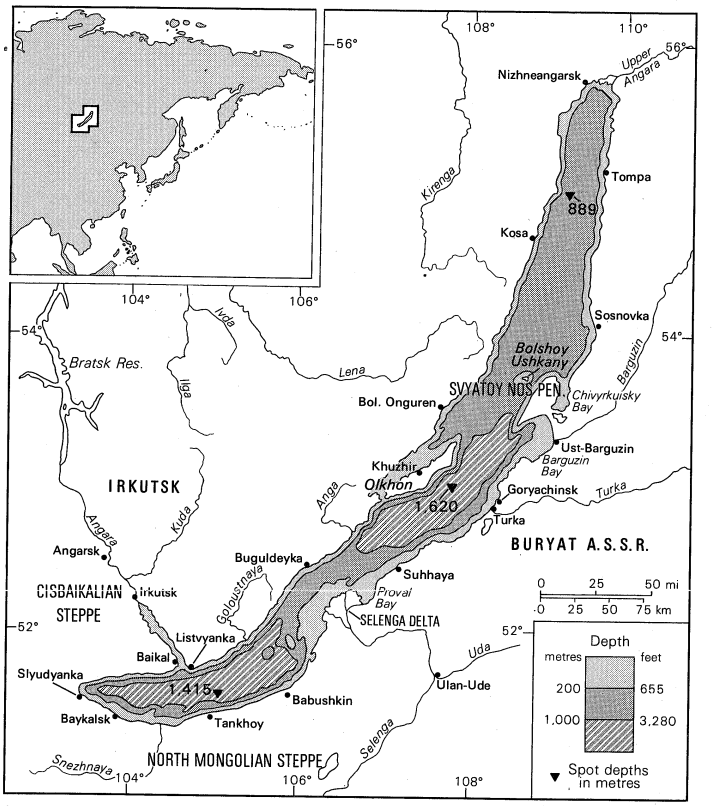
Baikal's geology

Baikal lies in a deep structural hollow surrounded by mountains, some of which tower 6,560 feet above it. The area is formed predominantly of metamorphic, sedimentary, and magmatic rock more than 500,000,000 years old. The sedimentary strata on the floor of the lake may be as much as 20,000 feet thick. Near the shore are remains of extinct volcanoes. Earth movements still continue, and there are occasional severe earthquakes; in 1862 a quake inundated 77 square miles in the northern Selenga River Delta, creating a new bay in Baikal known as Proval Bay. Breaking of the Earth's crust still produces hot mineral springs.

The lake hollow is not symmetrical, having steep slopes on the western shores and gentler slopes on the eastern. About 8 percent of the floor consists of shallows 160 feet deep. The meandering shoreline runs for 1,300 miles, with large indentations at the bays of Barguzin, Chivyrkuysky, and Proval and at Ayaya and Frolikha inlets; the Svyatoy Nos Peninsula juts out into the lake from the southeastern shore. Baikal contains 27 islands, of which five are periodically submerged; the largest are Olkhon (280 square miles) and Bolshoy Ushkany (over three square miles). The influx of water into the lake is primarily from rivers, chiefly the Selenga, while some comes from precipitation and subterranean sources. Most of the outflow is through the Angara River, a tributary of the Yenisey. The water level varies during the year by two or three feet, being highest in August–September and lowest in March–April.

Baikal's climate is much milder than that of the surrounding territory. January–February air temperatures average −2° F (−19° C), and the August temperatures average 52° F (11° C). The lake freezes in January and thaws in May. The water temperature at the surface in August is about 55° F (13° C) and reaches 68° F (20° C) in the offshore shallows. Waves sometimes measure over 15 feet. The water is very clear down to 130 feet, containing few minerals. Its salinity is low.

Plant and animal life in the lake are rich and various. There are over 1,200 animal species at different depths, and around 600 plant species living on or near the surface. About three-quarters of the species are peculiar to Baikal. There are 50 species of fish, belonging to seven families; the most numerous of these are bullheads of the family Cottidae, of which there are 25 species. The omul salmon is heavily fished; also important are the grayling and the lake whitefish. The largest of the fishes is the sturgeon, individuals of which sometimes measure 71 inches and weigh as much as 265 pounds. The one mammal is the Baikal seal. Unique to Baikal is a fish called the golomyanka, of the family Comephoridae, which gives birth to live young. There are 326 bird species in the Baikal area.



Lake Baikal.

Industries on the shores of Baikal include mining (mica and marble), cellulose and paper, shipbuilding, fisheries, and timber. There are many mineral springs, and visitors come to Goryachinsk and Khakusy for the curative properties of the spring waters. The lake is navigated from May to October by wooden rafts.

Anti-pollution measures

The Soviet government has begun to concern itself with the conservation of Baikal and its resources. A government decree concerning measures for the protection and rational use of those resources, and the prevention of polluting emissions from cellulose and other industrial plants, was adopted in June 1971.

In the town of Listvyanka is the Limnological Institute of the Siberian Department of the U.S.S.R. Academy of Sciences and the Baikal Sanatorium. In the town of Bolshiye Koty is the hydrobiological station of Irkutsk A.A. Zhdanov State University.

(G.I.G.)

Baking and Bakery Products

Baking, a dry-heat cooking process, is probably man's oldest cooking method. Bakery products are usually prepared from flour or meal derived from some form of grain. Bread, already a common staple in prehistoric times, provides many nutrients in the human diet.

HISTORY

The earliest processing of cereal grains probably involved parching or dry roasting of collected grain seeds. Flavour, texture, and digestibility were later improved by cooking whole or broken grains with water, forming gruel or porridge. It was a short step to the baking of a layer of viscous gruel on a hot stone, producing primitive flat bread. More sophisticated versions of flat bread include the Mexican tortilla, made of processed corn, and the chapatti of India, usually made of wheat.

Improved baking techniques

Baking techniques improved with the development of an enclosed baking utensil and then of ovens, making possible thicker baked cakes or loaves. The phenomenon of fermentation, with the resultant lightening of the loaf structure and development of appealing flavours, was probably first observed when doughs or gruels, held for several hours before baking, exhibited spoilage caused by yeasts. Most doughs exhibit some response to the puffing or leavening action of fermentation, but only wheat yields a flour producing a light, porous structure in baked products. Early baked products were made of mixed seeds with a predominance of barley, but wheat flour, because of its superior response to fermentation, eventually became the preferred cereal among cultural groups sufficiently advanced in culinary techniques to make leavened bread.

Brewing and baking were closely connected in early civilizations. Fermentation of a thick gruel resulted in a dough suitable for baking; a thinner mash produced a kind of beer. Both techniques required knowledge of the "mysteries" of fermentation and a supply of grain. Increasing knowledge and experience taught the artisans in the baking and brewing trades that barley was best suited to brewing, while wheat was best for baking.

By 2600 BC the Egyptians, credited with the first intentional use of leavening, were making bread by methods similar in principle to those of today. They maintained stocks of sour dough, a crude culture of desirable fermentation organisms, and used portions of this material to inoculate fresh doughs. With doughs made by mixing flour, water, salt, and leaven, the Egyptian baking industry eventually developed over 50 varieties of bread, varying in shape and using such flavouring materials as poppyseed, sesame, and camphor. Samples found in tombs are flatter and coarser than modern bread.

The Egyptians developed the first ovens. The earliest known examples are cylindrical vessels made of baked Nile clay, tapered at the top to give a cone shape and divided inside by a horizontal shelflike partition. The lower section is the firebox, the upper section is the baking chamber. The pieces of dough were placed in the baking chamber through a hole provided in the top.

Greek cultures made few improvements in bread and baking; most of their bread was in the form of flat cakes.

In the first two or three centuries after the founding of Rome, baking remained a domestic skill with few changes in equipment or processing methods. According to Pliny the Elder, there were no bakers in Rome until the middle of the 2nd century BC. As well-to-do families increased, women wishing to avoid frequent and tedious bread making began to patronize professional bakers, usually freed slaves. Loaves molded by hand into a spheroidal shape, generally weighing about a pound, were baked in a beehive-shaped oven fired by wood. *Panis artopticius* was a variety cooked on a spit, *panis testuatis* in an earthen vessel.

Although Roman professional bakers introduced technological improvements, many were of minor importance, and some were essentially reintroductions of earlier developments. The first mechanical dough mixer, attributed to Marcus Virgilius Euryasaces, a freed slave of Greek origin, consisted of a large stone basin in which wooden paddles, powered by a horse or donkey walking in circles, kneaded the dough mixture of flour, leaven, and water.

Guilds formed by the miller-bakers of Rome became institutionalized. During the 2nd century AD, under the Flavians, they were organized into a "college" with work rules and regulations prescribed by government officials. The trade eventually became obligatory and hereditary, and the baker became a kind of civil servant with limited freedom of action.

The revival of guilds

During the early Middle Ages baking technology advances of preceding centuries disappeared, and bakers reverted to mechanical devices used by the ancient Egyptians and to more backward practices. But in the later Middle Ages the institution of guilds was revived and expanded. Several years of apprenticeship were necessary before an applicant was admitted to the guild; often an intermediate status as journeyman intervened between apprenticeship and full membership (master). The rise of the bakers guilds reflected significant advances in technique. A 13th-century French writer named 20 varieties of bread varying in shape, flavourings, preparation method, and quality of the meal used. Guild regulations strictly governed size and quality. But outside the cities, bread was usually baked in the home. In medieval England rye was the main ingredient of bread consumed by the poor; it was frequently diluted with meal made from other cereals or leguminous seeds. Not until about 1865 did the cost of white bread in England drop below brown bread.

At that time improvements in baking technology began to accelerate rapidly, owing to the higher level of technology generally. Ingredients of greater purity and improved functional qualities were developed, along with equipment reducing the need for individual skill and eliminating hand manipulation of bread doughs. Automation of mixing, transferring, shaping, fermentation, and baking processes began to replace batch processing with continuous operations. The enrichment of bread and other bakery foods with vitamins and minerals was a major accomplishment of the mid-20th-century baking industry.

FUNCTION OF INGREDIENTS

Flour, water, and leavening agents are the ingredients primarily responsible for the characteristic appearance, texture, and flavour of most bakery products. Eggs, milk, salt, shortening, and sugar are effective in modifying these qualities, and various minor ingredients may also be used.

Flour. Wheat flour is unique among cereal flours in that, when mixed with water in the correct proportions, its protein component forms an elastic network capable of holding gas and developing a firm spongy structure when baked. The proteinaceous substances contributing these properties are known collectively as gluten. The suitability of a flour for a given purpose is determined by the type and amount of its gluten content. These characteristics are controlled by the genetic constitution and growing conditions of the wheat from which the flour was milled, as well as the milling treatment applied.

Low-protein, soft-wheat flour is appropriate for cakes, pie crusts, cookies, and other products not requiring great expansion and elastic structure. High-protein, hard-wheat flour is adapted to bread, hard rolls, soda crackers, and Danish pastry, all requiring elastic dough and often expanded to low densities by the leavening action.

Since most of the nitrogen in foods consists of protein, the protein content of flour is commonly estimated by the Kjeldahl determination, a process measuring the total nitrogen. Protein quality is often evaluated by measuring the elasticity and extensibility of flour-and-water doughs. Specifications for flours to be used in cakes, pies, and pastries may include limitations on the minimum or maximum viscosity developed in aqueous suspensions under standardized conditions. The effectiveness of the starch-digesting enzymes can be estimated by including a time variable.

Flour particle size and the extent of any damage to starch granules during milling are important factors, especially in cookie production, apparently affecting the hydration rate of the flour in low-moisture-content doughs.

Leavening agents. Pie doughs and similar products are usually unleavened, but most bakery products are leavened, or aerated, by gas bubbles developed naturally or folded in from the atmosphere. Leavening may result from yeast or bacterial fermentation, from chemical reactions, or from the distribution in the batter of atmospheric or injected gases.

Use of bakers' yeast

All commercial breads, except salt-rising types and some rye bread, are leavened with bakers' yeast, composed of living cells of the yeast strain *Saccharomyces cerevisiae*. A typical yeast addition level might be 2 per-

cent of the dough weight. Bakeries receive yeast in the form of compressed cakes containing about 70 percent water or as dry granules containing about 8 percent water. On a dry-weight basis, the gas-generating potentials of the two forms of yeast are about the same. Dry yeast, more resistant to storage deterioration, requires rehydration before it is added to the other ingredients.

Bakers' yeast performs its leavening function by fermenting such sugars as glucose, fructose, maltose, and sucrose. It cannot use lactose, the predominant sugar of milk, or certain other carbohydrates. The principal products of fermentation are carbon dioxide, the leavening agent, and ethanol, an important component of the aroma of freshly baked bread. Other yeast activity products also flavour the baked product and change the dough's physical properties.

The rate at which gas is evolved by yeast during the various stages of dough preparation is important to the success of bread manufacture. Gas production is partially governed by the rate at which fermentable carbohydrates become available to the yeast. The sugars naturally present in the flour and the initial stock of added sugar are rapidly exhausted. A relatively quiescent period follows, during which the yeast cells become adapted to the use of maltose, a sugar constantly being produced in the dough by the action of diastatic enzymes on starch. The rate of yeast activity is also governed by temperature and osmotic pressure, the latter primarily a function of the water content and salt concentration.

Leavening with sodium bicarbonate

Layer cakes, cookies (sweet biscuits), biscuits, and many other bakery products are leavened by carbon dioxide from added sodium bicarbonate (baking soda). Added without offsetting amounts of an acidic substance, sodium bicarbonate tends to make dough alkaline, causing flavour deterioration and discoloration and slowing carbon dioxide release. Addition of an acid-reacting substance promotes vigorous gas evolution and maintains dough acidity within a favourable range.

Carbon dioxide produced from sodium bicarbonate is initially in dissolved or combined form. The rate of gas release affects the size of the bubbles produced in the dough, consequently influencing the grain, volume, and texture of the finished product. Much research has been devoted to the development of leavening acids capable of maintaining the rate of gas release within the desired range. Acids such as acetic, from vinegar, or lactic, from sour milk, usually act too quickly; satisfactory compounds include cream of tartar (potassium acid tartrate), sodium aluminum sulfate (alum), sodium acid pyrophosphate, and various forms of calcium phosphate.

The chemical formulas of these compounds do not adequately indicate their functions in doughs. The addition of small quantities of additives during manufacture can greatly affect the reaction rate of the compound. Granule size and form also have modifying functions. Companies specializing in the manufacture of leavening acids usually offer several types of sodium acid pyrophosphate. The chemical formulas are apparently the same, and analysis will only reveal trace element variations, but the slowest reacting member of this series will produce an initial rate of gas evolution several times slower than the rate produced by the fastest acting compound.

Instead of adding soda and leavening acids separately, most commercial bakeries and domestic bakers use baking powder, a mixture of soda and acids in appropriate amounts and with such added diluents as starch, simplifying measuring and improving stability. The end products of baking-powder reaction are carbon dioxide and some blandly flavoured harmless salts. All baking powders meeting basic standards have virtually identical amounts of available carbon dioxide, differing only in reaction time. Most commercial baking powders are of the double-acting type, giving off a small amount of available carbon dioxide during the mixing and makeup stages, then remaining relatively inert until baking raises the batter temperature. This type of action eliminates excessive loss of leavening gas, which may occur in batter left in an unbaked condition for long periods.

Under baking conditions, ammonium bicarbonate, sometimes used to leaven cookies, decomposes entirely to carbon dioxide, water vapour, and ammonia, leaving no residue of solids. Control of its rate of decomposition is difficult, however, and its ammoniacal odour persists unless the finished product is baked almost to dryness.

Leavening by entrapped air

Angel food cakes, sponge cakes, and similar products are customarily prepared without either yeast or chemical leaveners. Instead, they are leavened by air entrapped in the product through vigorous beating. This method requires a readily foaming ingredient, capable of retaining the air bubbles, such as egg whites. To produce a cake of fine and uniform internal structure, the pockets of air folded in during beating are rapidly subdivided into small bubbles with such mixing utensils as wire whips, or whisks.

The small gas bubbles initially present in other doughs and batters also provide focuses for the evolution of gases produced by yeast or chemical reaction.

The vaporization of volatile fluids (*e.g.*, ethanol) under the influence of oven heat can have a leavening effect. Water-vapour pressure, too low to be significant at normal temperatures, exerts substantial pressure on the interior walls of bubbles already formed by other means as the interior of the loaf or cake approaches the boiling point. The expansion of such puff pastry as used for napoleons (rich desserts of puff pastry layers and whipped cream or custard) and *vol-au-vents* (puff pastry shells filled with meat, fowl, fish, or other mixtures) is entirely due to water-vapour pressure.

Shortening. Fats and oils are essential ingredients in nearly all bakery products. Shortenings have a tenderizing effect in the finished product and often aid in the manipulation of doughs. In addition to modifying the mouth feel or texture, they often add flavour of their own and tend to round off harsh notes in some of the spice flavours.

The common fats used in bakery products are lard, beef fats, and hydrogenated vegetable oils. Butter is used in some premium and specialty products as a texturizer and to add flavour, but its high cost precludes extensive use. Cottonseed oil and soybean oil are the most common processed vegetable oils used. Corn, peanut, and coconut oils are used to a limited extent; fats occurring in other ingredients, such as egg yolks, chocolate, and nut butters, can have a shortening effect if the ingredients are present in sufficient quantity.

Breads and rolls often contain only 1 or 2 percent shortening; cakes will have 10 to 20 percent; Danish pastries prepared according to the authentic formula may have about 30 percent; pie crusts may contain even more. High usage levels require those shortenings that melt above room temperature; butter and liquid shortenings, with their lower melting point, tend to leak from the product.

Desirable properties in general-purpose shortenings include bland flavour, white colour (or clarity in liquids), good plasticity (workability), and flavour stability. Shortening manufacturers achieve these characteristics by such physical and chemical treatments as hydrogenation, blending, catalytic rearrangement, fractionation, and controlled crystallization.

In addition to the usual chemical tests applied to fats, commercial bakeries may evaluate shortening by performance tests of such properties as plasticity, icing volume, water absorption, and characteristics produced in the finished cake.

Commercial shortenings may include antioxidants, to retard rancidity, and emulsifiers, to improve the shortening effect. Colours and flavours simulating butter may also be added. Margarines, emulsions of fat, water, milk solids, and salt, are popular bakery ingredients.

Fats of any kind have a destructive effect on meringues and other protein-based foams; small traces of oil left on the mixing utensils can deflate an angel food cake to unacceptably high density.

Liquids. Water is the liquid most commonly added to doughs. Milk is usually added to commercial prepara-

tions in dried form, and any moisture added in the form of eggs and butter is usually minimal. Water is not merely a diluent or inert constituent; it affects every aspect of the finished product, and careful adjustment of the amount of liquid is essential to make the dough or batter adaptable to the processing method. If dough is too wet it will stick to equipment and have poor response to shaping and transfer operations; if too dry, it will not shape or leaven properly.

The use of water in dough

Water hydrates gluten, permitting it to aggregate in the form of a spongy cellular network, the structural basis of most bakery products. It provides a medium in which yeast can metabolize sugars to form carbon dioxide and flavouring components, and allows diffusion of nutrients and metabolites throughout the mass. Water is an indispensable component of the baking-powder reaction, and it allows starch to gelatinize during baking and prevents interior browning of bakery products.

Water impurities affect dough properties. Water preferred for baking is usually of medium hardness (50 to 100 parts per million) with a neutral pH (degree of acidity), or slightly acid (low pH). Water that is too soft can result in sticky doughs, while very hard water may retard dough expansion by toughening the gluten (calcium ions, particularly, promote cross-linking of gluten protein molecules). Water sufficiently alkaline to raise the dough pH may have a deleterious effect on fermentation and on flour enzymes. Although strongly flavoured contaminants may affect the acceptability of the finished product, chlorides and fluorides at concentrations usually found in water supplies have little influence on bread doughs.

Eggs. The differences between yolks and whites must be recognized in considering the effect of eggs on bakery products. Yolks contain about 50 percent solids, of which 60 percent or more is strongly emulsified fat, and are used in bakery foods for their effect on colour, flavour, and texture. Egg whites, containing only about 12 percent solids, primarily protein, and no fat, are important primarily for their texturizing function and give foams of low density and good stability when beaten. When present in substantial amounts, they tend to promote small, uniform cell size and relatively large volume. Meringues and angel food cakes are dependent on egg white foams for their basic structure. Although the presence of fats and oils greatly diminishes its foaming power, the white still contributes to the structure of layer cakes and similar confections containing substantial amounts of both shortening and egg products.

Whole eggs form a relatively low-volume foam of poor stability but are otherwise intermediate in properties between yolks and whites. In mixed whole egg solids, the egg whites represent about two-thirds of the total weight.

Egg products are available to bakers in frozen or dried form. Few commercial bakers break fresh eggs for ingredients because of labour costs, unstable market conditions, and sanitary considerations. Most bakers use dried egg products because of their greater convenience and superior storage stability over frozen eggs. Processed and stored correctly, dried egg products are the functional equivalent of the fresh material, although flavour of the baked goods may be affected adversely at very high usage levels.

The use of dried eggs

Sweetening agents. Normal wheat flour contains about 1 percent sugars. Most are fermentable compounds, such as sucrose, maltose, glucose, and fructose. Additional maltose is formed during fermentation by the action of amylolytic enzymes (from malt and flour) on the starch. Glucose and sucrose are the sugars most frequently added to doughs and batters. The presence of yeast rapidly converts the sucrose to fructose and glucose (*i.e.*, invert sugar). Invert sugar can also be added to mixtures not containing yeast.

Sweetening power is the obvious property of added sugars, but sugars also provide fermentables for yeast activity. Crust colour development is related to the amount of reducing sugars present, and a dough in which the sugars have been thoroughly depleted by yeast will produce a poorly browned crust.

Doughs having high concentrations of dissolved substances retard fermentation because of the high osmotic pressure (low water activity) of the aqueous phase. Sugars constitute the bulk of dissolved materials in most doughs; and for this reason, sweet yeast-leavened goods develop gas and expand more slowly than bread doughs.

Sugar at high concentrations will tenderize the finished baked product by interfering with the hydration of the gluten network and the gelatinization of starch granules.

In addition to cane and beet sugar and corn syrups, honey and molasses often are used as sweeteners in crackers and cookies.

TYPES OF BAKERY PRODUCTS AND PRODUCTION METHODS

Yeast-leavened products. Breads and rolls constitute most of the bakery foods consumed throughout the world, and most bread and rolls are made from yeast-leavened doughs. The yeast-fermentation process leads to the development of desirable flavour and texture, and such products are nutritionally superior to products of the equivalent chemically leavened doughs, since yeast cells themselves add a wide assortment of vitamins and good quality protein.

Satisfactory white bread can be made from flour, water, salt, and yeast. (A "sourdough" addition may be substituted for commercial yeast.) Yeast-raised breads based on this simple mixture include Italian-style bread, and French or Vienna breads. Such breads have a hard crust, are relatively light in colour, with a course and tough crumb, and flavour that is excellent in the fresh bread but deteriorates with age. In the U.S., commercially produced breads of this type are often modified by the addition of dough improvers, yeast foods, mold inhibitors, vitamins, minerals, and small quantities of enriching materials such as milk solids or shortening. Formulas may vary greatly from one bakery to another and between different sections of the country. The standard low-density, soft-crust bread and rolls constituting the major proportion of breads and rolls sold in the U.S. contain greater quantities of enriching ingredients than the lean breads described above.

Breads designed to take advantage of consumer demands for unusual flavours or special nutritional qualities can assume an almost unlimited range of forms and compositions. Whole-wheat bread, using a meal made substantially from the entire wheat kernel instead of flour, is a dense, rather tough, dark product. Breads sold as wheat or part-whole-wheat products contain a mixture of whole grain meal with sufficient flour to produce satisfactory dough expansion. Bread made from crushed or ground whole rye kernels, without any wheat flour, such as pumpernickel, is dark, tough, and coarse textured. Rye flour with the bran removed, when mixed with wheat flour, allows production of a bread with better texture and colour. In darker rye bread, it is customary to add caramel colour to the dough, and most rye bread is flavoured with caraway seeds. Rye bread is sometimes made by the sourdough method, requiring addition of a small portion of a dough previously allowed to develop lactic-acid-producing bacteria. These bacteria are micro-organisms that ferment some of the sugars in the fresh dough batch, producing characteristic tastes and odours. Bakery supply houses offer dehydrated cultures as sourdough substitutes.

Sourdough method for rye bread

Salt-rising bread, a specialty product, derives its pungent aroma from combined yeast and bacterial fermentation. It was formerly made by a natural sourdough method, but commercial cultures are now available to provide the inoculum. The product name refers to the original use of a high-salt concentrate sourdough limiting growth of common yeasts, thus creating a better environment for growth of the desired bacteria species.

Potato bread, another variety that can be leavened with a primary ferment, was formerly made with a sourdough utilizing the action of wild yeasts on a potato mash and producing the typical potato-bread flavour but is now commonly prepared from a mixture of bakers' yeast, potato flour, and water.

The sponge and dough process is usually employed in making white bread and many specialty breads. The other conventional dough-preparation procedure, the straight dough method, in which all of the ingredients are mixed in one step, is rarely used for regular white bread. It is not sufficiently adaptable to allow compensation for the usual fluctuations in ingredient properties.

The sponge and dough mixing method consists of two distinct stages. In the sponge stage, the mixture, or sponge, usually contains one-half to three-fourths of the flour, all of the yeast, yeast foods, and malt, and enough water to make a stiff dough. Shortening may be added at this stage, although it is usually added later, and one-half to three-fourths of the salt may be added to control fermentation. The sponge is fermented until it begins to decline in volume. The time required for this process, called the drop or break, depends upon such variables as temperature, type of flour, amount of yeast, absorption, and amount of malt, which are frequently adjusted to produce a drop in about three to five hours.

At the dough stage, the sponge is returned to the mixer, and the remaining ingredients are added. The dough is developed to an optimum consistency, then returned to the fermentation room.

Judging dough developments

The ability to judge dough development must be acquired by experience. Mixing personnel often learn to estimate the stage of development by the sound of the dough as it slaps the mixer bowl. When objective measuring techniques are employed, optimum development is the stage at which the dough mass exhibits maximum resistance to shearing. Visually, development is optimal when the dough mass, exhibiting a silky sheen, is thrown around the mixer bowl in one piece, stretching elastically from the mixer arm but not breaking into pieces. A piece of this dough, removed from the mixer and stretched with the fingers, can be formed into a thin film having a webbed appearance when viewed by transmitted light.

Advantages of the sponge and dough method include: (1) a saving in the amount of yeast; about 20 percent less is required than for a straight dough; (2) the greater volume and more desirable texture and grain usually produced; and (3) the greater flexibility allowed in operations because, in contrast to straight doughs that must be taken up when ready, sponges can be held for later processing without marked deterioration of the final product.

The sponge method, however, involves extra handling of the dough, additional weighing and measuring, and a second mixing and thus has the disadvantage of increasing labour, equipment, and power costs.

Recommended variations in the straight dough process include the remixed straight dough process, with a small portion of the water added at the second mix, and the no-punch method, involving extremely vigorous mixing. Continuous mixing processes can be considered either straight dough methods or sponge methods, with the "sponge" consisting of the liquid ferment. Based on flour performance, they may be considered straight dough processes (Figure 1).

Yeast-leavened sweet goods, made from mixtures similar in many respects to bread doughs, include "raised" doughnuts, Danish pastries, and coffee cakes. The two main classifications are remix sweet doughs, corresponding to a sponge and dough bread product, and the straight doughs. With these two procedures, an almost infinite number of varieties can be produced by varying flavouring, filling, icing, and shape.

Sweet doughs, richer in shortening, milk, and sugar than bread doughs, often contain whole eggs, egg yolks, egg whites, or corresponding dried products. The enriching ingredients alter the taste, produce flakier texture, and improve nutritional quality. Spices such as nutmeg, mace, cinnamon, coriander, and ginger are frequently used for sweet-dough products; other common adjuncts include vanilla, nuts and nut pastes, peels or oils of lemon or orange, raisins, candied fruit pieces, jams, and jellies.

Although various portion-sized sweet goods are often called "Danish pastry," the name originally referred only to products made by a special roll-in procedure, in which

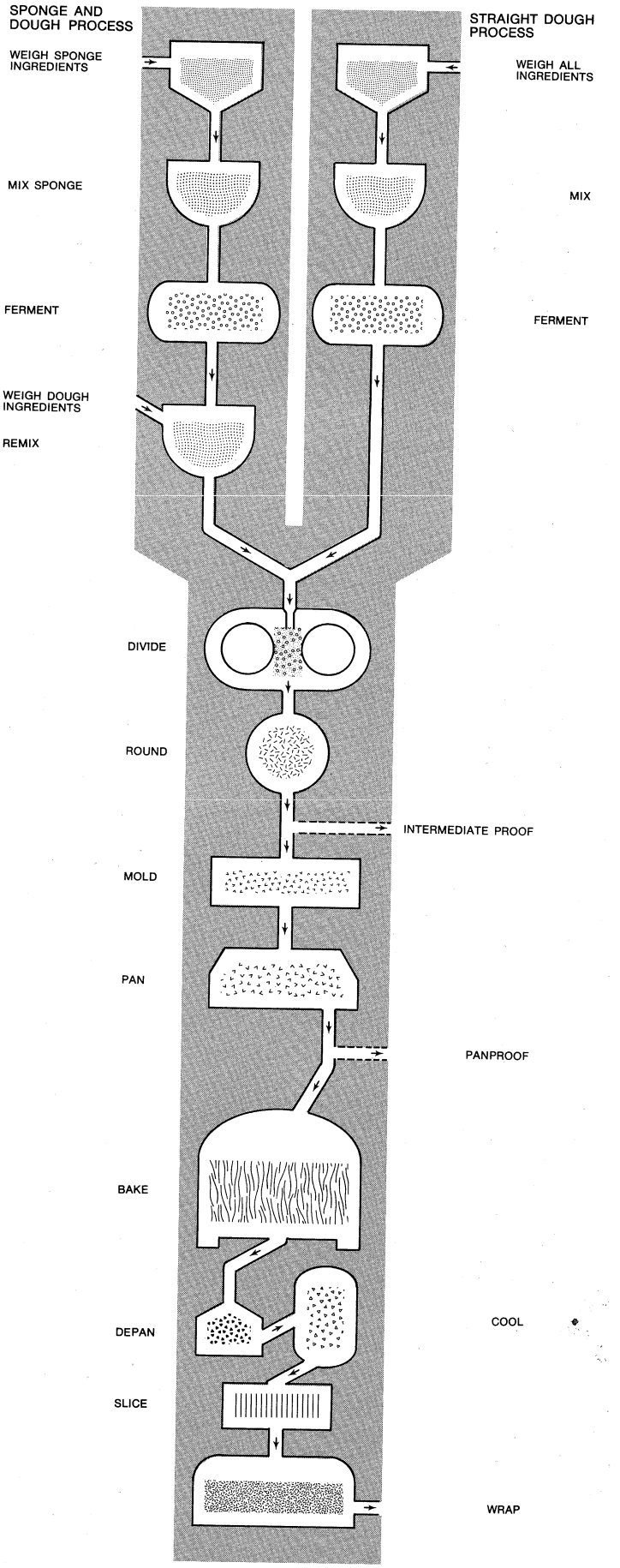


Figure 1: Steps in conventional (batch) production of bread.
Drawing by D. Meighan

Danish pastry production

yeast-leavened dough sheets are interleaved with layers of butter, and the layers are reduced in thickness, then folded to obtain many thin layers of alternating shortening and dough. Danish doughs ordinarily receive little fermentation. Before the fat is rolled in, there is an initial period of 20 to 30 minutes in the refrigerator, allowing some gas and flavour to develop. Proof time, fermentation of the piece in its final shape, is usually only 20 to 30 minutes, at lower temperatures. When properly made, these doughs yield flaky, baked products, rich in shortening, with glossy crusts.

Equipment. In most large, modern bakeries ingredients are mainly received and transferred by bulk-handling methods. Flour is pneumatically conveyed and stored in large tanks; shortening is received and stored in the liquid state; and water is continuously adjusted to the proper temperature and measured by precision pumps. Ingredients used in small quantities, such as milk solids, yeast, and eggs, are sometimes handled in bulk but are more often received and dispensed in small unit packages. In addition to the economic advantages of bulk handling, the mixing effects inherent in transporting and storing by these methods result in high ingredient uniformity.

Bread doughs customarily are mixed in large, horizontal dough mixers, processing about 2,000 pounds per batch, and usually constructed with heat-exchange jackets, allowing temperature control. The objectives of mixing are a nearly homogeneous blend of the ingredients, and "developing" of the dough by formation of the gluten into elongated fibres that will form the basic structure of the loaf. Because intense shearing actions must be avoided, the usual dough mixer has several horizontal bars, oriented parallel to the body of the mixer, rotating slowly at 35 to 75 revolutions per minute (rpm), stretching and kneading the dough by their action. A typical mixing cycle would be about 12 minutes.

The mixed bread dough is dumped into a trough, a shallow rectangular metal tank on wheels, and placed in an area of controlled temperature and humidity (*e.g.*, 80° F and 75 percent relative humidity), where the "sponge" undergoes about four hours of fermentation. The fermented sponge is next dumped into a horizontal dough mixer, often the same mixer that is used for the sponge mix, and the remaining dough ingredients are then incorporated (Figure 2).

By courtesy of the Kitchens of Sara Lee

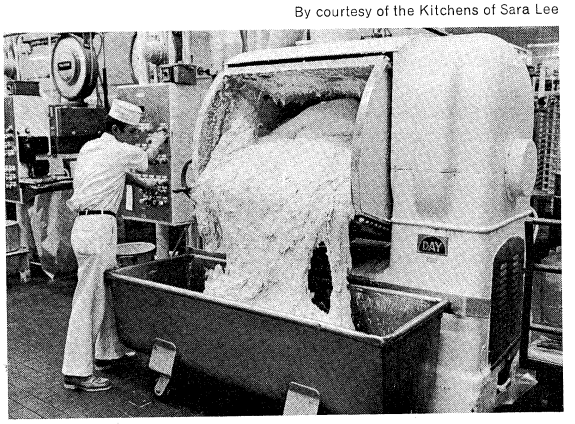


Figure 2: Dough mixer, showing a scientifically mixed batch of Danish dough being dropped into a stainless steel trough. When the mixer is closed, computer-measured liquid and dry ingredients enter at the top of the mixer, assuring uniformity of each batch.

Makeup equipment

The remixed dough is processed by a series of devices loosely classified as makeup equipment. In the manufacture of pan bread, makeup equipment includes the divider, the rounder, the intermediate proofer, and the molder. With the exception of the intermediate proofer, each of these devices changes the shape of the dough piece. Although the dough remains in approximately the same state of chemical and physical development throughout its processing in the makeup equipment, chemical and biological changes do occur in the intermediate proofer.

After the mass of dough has completed fermentation, and has been remixed if the sponge and dough process is employed, the filled trough is moved to the divider area or to the floor above the divider. The dough is dropped into the divider hopper, which cuts it into loaf-sized pieces. Since equipment that subdivides the dough satisfactorily on a weight basis is not yet available, the present apparatus operates on a volumetric basis, forcing the dough into pockets having a known volume. The pocket contents are cut off from the main dough mass, ejecting pieces of constant volume onto a conveyor leading to the rounder. A reciprocating division box is used for measuring. When density is kept constant, weight and volume of the dough pieces is the same, but since continual gas development may cause density variation, fluctuations may occur in the weight of the finished pieces.

Commercial dividers have from two to eight pockets in the cylinder, operate at speeds up to 25 strokes per minute, have a scaling range from 6 to 36 ounces, and use motors having up to 7½ horsepower.

By courtesy of General Host Corporation

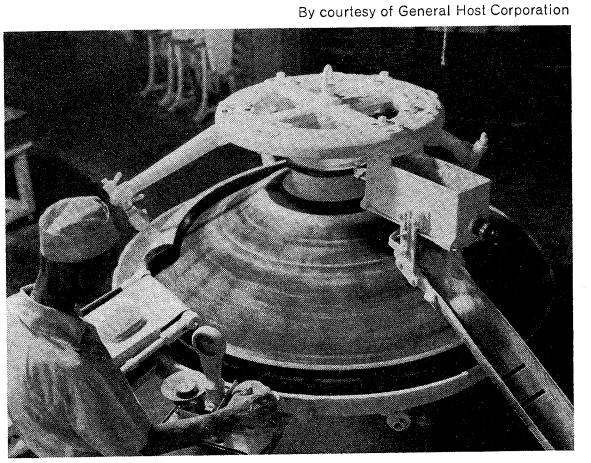


Figure 3: Rounder, which shapes the loaf-size pieces of bread dough into flour-sprinkled balls.

Function of the rounder

Dough pieces leaving the divider are irregular in shape, with sticky cut surfaces from which the gas can readily diffuse. Their gluten structure is somewhat disoriented and unsuitable for molding. The rounder closes these cut surfaces, giving each dough piece a smooth and dry exterior; forms a relatively thick and continuous skin around the dough piece, reorienting the gluten structure; and shapes the dough into a ball for easier handling in subsequent steps. It performs these functions by rolling the well-floured dough piece around the surface of a drum or cone, moving it upward or downward along this surface by means of a spiral track. As a result of this action, the surface is dried both by the even distribution of dusting flour and by dehydration resulting from exposure to air; the gas cells near the surface of the ball are collapsed, forming a thick layer inhibiting the diffusion of gases from the dough; and the dough piece assumes an approximately spherical shape (Figure 3).

Rounding machines vary in such features as configuration of the rounding surface; texture or composition of dough-contacting parts; mechanism for adjusting the relationship of the dough race, or hook, to the fixed drum or cone; and method of applying dusting flour.

Dough leaving the rounder is almost completely degassed. It lacks extensibility, tears easily, has rubbery consistency, and has poor molding properties. To restore a flexible, pliable structure that can be easily shaped by the molder, the dough piece must be allowed to rest while fermentation proceeds. This is accomplished by letting the dough ball travel through an enclosed cabinet, the intermediate proofer, for several minutes. Physical changes, other than gas accumulation, occurring during this period are not yet understood, but there are appar-

ently alterations in the molecular structure of the dough rendering it more responsive to subsequent operations. Upon leaving the intermediate proofer, the dough is more pliable and elastic, its volume is increased by gas accumulation, and its skin is firmer and drier.

Most intermediate proofers are the overhead type, in which the principal part of the cabinet is raised above the floor, allowing space for other makeup machinery beneath it. Intermediate proofers of all types are equipped with variable-speed controls to determine the length of time dough pieces spend within the cabinet. Although interior humidity and temperature control is desirable, most intermediate proofers, having no air-conditioning attachments, depend upon humidity accumulating from the loaves and upon ambient temperatures.

Shaping of dough by molder

The molder receives pieces of dough from the intermediate proofer and shapes them into cylinders ready to be placed in the pans. There are several types of molders, but all have four functions in common: sheeting, curling, rolling, and sealing. The dough as it comes from the intermediate proofer is a flattened spheroid; the first function of the molder is to flatten it into a thick sheet, usually by means of two or more consecutive pairs of rollers, each succeeding pair set more closely together than the preceding pair. The sheeted dough is curled into a loose cylinder by a special set of rolls or by a pair of canvas belts. The spiral of dough in the cylinder is not adherent upon leaving the curling section, and the next operation of the molder is to seal the dough piece, allowing it to expand without separating into layers. The conventional molder rolls the dough cylinder between a large drum, and a smooth-surfaced semicircular compression board. Clearance between the drum and board is gradually reduced, and the dough, constantly in contact with both surfaces, becomes transversely compressed (Figure 4).

By courtesy of the U.S. Department of Defense

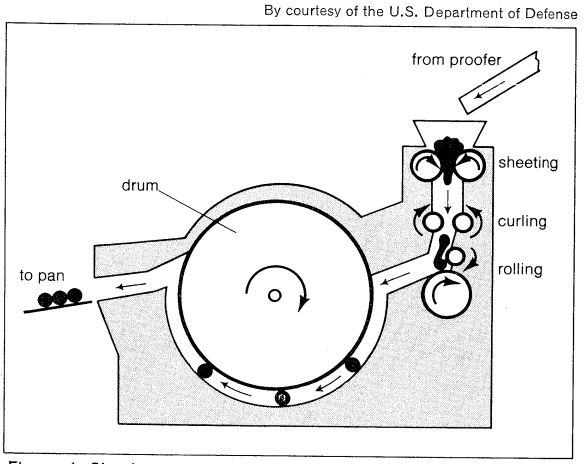


Figure 4: Simple type of drum molder.

The automatic panning device is an integral part of most modern molders. As the empty pans, carried on a conveyor, pass the end of the machine, the loaves are transferred from the molder and positioned in the pans by a compressed air-operated device. Before the filled pans are taken to the oven, the dough undergoes another fermentation, or pan-proofing, for about 20 minutes, at temperatures of 100° to 120° F (38° to 49° C) and relative humidity of about 82 percent.

The output of all other equipment is usually keyed to the oven, probably the most critical equipment in the bakery, operated at full capacity whenever possible. Most modern commercial bakeries use either the tunnel oven, consisting of a metal belt passing through a connected series of baking chambers open only at the ends, or the tray oven, with its rigid baking platform carried on chain belts. Other types include the peel oven, having a fixed hearth of stone or brick on which the loaves are placed with a wooden paddle or peel; the reel oven, with shelves rotating on a central axle in ferris-wheel fashion; the rotating hearth oven; and the draw plate oven. Most ovens are heated by gas burned within the chamber, although oil or electricity may be used. Burners are sometimes isolated from the main chamber, heat transfer then occurring through induced currents of air.

Reactions in the oven are both physical and chemical in nature. Physical reactions include film formation, gas expansion, reduction of gas solubility, and alcohol evaporation. Chemical reactions include yeast fermentation, carbon dioxide formation, starch gelatinization, gluten coagulation, sugar caramelization, and browning.

Automatic loaders and unloaders move the pans in and out of the ovens. Automatic depanners, removing the loaves from the pans, either invert the pans, jarring them to dislodge the bread, or pick the loaves out of the pans by means of suction cups attached to belts.

Continuous bread making. Many steps in the conventional bread-making procedure have been fully automated, but the batch process was the basic approach until about 1955, when two systems relying on truly continuous dough preparation were developed. In both systems the dough is handled continuously from the time the ingredients are mixed until it is deposited in the pan, although the initial fermentation process, still essentially a batch procedure, is conducted in tanks (Figure 5).

Continuous systems, operating without interruption for the addition of ingredients, feed the baker's formula to a device mixing all ingredients into a homogeneous mass. The batter-like material passes through a dough pump regulating the flow and delivering the mixture to a developing apparatus, where kneading work is applied to obtain desirable dough structure and proper gas-retention properties. The dough moves out of the developer into a metering device that constantly extrudes the dough and intermittently severs a loaf-sized piece, which falls into a pan passing beneath (Figure 5).

Although ingredients are generally the same as those used in batch processes, closer control and more rigid specifications are necessary in continuous processing to assure satisfactory operation of each unit. Changes in conditions cannot be readily made to compensate for changes occurring in ingredient properties. Oxidizers, such as bromate and iodate, are added routinely to compensate for the smaller amount of oxygen brought into the dough during mixing. Most of the flavour in continuous-mix breads results from a preferment, called the broth or liquid sponge, a mixture of water, yeast, sugar, and portions of the flour and other ingredients, fermented for a few hours before being mixed into the dough.

Advantages of continuous mixing

In addition to the labour-saving aspects of the continuous-mix process, there are also quality benefits. The grain, or cell structure, of the bread is small and regular, and the loaves are uniform in appearance. The flavour is sometimes considered blander than that of conventionally made loaves, but this characteristic apparently has little effect on consumer acceptance.

The developer, the key equipment in the production line, processing about 100 pounds (45 kilograms) each 90 seconds, changes the batter from a fluid mass having no organized structure, little extensibility, and inadequate gas retention, to a smooth, elastic, film-forming dough. The mechanical effort involved causes a temperature rise of 18° to 24° F (10° to 13° C) in the developer.

Chemically leavened doughs and batters. Many bakery products depend upon the evolution of gas from added chemical reactants as their leavening source. Items produced by this system include layer cakes, cookies, muffins, biscuits, corn bread, and some doughnuts.

The gluten proteins of the flour serve as the basic structural element in chemically leavened foods, just as they do in bread. The relatively smaller amounts of flour, the weaker (less extensible) protein in the soft-wheat flours customarily employed, and the lower protein content of the flour, however, result in a softer, crumblier texture. In most chemically leavened foods, the protein content of the flour, inadequate in quantity and quality to support the amount of expansion required in bread, produces a product of higher density.

Prepared dry mixes, available for home use and for

small- and medium-sized commercial bakeries, vary in complexity from self-rising flour, consisting only of salt, leavening ingredients, and flour, to elaborate cake mixes. Mixes offer the consumer ingredients measured with greater accuracy than possible with kitchen utensils, and special ingredients designed for functional compatibility.

Prepared doughs for such products as biscuits and other quick breads, packaged in cans of fibre and foil laminates, are available in refrigerated form. These products carry the mix concept two steps further; the dough or batter is premixed and shaped. Unlike ordinary canned products, refrigerated doughs are not sterile but contain microbiological organisms from normal ingredient contamination. Spoilage is retarded by low storage temperature, low oxygen tension, and the high osmotic pressure of the aqueous phase.

Drawing by D. Meighan

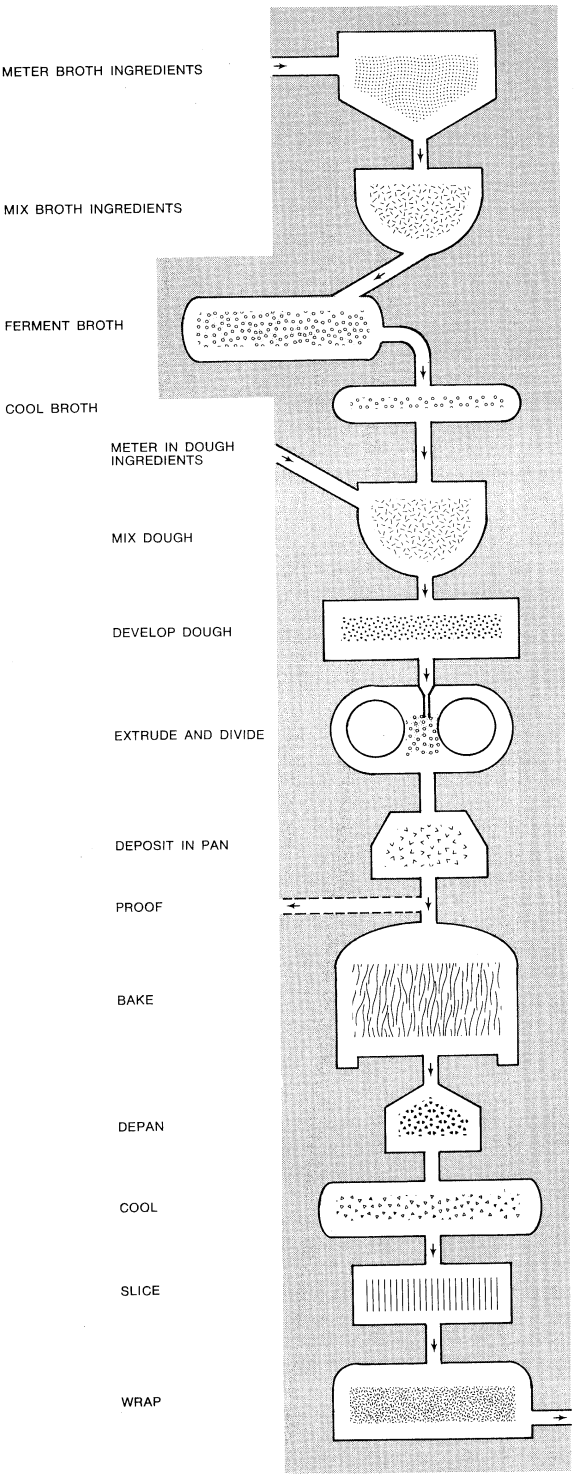


Figure 5: Essential steps in continuous production of bread.

Basic makeup of hot breads

Hot breads, such as biscuits, muffins, pancakes, and scones, constitute a large and important class of chemically leavened bakery foods. They consist of flour, baking powder, salt, and liquid, with varying amounts of eggs, milk, sugar, and shortening. Other variations include the addition of fruits such as raisins, condiments such as peppers, and adjuncts such as cheese. In corn breads a considerable proportion of the flour is replaced by corn meal. Mixing and forming methods, and the baking conditions applied, also affect product appearance, texture, and flavour. For example, a batter suitable for making corn bread might also be used to make muffins or pancakes, and each kind of finished product would vary not only in appearance but also in flavour and texture. Recipes for hot breads usually contain not more than about 15 percent shortening and 5 percent sugar. Eggs, when used, are customarily whole eggs. Milk is often used both for flavour and for its texturizing and crust coloration properties.

Pancake batter formulas are similar to other quick-bread recipes, except for the inclusion of larger amounts of liquid. Some muffins are much like layer cakes in composition, with increased sugar, shortening, and eggs producing more tender texture, sweeter flavour, and lower density. Flavourings, usually constituting only a small portion of the total batter, also differentiate cakes from quick breads. Vanilla is a common flavour in all kinds of cakes, and bitter chocolate (chocolate liquor, baking chocolate) or cocoa, at the 3 or 4 percent level, is used in chocolate or devil's food cakes.

There are traditional rules for assuring "formula balance," or the correct proportioning of ingredients, in layer cakes. For every ten parts of flour, yellow layer cakes should contain 10 to 16 parts sugar by weight, white layer cakes should contain 11 to 16 parts sugar. Shortening should range from three to seven parts for each ten parts of flour. The weight of liquid whole eggs should equal or exceed that of the shortening in the mixture. Total water, including the moisture in eggs and milk, should exceed the amount of sugar by 2½ to 3½ parts. Baking powder weight should equal from 3 to 6 percent of flour weight; salt should equal 3 to 4 percent of flour weight. If the amount of sugar in a formula is increased, the egg content should be increased an equal amount, and more shortening should be added when the percentage of eggs is increased. Additional water is rarely added when the formula contains dry milk, but if the formula water is not sufficient to equal the reconstitution water for the milk, about 1 percent of water for each additional percent of milk solids is added.

Rich formulas (those with large amounts of sugar, shortening, and eggs) are exceptions to the traditional rules, needing less chemical leavening because more air is incorporated during mixing. Their batters display lower specific gravity for the same reason and are baked at a lower temperature; also, batters baked in large-piece sizes require less water and less leavener than those baked in small size pans (*e.g.*, layer cakes as compared with cup cakes).

Common cake varieties include white cake, similar in formula to yellow cake, except that the white cake uses egg whites instead of whole eggs; devil's food cake, differing from chocolate cake chiefly in that the devil's food batter is adjusted to an alkaline level with sodium bicarbonate; chiffon cakes, deriving their unique texture from the effect of liquid shortening on the foam structure; and gingerbread, similar to yellow cake but containing large amounts of molasses and spices.

Variability of cookie recipes

Recipes for cookies (called biscuits or sweet biscuits in some countries) are probably more variable than those for any other type of bakery product. Some layer-cake batters can be used for soft drop cookies, but most cookie formulas contain considerably less water than cake recipes, and cookies are baked to a lower moisture content than any normal cake. With the exception of soft types,

the moisture content of cookies will be below 5 percent after baking, resulting in crisp texture and good storage stability.

Cookies are generally high in shortening and sugar. Milk and eggs are not common ingredients in commercial cookies but may be used in home recipes. Sugar granule size has a pronounced effect on cookie texture, influencing spread and expansion during baking, an effect partly caused by competition for the limited water content between the slowly dissolving sugar and the gluten of the flour.

Equipment. Commercial production methods and equipment for chemically leavened bakery foods are designed to produce specific items, although the horizontal dough mixers used for yeast-leavened products may also be used for mixing chemically leavened doughs and batters. Mixers may be the batch type, similar in configuration to the household mixer, with large steel bowls, open at the top, containing the batter while it is mixed or whipped by beater paddles of various conformations. In continuous mixers, the batter is pumped through an enclosed chamber while a toothed disk rapidly rotates and mixes the ingredients. The chambers may be pressurized to aerate the batter and surrounded with a flowing heat-transfer medium to adjust the temperature.

The shape of layer cakes and other items made from batters results mainly from the form of the container into which the batter is deposited. In production lines shaping or forming equipment is replaced by metering devices extruding the desired volume of batter into each cavity of the baking pan. Metering can be performed by calibrated rotary pumps or by piston-type dispensers. In the production of batter doughnuts, simple extruders form the mixture into ring shape just before it is dropped into the hot fat.

Chemically leavened doughs can be formed by methods similar to those used for yeast-leavened doughs of similar consistency. In the usual sequence, the dough passes between sets of rollers, forming sheets of uniform thickness; the desired outline is cut in the sheet by stamping pressure or embossed rollers; and the scrap dough is removed for reprocessing. Many cookies and crackers are made in this way, and designs may be impressed in the dough pieces by docking pins (used primarily to puncture the sheet, preventing formation of excessively large gas bubbles) or by cutting edges partially penetrating the dough pieces.

Specialized machines designed for the production of fancy sweet goods integrate and automate the operations of sheeting, application of liquid and solid flavouring materials, rolling the sheet into a continuous cylinder, sealing and cutting the roll into sections of the desired size.

Die forming and extrusion of cookies

In addition to the sheeting and cutting methods, cookies may be shaped by die forming and extrusion. In die forming a dough casing may be applied around a centre portion of jam or other material; or portions of dough may be deposited, forming such drop-type cookies as vanilla wafers, chocolate chip, and oatmeal cookies. Extrusion is accomplished by means of a die plate having orifices that may be circular, rectangular, or complex in outline. The mass of dough, contained in a hopper, is pushed through these openings, forming long strands of dough. Individual cookies are formed by separating pieces from the dough strand with a wire passing across the outer surface of the die, or by pulling apart the hopper and oven belt (to which the dough adheres). Fig bars and similar cookies are baked in a continuous strand, then sliced by disk cutters running across the lines of baked, extruded bars as they leave the band oven.

Cookies produced on rotary molders include sandwich-base cakes and pieces made with embossed designs. A steel cylinder, the surface covered with shallow engraved cavities, rotates past the opening in a hopper filled with cookie dough. The pockets are filled with the dough, which is sheared off from the main mass by a blade; and as the cylinder continues its revolution, the dough pieces are ejected onto a conveyor belt leading to the band oven.

Sugar-wafer bases are formed on large griddle-like molds. A thin batter is poured onto the bottom half of a hot mold, and the top section is pressed down, while baking is completed by externally applied heat. The mold is then opened and the wafer ejected.

Most commercial bakery ovens are the band types, although reel ovens are still used, especially in smaller shops or bakeries where short runs are frequent.

Air- and steam-leavened products. The tiny air bubbles mixed into doughs and batters during their preparation, or originating as occluded gases in ingredients such as flour, are highly important as focuses for gas evolution from the dissolved and combined state in both chemically leavened and fermented goods. In addition, the air expands during the baking process, contributing substantially to the leavening action. In foam products, the foams may be the sole source of leavening action, with air entrapped in an agent capable of forming thin vesicle walls (*e.g.*, egg white).

Foams and sponges

The albumen of egg white, a protein solution, foams readily when whipped. The highly extended structure has little strength and must be supported during baking by some other protein substance, usually the gluten of flour. Because the small amount of lipids in flour tend to collapse the albumen foam, flour is gently folded into egg white foams, minimizing contact of fatty substances with the protein. Gluten sponges are denser than the lightest egg-white foams but are less subject to fat collapse. The foam of egg yolks and whole eggs, as in pound cakes, is an air-in-oil emulsion. Proteins and starch, scattered throughout the emulsion in a dispersed condition, gradually coalesce as the batter stands or is heated. Fats and oils, in addition to yolk lipids, can be added to such systems without causing complete collapse but never achieve the low density possible with protein foams and usually have a tender, crumbly texture, unlike the more elastic structure of albumen-based products. Air-leavened bakery products, avoiding the flavours arising from chemical and yeast-leavening systems, are particularly suitable for delicately flavoured cakes. Since the batters can be kept on the acidic side of neutrality, the negative influence of chemical leaveners on fruit flavours and vanilla is avoided.

Rye wafers made of whipped batters are modern versions of an ancient Scandinavian food. High-moisture dough or batter is whipped, extruded onto an oven belt, scored and docked, then baked slowly until almost dry.

Beaten biscuits, an old Southern U.S. specialty, are also made from whipped batter. Air is beaten into a stiff folded dough with many strokes of a rolling pin or similar utensil. Round pieces cut from the dough are pricked with a fork to prevent development of large bubbles, then baked slowly. The baked biscuit is similar to a soft cracker.

All leavened products rely to some extent on water-vapour pressure to expand the vesicles or gas bubbles during the latter stages of baking, but some items also utilize the leavening action produced by the rapid build-up of steam as the interior of the product reaches the boiling point. These foods include puff pastries, used for patty shells and napoleons, and chou pastes, often used for cream-puff cases.

Puff pastry, often used in French pastries, is formed from layered fat and dough. The proportion of fat is usually high, rarely less than 30 percent of the finished raw piece. The dough should be extensible but not particularly elastic; for this reason mixtures of hard and soft wheat flour are often used. The fat should have an almost waxy texture and must remain solid through the sheeting steps. Butter, although frequently used, is not particularly suitable for puff pastry because of its low melting point. Bakers specializing in puff pastry often use special margarines containing high-melting-point fats.

Puff pastry production methods

There are several methods of making puff pastry. In the basic procedure dough is rolled into a rectangular layer of uniform thickness, and the fat is spread over two-thirds of the surface. Although the usual practice is to place lumps of margarine on the dough, a more efficient practice is to place fat slices of equal thickness close to-

gether on the dough surface. The dough is next folded, producing three dough strata enclosing two fat layers. This preparation is next chilled by refrigeration, then rolled, reducing thickness until it reaches approximately the area of the original unfolded dough. The folding, refrigeration, and rolling procedure is repeated several times, and after the final rolling the dough is reduced to the thickness desired in the shaped raw piece.

The goal of this process is reduction of the layers of fat and dough to the minimum thickness consistent with their retention in continuous sheets. Refrigeration helps maintain the fat in solid form throughout the rolling steps; rolling would otherwise tend to mix the fat into the dough. An initial baking period at high temperature causes the puffing and is followed by a lower temperature stage to complete drying.

Correctly prepared puff pastry will expand as much as ten times during baking because of the evolution of large volumes of steam at the interface between shortening and dough. The focuses for production are the microscopic air bubbles rolled into the dough during the layering process. If layering has been properly conducted, the finished pieces will be symmetrical and well shaped, with crisp, flaky outer layers.

Chou paste, used for cream puffs, is made by an entirely different method. Flour, salt, butter, and boiling water are mixed together, forming a fairly stiff dough, and whole eggs are incorporated by beating. Small pieces of the dough are baked on sheets, initially at high temperature. The air bubbles formed during mixing expand rapidly at baking temperatures, filling the interior with large, irregular cells, while the outside browns and congeals, forming a rather firm case. The interior, largely hollow, can be injected with such sweet or savoury fillings as whipped cream or shrimp in sauce.

Unleavened bakery products. The tortillas of Central America and the chapatties of India are later examples of ancient flatbread, which was, for all practical purposes, unleavened. Pie crusts, the major volume item in this category prepared by modern bakeries, usually are unleavened. Small amounts of baking powder or soda are sometimes added to pie-crust doughs, mostly in domestic cookery. This addition, although increasing tenderness, tends to eliminate the desirable flakiness and permits the filling liquid to soak into the crust more rapidly.

Pie crusts are usually simple mixtures of flour, water, shortening, and salt. The shortening proportion is about 30 to 40 percent of the dough. The amount of water is kept low, and the mixing process is kept short to minimize development of elasticity, leading to shrinkage and development of toughness on baking. For flaky crust, the fat should not be completely dispersed through the dough, but should remain in small particles. Commercial producers often employ special mixers using reciprocating, intermeshing arms to gently knead the dough. The doughs are chilled before mixing and forming, to reduce smearing of the shortening.

Flakiness is also related to the type of shortening used. Lard is popular in home cookery for this reason, and also because of its satisfying flavour. Because shortening should be solid at the temperature of mixing, oils are undesirable.

Milk or small amounts of corn sugar may be added to improve crust browning and for their flavour effect. About 1 to 2 percent of the dough will be salt.

MARKETING PREPARATION

Slicing of bread

Bread often is marketed in sliced form. Slicing is performed by parallel arrays of saw blades through which the loaves are carried by gravity or by conveyors. The blades may be endless bands carried on rotating drums, or relatively short strips held in a reciprocating frame. Most bread is sliced while still fairly warm, and the difficulty of cutting the sticky, soft crumb has led to development of coated blades and blade-cleaning devices. Horizontal slicing of hamburger rolls and similar products is accomplished by circular (disk) blades, usually two blades in a slicer, between which four or six rolls are carried by a belt. The cutter blades are separated to avoid cutting completely through the roll, in order to leave a "hinge."

Freezing is an indispensable bakery industry process. Ordinary bread and rolls are rarely distributed and sold in frozen form because of the excessive cost in relation to product value, but a substantial percentage of all specialty products is sold in frozen form. Most bakery products respond well to freezing, although some cream fillings must be especially formulated to avoid syneresis, or gel breakdown. Rapid chilling in blast freezers is preferred, although milder methods may be used. Storage at 0° F (−18° C) or lower is essential for quality maintenance. Thawing and refreezing is harmful to quality. Frozen bakery products can dehydrate under freezer conditions and must be packaged in containers resistant to moisture-vapour transfer.

The use of doughs and batters frozen and stored for later baking facilitates both commercial bakery production and distribution to consumers. Products can be prepared in advance when labour is available, stored frozen, then baked as required to meet customer demand. The unbaked dough can be prepared in central plants, frozen, and transferred to small branch bakeries, reducing the equipment and skilled labour requirements to make freshly baked foods available at branch outlets.

Most U.S. consumers prefer wrapped bread, and the trend toward wrapping is growing in other countries. Sanitary and aesthetic considerations dictate protection of the product from environmental contamination during distribution and display. Waxed paper was originally the only film used to package bread; then cellophane became popular; and polyethylene and combination laminates became common in the late 1960s. Other bakery products are packaged in a variety of containers ranging from open bags of greaseproof material to plastic trays with sealed foil overwraps.

Canning bakery products

In the early 1970s the market for bakery products in tin cans was small. Hunters and campers found canned foods convenient. Canning protects against drying and environmental contamination, but at a cost in texture staling and some degree of flavour staling. In processing, an amount of dough or batter known to fill exactly the available space after baking is placed in a can and the cover is loosely fastened to allow gases to escape. The product is then baked in a conventional oven; the lid is hermetically sealed immediately after baking; and the sealed can is sprayed with water to cool it. Vacuums of 25 inches or more (needed to assure storage stability) can be routinely achieved by this method. Special can linings and sealing compounds are needed to survive oven temperatures, and the exterior must be dark coloured (*e.g.*, olive drab) in order to absorb radiant heat in the oven, avoiding long baking times. Spores of some pathogens are not killed by the conditions reached in the centre of the baked product, but pH and osmotic pressure can be adjusted to prevent growth of spoilage organisms. There is no record of food poisoning attributable to canned bakery food.

QUALITY MAINTENANCE

Bakery products are subject to the microbiological spoilage problems affecting other foods. If moisture content is kept below 12 to 14 percent (depending on the composition), growth of yeast, bacteria, and molds is completely inhibited. Nearly all crackers and cookies fall below this level, although jams, marshmallow, and other adjuncts may be far higher in moisture content. Breads, cakes, sweet rolls, and some other bakery foods may contain as much as 38 to 40 percent water when freshly baked and are subject to attack by many fungi and a few bacteria.

To obtain maximum shelf life free of mold spoilage, high levels of sanitation must be maintained in baking and packing areas. Oven heat destroys all fungal life forms, and any spoilage by these organisms is due to reinoculation after baking. Strict observation of sanitary practices, including the use of fungicides and ultraviolet

lights where possible, reduces contamination to a low level. Mold spores cannot be entirely eliminated, since they are ubiquitous and can be drawn into the package from the outside atmosphere after the package leaves the bakery.

A number of compounds have been proposed for use as fungistats in bread. Some have proven to be innocuous to molds, toxic to man, or both. Soluble salts of propionic acid, principally sodium propionate, have been accepted and extend shelf life appreciably in the absence of a massive inoculum. Acetic acid also has a protective effect.

Bacterial bread spoilage

Bacteria associated with bread spoilage include *Bacillus mesentericus*, responsible for "ropy" bread, and the less common but more spectacular *Micrococcus prodigiosus*, causative agent of "bleeding bread." Enzymes secreted by *B. mesentericus* change the starch inside the loaf into a gummy substance stretching into strands when a piece of the bread is pulled apart. In addition to ropiness, the spoiled bread will have an off-aroma sometimes characterized as fruity or pineapple-like. Formerly, when ropiness occurred, bakers acidified doughs with vinegar as a protective measure, but this type of spoilage is rare in bread from modern bakeries.

Micrococcus prodigiosus causes red spots to appear in bread. At an advanced stage these spots of high bacterial population may liquefy, emphasizing the similarity to blood, which has sometimes led the superstitious to attribute religious significance to this phenomenon. The organism will not survive ordinary baking temperatures, unlike *B. mesentericus*, which forms spores capable of survival in the centre of the loaf, where the temperature rises only to about 212° F (100° C).

Neither ropy nor bleeding bread is particularly toxic to man; the only widespread food poisoning in which bread has been a vector has resulted from ergot, a fungus infection of rye. Ergot contamination of bread made from rye, or from blends of rye and wheat, has caused epidemics leading to numerous deaths.

Baked goods containing such high-moisture adjuncts as pastry creams and pie fillings are susceptible to contamination by food-spoilage organisms, including *Salmonella* and *Streptococcus*. Cream and custard pies are recognized health hazards when stored at room temperature for any length of time, and some communities ban their sale during summer. Storage in frozen form eliminates the hazard.

Undesirable changes in bakery products can occur independently of microbial action. Staling involves changes in texture, flavour, and appearance. Firming of the interior, or "crumb," is a highly noticeable alteration in bread and other low-density, lean products. Elasticity is lost and the structure becomes crumbly. Although loss of moisture produces much the same effect, texture staling can occur without any appreciable drying. Such firming is due to changes in the molecular status of the starch, specifically to a kind of aggregation of sections of the long-chain molecules into micelles, making the molecules more rigid and less soluble than in the newly gelatinized granule. Stale bread can be softened to a state approximating the texture of fresh bread by heating to about 140°–150° F (60°–65° C). Care must be exercised to prevent drying during heating.

Starch retrogradation, the cause of ordinary texture staling of the crumb, can be slowed by the addition of certain compounds to the dough. Most of the effective chemicals are starch-complexing agents. Monoglycerides of fatty acids have been widely used as dough additives to retard staling in the finished loaf.

The crust, dry and crisp in the fresh state, becomes soft and leathery when stale, due to redistribution of moisture between the loaf components. Crust softening is accelerated when the product is enclosed in moisture-proof packaging.

Flavour changes during storage

Changes in flavour occur during the storage of baked products. Some volatile substances evaporate; others oxidize or react with other dough constituents to yield less flavourful compounds. Ethanol contributes to the aroma of freshly baked yeast-leavened products; once it is lost, reheating cannot revive the original flavour. Fortunately, most volatile compounds are adsorbed on other bread constituents and are thus hindered from diffusing away.

QUALITY TESTING

Bakery food testing is mainly oriented toward the determination of sensory attributes. Expert evaluation of the appearance, texture, and flavour of finished products is essential in quality control, but objective measurements are necessary for both finished goods and ingredients.

Bread scoring is a daily bakery routine. Plant production executives and quality-control laboratory personnel assign numerical scores to the volume, shape, crust colour, crumb texture and colour, aroma, taste, and other characteristics of cut loaves representative of the day's production. The numerical scores are usually entered on a form sheet listing the quality factors. The final score is totalled and compared with a minimum representing lowest loaf quality permissible for sale.

Cakes and pies, less uniform and more difficult to evaluate on a rigid scale than other bakery products, are graded less formally, generally on a pass or fail basis.

Objective tests conducted on finished products include chemical analyses for moisture content and the amounts of such important ingredients as shortening, sugar, vitamins and minerals, and physical measurements of density, colour, and texture. In general, the chemical tests are much the same as those conducted on other foods. Moisture content for ordinary white bread is kept just below 38 percent by most producers to meet legal requirements and still obtain the soft, elastic texture preferred by the majority of consumers.

Testing bread texture

Attempts have been made to devise objective tests of bread texture. An instrument designed to measure the yielding of crumb to compression under controlled load, the Baker Compressimeter, applies force from a synchronous motor to a compound lever forcing a plunger of square section into the surface of a slice of bread. A spring and scale arrangement in the lever system measures the deformation of the bread under constant load, or the load required to cause a constant deformation. Other devices simulate chewing to measure the force required to tear the crumb.

Reflectance spectrophotometers provide objective measurement of crust and crumb colour. Standard instruments must be modified to scan areas sufficiently large to average out small-scale nonuniformities.

There are special tests for functional quality applicable to bakery product ingredients. Flour, shortening, leavening, and dried milk are the principal ingredients affecting dough performance during processing. Much experimentation has been applied to development of analytical methods to predetermine accurately the suitability of a flour for a given use. The only conclusive test, however, is use of the flour under conditions encountered during normal processing. Since large-scale tests of this type are expensive, inconvenient, and time-consuming, small-scale evaluations duplicating production conditions as closely as possible usually are substituted. Such empirical procedures are useful, although subject to many theoretical objections.

Instruments have been devised to measure physical properties of standard doughs, including resistance to intense sheer, elasticity of a dough cylinder under elongation, and maximum size of a bubble blown from a film of dough.

Diastatic activity—the rate at which starch is converted to reducing sugars by the enzymes present in the flour or in added malt—governs the gas-production rate, especially in the later stages of dough processing. It therefore affects the inner structure and volume of the loaf or roll and often influences crust colour. Diastatic activity, affected both by amylolytic enzyme activity and the susceptibility to attack of the starch granules, can be expressed as the amount of reducing sugar developed in a given time at controlled temperatures by a standard mixture of flour and water.

Quality standards established by government

Governments may establish quality standards for various bakery products, issuing product definitions and composition requirements, specifying weights, requiring certain hygienic measures, stating labelling requirements, and providing for testing and sampling methods. For example, in the U.S., the Federal Standard of Identity for Bread and Rolls establishes the composition of white, enriched, milk, raisin, and whole-wheat breads, as well as that of the same types of rolls and buns. Virtually all countries with modern baking industries regulate the size and composition of the principal bakery products (*e.g.*, bread).

BAKERY PRODUCTS IN THE HUMAN DIET

Bakery-product consumption varies so greatly among cultures that it is not possible to generalize on a worldwide basis. The United States has shown a steady downward trend in per capita consumption of cereal-based foods for many years, and the consumption of sweet bakery foods, such as cakes and pastry, has been increasingly displacing bread consumption. The worldwide consumption of bakery products is increasing. Many populations that formerly relied solely on rice or coarser grains as their main source of carbohydrates show a preference for compounded bakery products as new industry and increased incomes make them more widely accessible. Japan is an outstanding example of countries following this trend.

Probably 95 percent of the white bread sold in the United States is enriched with thiamine, niacin, riboflavin, and iron, and about 30 states have laws requiring white-bread enrichment. The most common method of vitamin and mineral supplementation is the addition of an enrichment tablet to the dough mixture, but enriched salt and packets formed of water-soluble, edible films are also used. Formerly, calcium and vitamin D were frequently added to enriched bread, but lack of consumer demand and questionable nutritional benefit led to gradual phasing-out of these enrichments.

Lower income families have the highest per capita consumption of bakery foods in the U.S.; a considerable portion of their caloric intake, as well as much of their calcium, iron, thiamine, and protein, comes from this source. Consumption of enriched flour, cornmeal, and bread continues to decrease, however, while unenriched baked-goods consumption is increasing. The contribution of cereals to the diet could be improved by enrichment of consumer mixes and sweet bakery goods. Most bakery products, composed primarily of cereal derivatives, are subject to many of the nutritional generalizations applied to cereals.

Home baking has been in a continual decline for decades, decreasing the sale of enriched flour at the retail level. Cornmeal usage tends to decline as incomes increase and as exposure to convenience foods and to more elaborate cuisines becomes more general. Calories consumed as starchy or bulk foods decrease and "sweet goods," confectionery, or rich bakery items increase as disposable income increases.

Protein concentrates suitable for bakery-product fortification can be made from a wide variety of raw materials. Oilseeds are the largest single protein concentrate source; milk is another excellent source of supplementary protein for bakery foods. India has attempted to encourage consumption of protein-enriched bread, with some success. Except for research projects, little has been done in other countries, however.

Feeding trials show that, with the addition of about 0.5% L-lysine and 0.1% to 0.2% threonine, the nutritional value of the protein in bread would equal that of casein. Addition of nonfat-milk solids to bread is an effective but relatively expensive way of adding the needed lysine, and production of bread of normal appearance and texture is difficult if the dough contains more than 6 percent of skim-milk solids, an insufficient amount to bring the lysine content to the desired level.

In areas where bakery products comprise a large proportion of the diet, they provide an ideal vehicle for nutritional supplementation. The need for additional or better quality protein in the daily diet occurs mostly in the developing countries, where carbohydrate foods—usually cereals—are the diet staples and logical protein carriers. Adding protein supplements to compounded bakery foods, such as bread, is less complicated than adding these enrichments to grains, such as rice, but unfortunately, compounded bakery foods are not much used in many developing countries.

BIBLIOGRAPHY. General discussions of the properties of cereal grains including quality factors affecting their suitability for milling and baking may be found in N.L. KENT, *Technology of Cereals; with Special Reference to Wheat* (1966); S.A. MATZ (ed.), *Chemistry and Technology of Cereals as Food and Feed* (1959), *Cereal Science* (1969), and *Cereal Technology* (1970). Specialized in-depth discussions of the methods, materials, and products of commercial baking may be found in S.A. MATZ, *Bakery: Technology and Engineering* (1960); and E.J. PYLER, *Baking Science and Technology*, 2 vol. (1952).

(S.A.M.)

Bakunin, Mikhail Aleksandrovich

Mikhail Bakunin, a prominent Russian revolutionary agitator and the chief propagator of 19th-century anarchism, was born on May 30 (new style; May 18, old style), 1814, the eldest son of a small landowner of Premukhino in the province of Tver (now Kalinin). He grew up in idyllic surroundings, romantically devoted to four sisters who were nearer to him in age than his younger brothers. His lifetime of revolt began when he was sent to the Artillery School in St. Petersburg and later was posted to a military unit on the Polish frontier. In 1835 he absented himself without leave, and resigned his commission, narrowly escaping arrest for desertion. For the next five years he divided his time between Premukhino, where he plunged into the study of the German philosophers Johann Fichte and Hegel, and Moscow, where he moved in the literary circles of the critic V.G. Belinsky, the novelist Ivan Turgenev, and the publicist Aleksandr Herzen. In 1840, his opinions still in a state of fluid turbulence, he journeyed to Berlin to complete his education. There he fell under the spell of the Young Hegelians, the radical followers of Hegel, and, having moved to Dresden, in 1842 published in a radical journal his first revolutionary credo, ending with the now-famous aphorism: "The passion for destruction is also a creative passion." This brought him a peremptory order to return to Russia and, on his refusal, the loss of his passport. After brief periods in Switzerland and Belgium, Bakunin settled in Paris, where he consorted with French and German Socialists, including Pierre-Joseph Proudhon and Karl Marx, and with numerous Polish émigrés who inspired him to combine the cause of the national liberation of the Slav peoples with that of social revolution. The February Revolution of 1848 in Paris gave him his first taste of street fighting; and after a few days of eager participation he travelled eastward in the hope of fanning the flames in Germany and Poland. In Prague in June 1848, he attended the Slav congress, which ended when Austrian troops bombarded the city; and later in the year, in the secure retreat of Anhalt-Köthen, in Germany, he wrote his first major manifesto, *An Appeal to the Slavs.* He denounced the bourgeoisie as a spent counter-revolutionary force; he called for the overthrow of the Habsburg Empire and the creation in central Europe of a free federation of Slav peoples; and he counted on the peasant and especially on the Russian peasant, with his tradition of violent revolt, as the agent of the coming revolution.

Early revolutionary activity

Tired of inaction, Bakunin once more plunged into revolutionary intrigues and, engaging in the Dresden insurrection of May 1849, failed this time to escape arrest. The Saxon authorities handed him over to Austria and Austria, after a further period of incarceration, to Russia. In May 1851 he was back on Russian soil in the Peter-Paul Fortress in St. Petersburg. There, at the invitation of the chief of police, he wrote an enigmatic *Confession,* which was not published until 1921. Much of it consists of expressions of repentance for misdeeds and abject ap-



Bakunin.
EB Inc.

peals for mercy. But it includes some gestures of defiance and plays heavily on Bakunin's devotion to the Slavs and hatred of the Germans—sentiments that were noted with interest and approval by the Tsar. They did not, however, help the prisoner. He remained for three years in the Peter-Paul Fortress and for three further years in another fortress, the Schlisselburg, in conditions of rapidly deteriorating health. Finally, in 1857 he was released to live in Siberia. There he contracted a marriage, which was not consummated, with the daughter of a Polish merchant. The governor of Eastern Siberia was a cousin of Bakunin's mother, and it was probably through this connection that he obtained permission in 1861 to travel down the Amur, ostensibly on commercial business. Having reached the coast in a Russian ship, he transferred to an American vessel bound for Japan and travelled via the United States to Great Britain.

Bakunin's arrival in London at the end of 1861 reunited him with Herzen, whom he had last seen in Paris in 1847 and who now occupied a pre-eminent position among Russian émigrés as editor of *Kolokol* ("The Bell"). Bakunin's 14-month stay in London led to an irretrievable rift with Herzen, who had shed some of the revolutionary ardour of his youth and had already crossed swords with the critic and novelist Nikolay Chernyshevsky and other extreme radicals of the rising Russian generation. He now found Bakunin's financial, as well as political, irresponsibility hard to bear. When the Polish insurrection broke out early in 1863, Bakunin eagerly embarked with a shipload of Polish volunteers for the Baltic. He got only as far as Sweden, where he spent a fruitless summer. At the beginning of 1864 he established himself in Italy, which became his residence for four years. It was there that he framed the main outlines of the anarchist creed that he preached with unsystematic but unremitting vigour for the rest of his life. It was there, too, that he began to weave that complex network, part real, part fictitious, of interlocking secret revolutionary societies that absorbed his energies and bewildered the followers whom he enrolled in them.

Quarrel with Marx

The most famous episode of Bakunin's later years was his quarrel with Marx. In 1868, then settled in Geneva, he joined the First International, a federation of working-class parties aiming at transforming the capitalist societies into Socialist commonwealths and their eventual unification in a world federation. At the same time, however, he enrolled his followers in a semisecret Social Democratic Alliance, which he conceived as a revolutionary avant-garde within the International. The same organization could not hold two such powerful and incompatible personalities; and at The Hague congress in 1872 Marx, by an intrigue that had little relation to the causes of the quarrel, secured the expulsion of Bakunin and his followers from the International. The breach split the revolutionary movement in Europe for many years to come. Two of Bakunin's major writings, *The Knouto-Germanic Empire* (1871) and *State and Anarchy* (1873), directly reflected his conflict with Marx. Bakunin was as uncompromising a revolutionary as Marx and never ceased to preach the overthrow of the existing order by violent means. But he rejected political control, centralization, and subordination to authority (while making an unconscious exception of his own authority within the movement). He denounced what he regarded as characteristically Germanic ways of thought and organization and opposed to them the untutored spirit of revolt that he found embodied in the Russian peasant. Bakunin's anarchism took final shape as the antithesis of Marx's communism.

During his last years, which he spent in penury in Switzerland, Bakunin reverted to his preoccupation with central and eastern Europe. He was compromised by a short-lived enthusiasm for S.G. Nechayev, a young Russian nihilist who paraded his contempt for conventional morality, achieved notoriety by murdering a fellow conspirator whom he suspected of intending to betray or desert the cause, and for this crime was eventually extradited to Russia by the Swiss authorities. Bakunin consorted with Russian, Polish, Serb, and Romanian émigrés, among whom he found eager disciples; drafted proclamations; and planned revolutionary organizations. His health grew worse; his financial embarrassments became ever more acute, and he depended on the bounties of a few Italian and Swiss friends. But he never wholly lost the resilience of his revolutionary convictions. He died in Bern on July 1 (June 19, O.S.), 1876, aged 62.

Proudhon and Bakunin rank as the founding fathers of 19th-century anarchism. Bakunin formulated no coherent body of doctrine. His voluminous and vigorous writings were often left incomplete. But his fame and personality inspired a large and widely dispersed following. Small anarchist groups existed in Great Britain, Switzerland, and Germany, whereas the powerful anarcho-syndicalist wing of the French trade unions owed more to Proudhon than to Bakunin. Anarchist movements owing allegiance to Bakunin continued to flourish in Italy and especially in Spain, however, where as late as 1936 the anarchists were the strongest revolutionary party.

BIBLIOGRAPHY. The first collection of Bakunin's writings was published in six volumes in French between 1895 and 1913; a complete Russian edition of writings and letters was planned, but only the first four volumes down to 1861 were published (1934–36). A complete edition of the later writings, edited by ARTHUR LEHNING, is in course of publication by the Instituut voor Sociale Geschiedinis, Amsterdam. Bakunin has not been well served by biographers: MAX NETTLAU'S manuscript biography (in German), *Michael Bakunin*, 3 vol. (1896–1900), based on a large mass of documents, was distributed to the main European libraries; Y. STEKLOV'S biography (in Russian), 4 vol. (1926–27), is unfriendly and Marxist; see also E.H. CARR, *Michael Bakunin* (1937).

(E.H.C.)

Balanchine, George

As artistic director and chief choreographer of the New York City Ballet, George Balanchine, in the 20th century, became the most influential name associated with classical ballet choreography in the United States. Working in the classic dance idiom, he created many of his finest ballets, such as *Serenade* (first performed 1935), *Symphony in C* (1948), and *Agon* (1957), with no dramatic themes but with highly original and imaginative dance patterns.

Born on January 9, 1904, in St. Petersburg (Leningrad), of a Georgian family, Georgy Melitonovich Balanchivadze was one of a generation of dancers who spent the World War I years at the Imperial School of Ballet at the Mariinsky Theatre. The theatre closed for some months in 1917, and, until the Imperial School reopened in 1918 as the Soviet State School of Ballet, Balanchine had to support himself with unskilled jobs or by playing

piano in a cinema. After three more years of study, he graduated. The son of a composer, Balanchine had also studied music at the Petrograd (St. Petersburg) Conservatory (1921–24).

As a student Balanchine had already tried choreography. His first work, as early as 1920, was a short piece danced to Anton Rubinstein's *Nuit*. He also choreographed works for evenings of experimental ballet performed by himself and his colleagues at the State School of Ballet. The State's directors discouraged this activity, however. He mounted some new and experimental ballets for the Mikhailovsky Theatre in Petrograd. Among them were *Le Boeuf sur le toit* (1920) by Jean Cocteau and Darius Milhaud and a scene for *Caesar and Cleopatra* by George Bernard Shaw.

Balanchine was one of the first ballet dancers to leave the Soviet Union, initially to tour with a small group, the Soviet State Dancers, which included Aleksandra Danilova, Tamara Gevergeva (later Geva), Nicolas Efimov, and himself. They toured Germany, London, and Paris, where in June 1925 Balanchine joined Sergey Diaghilev's Ballets Russes. (It was Diaghilev who simplified Balanchivadze to Balanchine.)

Period with Diaghilev

It was as a choreographer that Diaghilev envisaged Balanchine—Bronisława Nijinska had recently left Diaghilev and Balanchine assumed her duties—and in 1925 the Ballets Russes danced Balanchine's *Barabau*, the first of ten ballets Balanchine was to mount for Diaghilev. Of the ballets he choreographed for Diaghilev, two survive notably in the world repertoire: *Apollo* (1928), the first example of his individual neoclassical style, and *Le Fils prodigue* (*The Prodigal Son*, 1929).

Martha Swope



Balanchine Instructing Allegra Kent, *c.* 1960.

When Diaghilev died in 1929, Balanchine was sufficiently established to have little difficulty in continuing as a choreographer and ballet master. He worked successively with the Royal Danish Ballet and with the Ballet Russe de Monte Carlo, adding significantly to his reputation by composing *La Concurrence* (1932) and *Cotillon* (1932). In 1933 he was one of the founders of the avant-garde company Les Ballets 1933, whose work so enormously impressed the American dance enthusiast, Lincoln Kirstein, that he invited Balanchine to organize the School of American Ballet and the American Ballet company (of which Kirstein was cofounder and director), thus beginning the association of "Mr. B.," as the ballet world knows him, and the U.S.A.

The American Ballet became the resident ballet company at the Metropolitan Opera in New York, and, during its tenure there, Balanchine produced among other works *Le Baiser de la fée* (1937; *The Fairy's Kiss*). He was also creative in a totally different sphere, as pioneer choreographer for Broadway musicals and Hollywood movies, including the celebrated *Slaughter on Tenth Avenue* ballet in *On Your Toes* (1936).

The end of the largely unsatisfactory association between the American Ballet and the Metropolitan Opera came in 1938. Kirstein founded Ballet Caravan in 1936, with a repertoire of ballets by American choreographers. In 1941 this company and what remained of the American Ballet were united for a Latin American tour, for which Balanchine composed *Concerto barocco* and *Ballet Imperial*. During the World War II period, Balanchine worked in the U.S. for the Original Ballet Russe, Ballet Russe de Monte Carlo, and in Hollywood or on Broadway. In 1947 he was guest ballet master at the Paris Opéra.

Kirstein's determination, however, to establish American ballet under Balanchine's artistic direction never faltered. In 1946 he succeeded in founding the Ballet Society, which developed in 1948 into the New York City Ballet. First centred at the New York City Center and later at the New York State Theatre at Lincoln Center, this company has become particularly identified with Balanchine. A prolific creator in various styles, he has been responsible for most of the New York City Ballet's extensive repertoire. Among the works choreographed for the company were the full-length versions of *The Nutcracker* (1954) and *Don Quixote* (1965).

In 1964 the U.S. dance world was stirred when the Ford Foundation, having granted nearly $8,000,000 to strengthen professional ballet in the United States, presented the entire amount to the New York City Ballet, its affiliated School of American Ballet, and six other ballet companies—all under the direct or indirect influence of Balanchine.

Although he has worked mainly in the U.S., Balanchine is very much an international choreographer, and almost every leading ballet company in the world has mounted at least one of his ballets. Best known abroad are his interpretations of musical compositions, either in a serious vein, such as Brahms' *Liebeslieder Walzer* (1960), or broadly comic, such as Hershy Kay's *Western Symphony* (1954).

Association with Stravinsky

Balanchine had a special artistic relationship with the composer Igor Stravinsky. Stravinsky's connection with the ballet started with Diaghilev, and Balanchine's first association with his music was in choreographing a new version of *Le Chant du rossignol* (*The Song of the Nightingale*) for the Ballet Russe in 1925. A long series of Stravinsky–Balanchine ballets followed; some of them were composed in collaboration.

Other modern composers whose music Balanchine has set to the dance are Arnold Schoenberg (*Opus 34*) and the American composer Charles Ives (*Ivesiana*).

Balanchine studies his scores intensively as a preliminary to composition and begins dance creation only at the first rehearsal. He has said that his ideas come from working with his dancers, but he rarely discusses his ideas with them. He invents rapidly and without indulging in fits of temperament. This approach, together with a predominating impression of cool intellectuality rather than warm emotion in the body of his work, has established Balanchine, to an outside view, as a slightly remote and superhuman personality. The other side of the coin shows a man who has enjoyed working with every kind of musical entertainment in the theatre and in motion pictures.

BIBLIOGRAPHY. An autobiographical chapter is contained in *Balanchine's New Complete Stories of the Great Ballets*, ed. by FRANCIS MASON (1968). See also BERNARD TAPER, *Balanchine* (1962), a full and interesting, if slightly adulatory, account of the man and his work; and ANATOLE CHUJOY, *The New York City Ballet* (1953), an account to date of publication of Balanchine and his company.

(K.S.W.)

Balaton, Lake

Lake Balaton, the largest lake of central Europe, is located in central Hungary, at 46° 50′ N and 17° 45′ E. It has an area 230 square miles (596 square kilometres) wide, and extends for 48 miles (78 kilometres) in a south-

west to northeast direction along the southern foothills of the Bakony Mountains of Hungary. Its average depth is only 10.8 feet (3.3 metres), with a maximum depth of 37 feet (11 metres). The eastern third of the basin is separated to some extent by the Tihany Peninsula, which projects from the northern shore, narrowing the lake at this point to about 1 mile (1.5 kilometres). The Zala River, entering the lake from the eastern Alpine spurs to the west, provides the largest inflow of water, which it transports at the low water-freight rate of six cubic metres per second. Water outflow is through the sluice gates of Siófok, toward the eastern end of the lake, and the entire contents of the lake are replenished about every two years.

Scientific study. The earliest researches on Lake Balaton were carried out by geologists, notably by Lajos Lóczy (1849–1920), who published the first detailed topographical and geological map of the region. The Biological Research Institute at Tihany (established 1927) is the main centre for hydrochemical and biological studies, supplemented by the activities of the Research Institute for Water Resources, which was opened in Balatonszemes in 1952. In 1966 the Hungarian State Geological Institute organized a team for the study of the engineering geology of the Balaton environment.

Geological background. The bed of the Balaton is relatively young; it was formed at the end of the Pleistocene Epoch, less than 1,000,000 years ago, when a structural downwarp of the underlying rocks occurred, caused by the earth movements that dissected the nearby foothills with numerous cross-faults. Originally, there were five small lakes that extended in a north–south chain, but these coalesced when the erosive action of wind, rain, and ice broke down the dividing ridges. Traces of these former lakes can still be seen in the configuration of Lake Balaton today, and the Tihany peninsula is the remnant of one of the dividing ridges.

Origin of the lake

The northern border of the lake is hilly and formed of much older rocks, dating from the Paleozoic and Mesozoic eras, of up to 500,000,000 years ago. The lake itself is bordered by a narrow strip of sedimentary rocks of the Quaternary Period, less than 2,000,000 years old. To the northwest, the Tapocla Bay area is surrounded by hard basalt caps—remnants of volcanic eruptions that took place in the late Pliocene Epoch, 7,000,000 years ago. The area bordering the lake to the south is flat and covered by sand and loess—windblown deposits—of Pleistocene age.

Climate and hydrology. The climate of the region is rather continental, with warm and sunny weather prevailing from May to October. In summer the temperature of the lake ranges from 75° to 82° F (24° to 28° C), but in winter the lake is covered with a sheet of ice about 8 inches (20 centimetres) thick. As the prevailing winds are from the northwest, the southeastern shore of the lake is subject to erosion of its banks by wave action. Oscillations in the levels of the water surface known as seiches, the product of local variations of atmospheric pressure aided by currents in the water, increase the erosive effect. In the Tihany Narrows, the water currents flow with a speed of up to 5 feet (1½ metres) per second. This almost incessant flow scours the lake bed at the tip of the Tihany Peninsula, keeping the two sections of the lake connected with each other.

The chemical composition of the lake differs greatly from that of most central European lakes. The predominant anions, of negatively charged chemical components, are carbonate and sulphate; while the corresponding cations, or positive components, are magnesium, calcium, and sodium: their interaction has given the lake its sulfocarbonate character.

Plant and animal life. The regions around the lake are inhabited by a rich and interesting variety of plant and animal life. There is a wildlife reserve in the Tihany Peninsula, and another one in the extensive reedbeds near Keszthely, where rare water birds nest. The loess region at the southern border of the lake is very fertile, and the volcanic soils to the northwest form the basis of a noted wine-growing region.

Agriculture has nevertheless become less significant as a result of the development of the tourist industry in the decades of the 1960s and 70s. A number of watering places sprang up, notable among which were Siófok, on the southern shore, and Balatonfüred, on the northern shore. The town of Balatonfüred is also famous as a health centre, with medicinal springs for the treatment of heart diseases, and near Keszthely the thermal spring of Hévíz reached temperatures of 97° F (36° C). The oldest and best known settlement is Tihany, noted for its museum and biological station. In all of the resorts, numerous modern hotels have been constructed in recent years, and the annual number of visitors from Hungary and elsewhere in Europe was in excess of 2,000,000 by the start of the 1970s. Watersport and fishing—particularly for the *fogas*, a local delicacy similar to a pike in appearance—were among the main attractions of the lake and its environs.

Development of tourism

BIBLIOGRAPHY. Information on Lake Balaton may be found in J. REISMANN, *Der Balaton* (1962), a travel guide in German; and in two articles in the Hungarian journal, *Földrajzi Közlemények* (1958): B. BULLA, "Geographical Exploration of the Lake Balaton and its Surroundings," and I. KAKAS, "Climatic Problems of Lake Balaton, Hungary" (both with English summaries).

(I.F.)

Balboa, Vasco Núñez de

Vasco Núñez de Balboa, Spanish conquistador and explorer and the European discoverer of the Pacific Ocean, ranks with Christopher Columbus, Hernán Cortés, and Francisco Pizarro among the great leaders who founded the Spanish Empire in America. His discovery of the Pacific led to the discovery and conquest of the Inca Empire.

Balboa was born about 1475 in Jerez de los Caballeros, or in Badajoz, according to some sources, in the province of Extremadura. He came from the ranks of that lower nobility whose sons, "men of good family who were not reared behind the plow," in the words of the chronicler Gonzalo Fernández de Oviedo y Valdés, often sought their fortunes in the Indies. In 1500 he sailed with Rodrigo de Bastidas on a voyage of exploration along the coast of present-day Colombia. Later, he settled in Hispaniola (Haiti), but he did not prosper as a pioneer farmer and had to escape his creditors by embarking as a stowaway on an expedition organized by Martín Fernández de Enciso (1510) to bring aid and reinforcements to a colony founded by Alonso de Ojeda on the coast of Urabá, in modern Colombia. The expedition found the survivors of the colony, led by Francisco Pizarro, but Ojeda had departed. On the advice of Balboa the settlers moved across the Gulf of Urabá to Darién, on the less hostile coast of the Isthmus of Panama, where they founded the town of Santa María de la Antigua, the first stable settlement on the continent, and began to acquire gold by barter or war with the local Indians. The colonists soon deposed Enciso, Ojeda's second in command, and elected a town council; one of its two alcaldes, or magistrates, was Balboa. With the subsequent departure of Enciso for Hispaniola, Balboa became the undisputed head of the colony. In December 1511 King Ferdinand sent orders that named Balboa interim governor and captain general of Darién.

Early years in the New World

Balboa meanwhile had organized a series of gold- and slave-hunting expeditions into the Indian chiefdoms of the area. His Indian policy combined the use of barter, every kind of force, including torture, to extract information, and the tactic of divide and conquer by forming alliances with certain tribes against others. The Indians of Darién, less warlike than their neighbours of Urabá and without poisoned arrows, were not formidable foes and often fled at the approach of the Spaniards. The Spanish arsenal included their terrible war dogs, sometimes used by Balboa as executioners to tear Indian victims to pieces.

The Spaniards were told by Indians that to the south lay a sea and a province infinitely rich in gold—a reference to the Pacific and perhaps to the Inca Empire. The

conquest of that land, their informants declared, would require 1,000 men. Balboa hastened to send emissaries to Spain to request reinforcements; the news they brought created much excitement, and a large expedition was promptly organized. But Balboa was not given command. Charges brought against him by his enemies had turned King Ferdinand II against him, and, as commander of the armada and governor of Darién, the King sent out the elderly, powerful nobleman Pedro de Arias de Ávila (usually called Pedrarias). The expedition, numbering 2,000 persons, left Spain in April 1514.

The discovery of the Pacific

Meanwhile, Balboa, without waiting for reinforcements, had sailed on September 1, 1513, from Santa María for Acla, at the narrowest part of the isthmus. Accompanied by 190 Spaniards and hundreds of Indian carriers, he marched south across the isthmus through dense jungles, rivers, and swamps and ascended the cordillera; on September 25 (or 27), 1513, "standing on a peak in Darién," he sighted the Pacific. Some days later he reached the shore of the Gulf of San Miguel and took possession of the Mar del Sur ("South Sea") and the adjacent lands for the king of Castile. He then recrossed the isthmus, arriving at Santa María in January 1514. His letters and those of a royal agent who had been sent to Darién to prepare the ground for the coming of Pedrarias, announcing the discovery of the "South Sea," restored Balboa to royal favour; he was named *adelantado* (governor) of the Mar del Sur and of the provinces of Panamá and Coiba but remained subject to the authority of Pedrarias, who arrived in Darién, now a crown colony and renamed Castilla del Oro, in June 1514.

Troubled relations with Pedrarias

Relations between the two men were, from the first, troubled by the distrust and jealousy of the ailing, ill-natured Pedrarias toward the younger man. The first bishop of Darién, Juan de Quevedo, sought to act as peacemaker and arranged a temporary reconciliation; in a turnabout Pedrarias by proxy betrothed his daughter María in Spain to Balboa. But the underlying causes of friction remained. The suspicious Pedrarias pursued a tortuous policy designed to frustrate Balboa at every turn; but he at last gave Balboa grudging permission to explore the South Sea. By dint of enormous efforts Balboa had a fleet of ships built and transported in pieces across the mountains to the Pacific shore, where he explored the Gulf of San Miguel (1517–18). Meantime, the stream of charges of misconduct and incapacity levelled against Pedrarias by Balboa and others had finally convinced the crown of Pedrarias' unfitness to govern; news arrived in Darién of his imminent replacement by a new governor who would subject Pedrarias to a *residencia* (judicial review of his conduct in office). Pedrarias doubtless feared that Balboa's presence and testimony would contribute to his total ruin and decided to get rid of his rival. Summoned home on the pretext that Pedrarias wished to discuss matters of common concern, Balboa was seized and charged with rebellion, high treason, and mistreatment of Indians, among other misdeeds. After a farcical trial presided over by Gaspar de Espinosa, Pedrarias' chief justice, Balboa was found guilty, condemned to death, and beheaded with four alleged accomplices in January 1519.

BIBLIOGRAPHY. The standard biography in English, well documented and written, is KATHLEEN ROMOLI, *Balboa of Darién: Discoverer of the Pacific* (1953), which also contains an extensive bibliography of sources. There is a judicious survey of Balboa's career in AMANDO MELON Y RUIZ DE GORDEJUELA, *Los primeros tiempos de la colonización. Cuba y las Antillas. Magallanes y la primera vuelta del mundo* (1952). For a discussion of Balboa's route to the Pacific, see ANGEL RUBIO, *La ruta de Balboa y el descubrimiento del Océano Pacífico* (1965). All modern accounts take as their points of departure the great chronicles of PIETRO MARTIRE D'ANGHIERA, *De orbe novo*, trans. by FRANCIS A. MACNUTT, 2 vol. (1912, reprinted 1970); GONZALO FERNANDEZ DE OVIEDO Y VALDES, *Historia general y natural de las Indias*, ed. by JUAN PEREZ DE TUDELA BUESO, 5 vol. (1959); and BARTOLOME DE LAS CASAS, *Historia de las Indias*, ed. by AGUSTIN MILLARES CARLO and preliminary study by LEWIS HANKE, 2nd ed., 3 vol. (1965). See also C.L.G. ANDERSON, *Life and Letters of Vasco Núñez de Balboa* (1941).

(B.K.)

Balkans, History of the

Since the early 19th century, the name Balkan, a Turkish word meaning mountain, has been applied to the easternmost of the three great southern peninsulas of Europe. The peninsula shades gradually into the European mainland, and it is therefore difficult to assign its exact geographical boundaries; but for the purposes of this article the Balkans are taken to mean the territory of the modern states of Greece, Albania, Yugoslavia, Bulgaria, and Romania. The article traces Balkan history from its earliest settlement in the Old Stone Age to the present. The great Bronze Age civilization of the Aegean and the period of classical Greece, however, are treated in separate articles (see AEGEAN CIVILIZATIONS; GREEK CIVILIZATION, ANCIENT). The article also is intended to cover the national histories of the five modern Balkan states, which, despite their deep ethnic, religious, and political differences, share a geographical unity and a common political heritage of nearly five centuries of Turkish rule.

This article is divided into the following sections:

I. The Balkans to 1815
- Old European civilization
  - The central Balkan region
  - The Adriatic
  - The middle Danube
  - The east Balkan area
- From the Bronze Age to the coming of the Slavs
  - Indo-Europeanization during the Bronze Age
  - Illyrians and Thracians during the 1st millennium BC
  - Roman conquest and barbarian colonization
- The Balkans in the Middle Ages
  - The repopulation of the Balkans
  - Greece from the 7th to the 15th century
  - Bulgaria
  - Serbs and Croats
  - Romania and Albania
- The Balkans under Ottoman rule
  - The Ottoman conquest
  - The character of Ottoman rule
  - The increased role of the European powers
  - The French Revolution and the Napoleonic era

II. The Balkans from 1815 to 1914
- National revolutions (1804–30)
  - The Yugoslavs
  - The Greek revolution
  - The Romanians at the beginning of the 19th century
  - The Bulgarians in the early 19th century
- The development of Balkan states and societies (1830–78)
  - The Yugoslavs
  - Greek politics from 1830 to 1878
  - The birth of Romania
  - The emergence of Bulgaria
  - International politics
- The age of imperialism (1878–1903)
  - The Macedonian question
  - The Albanian national awakening
  - The Yugoslavs
  - Bulgaria
  - Greece
  - Romania
- The Balkans before World War I (1903–14)
  - The crisis of 1908–09
  - The Balkan Wars of 1912–13
  - Results of the wars

III. The Balkans after 1914
- World War I and the peace settlements, 1914–23
  - The role of the Balkan states in the war
  - Results of the peace conferences
- Interwar developments, 1923–39
  - New revolutionary movements
  - Government reactions
  - The Balkans in the 1930s
- World War II, 1939–45
  - Axis victories and occupation
  - The resistance movements
- World War II to the present
  - Postwar settlement, 1945–49
  - Recent developments

I. The Balkans to 1815

The Balkan Peninsula has been inhabited since Lower Paleolithic (Old Stone Age) times (*c.* 200,000–100,000 BC), but how continuously is not yet known. The Mousterian industries of the Neanderthal man from the early

part of the last glaciation (*c.* 70,000–40,000 BC) are well represented, followed by those of the Upper Paleolithic Aurignacian and Gravettian, *c.* 40,000–*c.* 8000 BC, associated with the physical remains of modern man (*Homo sapiens sapiens*). Evidence of habitation layers of postglacial food-gatherers in caves includes the chipped-stone industry developed from the Gravettian, and it continues into the beginning of the food-producing era.

The development of the Neolithic Period around 7000 BC was characterized by increased sedentariness and reliance upon domesticated plants and animals, permitting the existence of larger demographic units. From this period onward, one can speak of the Balkan culture in more detail.

Chronology of Balkan Neolithic and Chalcolithic Complexes*

radiocarbon years BC	(5500)	(5000)	(4500)	(4150)	(3750)	(3250)	(2800)		
years BC	7000	6500	6000	5500	5000	4500	4000	3500	3000
Aegean -central Balkan	proto-Sesklo→	Sesklo→			north Greek late neolithic Dhimini→				
	Starčevo→	Vinča→							
Adriatic				Butmir→			Diana→		
			Matera→				Serra D' Alto→		
	Impresso→	Danilo→					Hvar→		
East Balkan		Karanovo I + II→	III→		IV + V→		VI→		
				(Veselinovo)		(Boian Maritsa)		(Gumelnita)→	
				Hamangia					
years BC	7000	6500	6000	5500	5000	4500	4000	3500	3000

*Based on radiocarbon dates adjusted to true age by dendrochronological technique.

OLD EUROPEAN CIVILIZATION

Together with east central Europe, the Balkans formed the core of the earliest European civilization, Old Europe, dated to the period 7000–3500 BC. Old Europe is not merely a provincial reflection of the great civilization of the ancient Near East but rather a distinct culture with roots of its own and a unique identity. The Old Europeans discovered the use of copper in the 6th millennium BC and formed population aggregates that often amounted to small townships, which inevitably involved craft specialization and the creation of religious and governmental institutions. The Old Europeans also evolved a written script.

Regional groupings

Old European civilization may be divided into five regional groups: (1) the Aegean and central Balkan, (2) the Adriatic, (3) the middle Danube, (4) the east Balkan, and (5) the Moldavian–western Ukrainian. The first region embraces the culture of the Neolithic Period in what is presently Greece, Yugoslavia (except the Adriatic seaboard), southeastern Hungary, and western Romania, succeeded by the Chalcolithic Vinča, Butmir, and Tisza cultures. The second area consists of the painted-pottery complex on either side of the Adriatic, including the Matera-Serra d'Alto in southern Italy and the Danilo in Dalmatia. The third variant includes the Lengyel culture of the middle Danube Basin, which displays similarities to the Adriatic complex. To the eastern bloc belong three cultural groups with individual traditions in art: the Hamangia on the Black Sea coast; the Karanovo in Thrace, and its Chalcolithic successors, and the Boian and Gumelniţa complexes, which occupied most of modern Macedonia, Bulgaria, and Romania. In the fifth region, Moldavia and the western Ukraine, local Neolithic culture was followed by the formation of the Cucuteni (Tripolye in Russian) civilization. On coastal and inland plains, mounds created by the accumulation of cultural debris attest to the permanence of the farming communities, whereas in the Danubian region a wandering agriculture was practiced. Houses were arranged either around a central structure or in parallel rows. The typical house was a rectangular structure of timber posts completed with wattle-and-daub walls, semi-subterranean dwellings, trapezoidal structures; cultures living in dry Mediterranean conditions built characteristic houses with mud-brick walls, stone foundations, and internal buttresses.

The Neolithic complex left a remarkably uniform artifactual expression, including polished-stone axes, adzes, and small ornaments; clay stamp seals; bone spatulas; bird-shaped vases; and clay models of shrines. Clay and marble votive figurines reflect several types of goddesses, notably the Snake Goddess, the Bird Goddess, and the Great Goddess, and their epiphanies in the form of a bird, toad, bee, butterfly, deer, bear, and dog. A male god is also represented. Loom weights, spindle whorls, and woven-mat impressions on pot bases attest the production of textiles. Typical of the fine pottery is a highly burnished, hemispherical bowl, often on a ring base; the vessel is commonly coated with a red slip, with red-on-white or white-on-red-painted decoration. Cemeteries were unkown and interments are found within the settlement. Infants were buried inside egg-shaped vessels.

Differentiation into regional groups progressed steadily. By the mid-6th millennium BC, there appear the first traces of metallurgy, the Neolithic thus yielding to the Chalcolithic, the age of a mixed copper and stone technology.



Prehistoric Balkan sites.

The central Balkan region. *Proto-Sesklo and Starčevo.* The earliest Neolithic occupation of the Aegean and central Balkans is differently named in each of the modern European countries in which it is distributed—Proto-Sesklo, Starčevo, Körös, Criş—but these are regional variants of a widespread, comparatively uniform cultural complex.

Early Neolithic technology

The subsistence bases relied upon emmer, a primitive variety of wheat, and the domesticated sheep or goat. The identical bone structure of the sheep and goat makes it impossible to distinguish which animal is represented by the remains, but, because neither species was indigenous to Europe in the terminal Pleistocene, or Ice Age, the animals in question must have been introduced from their natural habitat in Anatolia and the Near East. Though their initial introduction was probably a function of ethnic movement across the Aegean, subsequent diffusion may have resulted from acculturation of the Mesolithic population under the stimulus of contact and trade with early farming communities. Radioactive-car-

bon dates and typological study show the Thessalian and Macedonian pre-pottery and early ceramic sites to be Europe's earliest Neolithic settlements.

Excavations conducted between 1965 and 1971 on the Danube above the Iron Gate provide new insight into the process of economic transition to a Neolithic way of life. The early levels comprise permanent settlements without ceramics; their subsistence was based on fishing and on hunting with the dog, the only domesticated animal. The subsequent phases belong to the characteristic central Balkan Neolithic (Starčevo), with the domesticated sheep or goat and wheat. Starčevo farmers increasingly exploited domestic cattle and pigs and hunted and fished in response to the better forested environment. The foremost systematically excavated sites in this region are Lepenski Vir excavated by D. Srejovic in 1965–69 and Padina, excavated by B. Joranovic in 1968–71.

Sesklo. Sites containing Sesklo materials have been discovered throughout Greece; most of these were newly founded settlements, implying an increase in population density. The beginning of the Sesklo phase is defined by the appearance of fine white-slipped pottery, decorated with flame and stair designs, painted in dark red. Shapes included globular-necked jars, footed bowls, straight-walled bowls, and cups and jugs.

The site of Otzaki, in Thessaly, has disclosed a three-phase development of Sesklo ceramic art. Excavation has uncovered two close-standing rows of square houses constructed of rectangular mud bricks. These houses frequently rest upon a stone foundation and are occasionally strengthened by internal buttresses.

Vinča. The Vinča sequence is best documented at the eponymous site, situated 14 kilometres (nine miles) east of Belgrade, which has yielded seven metres of stratified Vinča deposit overlying the Starčevo levels excavated by the Yugoslav archaeologist Miloje Vasić, intermittently from 1908 to 1932. Fine Vinča ceramic wares are burnished in orange or black and decorated with a shallow linear channeling.

The origin of the Vinča black burnished ware need be sought no farther afield than the Maritsa Valley of central Bulgaria. At its greatest extent, the Vinča complex occupied the central Balkans from the Rhodope Mountains north as far as the Banat; west to northeast Bosnia; and eastward to western Bulgaria, southwest Romania, and Transylvania. In architecture and subsistence economy, no marked difference has been discerned between Vinča and the Neolithic Period. Wattle-and-daub architecture continued to be employed, with evidence of the use of split planking in floor and wall construction. Floors were clay plastered. Increasing trade is evidenced by the widespread occurrence of Aegean Spondylus shell, which was used for beads and bracelets. The archaeological data indicate gradual social and economic change and an attendant population increase. Vinča sites occupy as much as 20 acres of river terrace, with houses organized into lined streets.

Concomitant with this expansion was a remarkable artistic development of the Vinča culture, marked by an increase in the quality and number of figurines and other objects that served a ritual function. These symbolic figures testify to an important intensification of spiritual life during the period. Vinča art remained distinct from that of neighbouring groups throughout a millennium or more; and its gradual evolution reflects a remarkably stable and well-organized social structure.

Linear writing

The 1961 discovery at Tărtăria in Transylvania of three clay tablets inscribed with pictographs has encouraged the explanation of import from Mesopotamia at about 3000 BC, but this hypothesis has been invalidated by the mutually reinforcing evidence of stratigraphic typologies and radioactive-carbon dating, which locate the early Vinča period in the latter part of the 6th millennium BC. The Tărtăria tablets were found in a secure Vinča context, and inscribed linear signs on figurines, vases, and spindle whorls of definitively local manufacture appear to confirm that the late 6th and 5th millennia witnessed the development of linear writing. Equivalent inscriptions occur in the east Balkan civilization.

Tisza. Tisza settlements, named after the River Tisza, contemporary to Vinča and distributed in eastern Hungary and the Yugoslav Banat, are smaller than Vinča, consisting of rectangular and semi-subterranean dwellings with pitched roofs apparently resting at ground level. Local domestication of cattle on a scale unprecedented in temperate Europe was proved here.

Butmir. The inland Bosnian Butmir variant was strongly influenced by the Vinča culture and was already producing copper. It is best known from the eponymous site at Butmir (Sarajevo), excavated between 1893 and 1896 by W. Radimsky and F. Fiala, which yielded a significant number of sculptures and fine ceramics, including black-polished ones with incised triangle and spiral decoration encrusted with red and white paint. Typical shapes were biconical bowls; high-pedestal, footed "wine" cups; and hole-mouth pyriform vases. At Obre II, a Butmir site in a forested upland environment near the upper Bosna, excavated in 1967–68 by A. Benac and M. Gimbutas, faunal analysis revealed that over half the domestic animals were cattle and that a considerable amount of hunting was practiced. Einkorn wheat was the chief cereal crop, supplemented by barley and lentils; at the Butmir site itself, pear and apple pips have been found.

At Obre II, wattle-and-daub houses of two rooms, about 15 metres long, containing beehive-shaped bread ovens, wooden platforms, ash pits, storage and ritual vases, spindle whorls, loom weights and tools, were revealed. The calibrated radioactive-carbon dates from this site place the three periods of the Butmir culture at 5000–4000 BC.

The Adriatic. *Impresso.* The Early Neolithic culture of the Adriatic region is known as the Impresso complex; it is characterized by grit-tempered wares impressed with shells or with a stabbing tool. The simple ornamented bowls and the farming economy they served are believed to have developed as a result of rapid diffusion, a corollary to maritime movement and trade along the Adriatic littoral. The Impresso culture of western Yugoslavia represents only a part of a complex widely dispersed throughout the Mediterranean world. The initial stimulus to agriculture probably came through eastern seaborne contact with the indigenous Mesolithic peoples of the Mediterranean coastlands. No considerable ethnic migration need be invoked in explanation.

Impresso sites occupy both caves and open settlements. The Dalmatian open-air settlement of Smilčić near Zadar consisted of wattle-and-daub houses and was enclosed by a deep ditch.

Danilo. The beginning of the Danilo period is marked by the appearance of elaborate red-on-cream-painted wares, with geometric motifs akin to Sesklo designs. This was a flourishing Neolithic culture of the end of the 6th and 5th millennia BC, with contacts with Greece and southern Italy.

The middle Danube. The local indigenous population of Cro-Magnon type prevailed in the middle Danube until the 5th millennium BC, when the periodically migrating farmers were supplanted by the Lengyel complex, named after a site in western Hungary that displays different architectural and artistic traditions. The Lengyel physical type contrasts markedly with the central European Cro-Magnon and is related, rather, to the people of the Adriatic region. Lengyel settlements are fortified with wide ditches and palisades, and the typical painted piriform vases and "fruit stands" have analogies in the Danilo complex. These similarities may reflect an ethnic infiltration from the Adriatic to the Sava basin and the region east of the Alps. The villagers cultivated wheat, barley, and Italian millet; and in addition to tending sheep or goats, they hunted and locally domesticated both cattle and pigs.

The east Balkan area. The east Balkan civilization began around 6000 BC with the first appearance of Neolithic occupation along the Maritsa Valley in Bulgaria. The best known sites are Karanovo, Azmak, and Kazanlŭk. The Karanovo sequence, phases I–VI, has become universally adopted as a chronological yardstick

for the development of east Balkan civilization. (Roman numerals are used by archaeologists to designate the successive layers of occupation found at a particular site, in this case phase I representing the earliest known settlement and higher numerals representing later settlements.)

Karanovo I–III. In the villages of the Karanovo I–II period, rectangular one-roomed houses of six or seven metres to a side, with thick, wattle-and-daub walls, an aligned plank floor, and an internal oven, were arranged in parallel rows. The ceramics show affinities with central Balkan wares, and a rich tool kit includes numerous millstones and sickles of deer antler with flint blades inserted like sawteeth. The plentiful remains of einkorn and emmer wheat, lentils, and bones of domesticated animals confirm the role of mixed farming. Karanovo III pottery typically displays a dark, lustrous black surface, sometimes decorated with plastic bands and knobs or linear incision. The most diagnostic forms are cylindrical and pear-shaped vases with single handles. Karanovo III complex is also known as Veselinovo, after another site in eastern central Bulgaria. Karanovo III corresponds to a considerably increased population; during this phase, elements of the Karanovo III assemblage were carried, probably by a movement of peoples, northwest into the lower Danube region and south beyond the Rhodope Mountains to eastern Macedonia and Thrace. In the north, the intruders came in contact with the Hamangia group on the coast of the Black Sea and with the settlers of the central European Linear Pottery culture. This contact along the lower Danube resulted in the amalgamation of the individual ceramic traditions within the assimilating expansion of the Karanovo III population, bringing about the formation of the "Boian" tradition, a northern variant of east Balkan civilization (the name Boian is derived from an island settlement near the Danube south of Bucharest). The "Vădastra culture" in western Romania represents a sister branch of the same civilization.

Boian. The Boian farmers cultivated einkorn, millet, beans, and flax; they also domesticated and kept cattle, pigs, sheep, and goats, but cattle were by far the most important. Houses arranged in parallel rows in compact villages were fortified by ditches. The Boian pottery was ornamented with excised, white, encrusted spiral and meander designs. There are five subbases of Boian.

During its second phase, Boian material culture spread as far as Moldavia in the northeast, influencing the formation and development of the Cucuteni civilization; it was this vigorous phase that witnessed the first signs of metallurgy. A late Boian temple edifice with trichrome wall decorations and with two exquisitely painted pillars inside has been discovered in the island site of Căscioarele, southeast of Bucharest. Skeletons from a large cemetery at Cernica (near Bucharest) are predominantly of a small-statured Mediterranean type.

Gumelniţa. The advanced stage of east Balkan culture is the Gumelniţa civilization, known from at least 500 tells (mounds of cultural debris) in Romania, Bulgaria, and eastern Macedonia. Gumelniţa itself lies southeast of Bucharest. Other important sites are: Căscioarele; Tangîru and Hîrşova, in the lower Danube region of Romania; and Ruse and Hotnica, in northern Bulgaria. North of the Aegean, the most noteworthy sites are the mounds of Sitagroi and Dikili Taih, excavated in 1968–70 on the plain of Drama. Sedentary Gumelniţa communities occupied compact villages or small towns for a millennium or more. Ceramic models of houses show gabled roofs, round windows, and plastered walls with designs painted in red, yellow, and white. These models all attest to the existence of two-story buildings. A model of a temple edifice discovered on the Danube island of Căscioarele shows four shrines standing on a large substructure.

East Balkan culture

Gumelniţa fine ceramic vessels are distinguished by their technological sophistication. Graphite painting required special kilns to prevent oxidation of the graphite.

Metal production and trade exhibited steady growth during the period: needles, awls, fishhooks, and pins were of copper, and, at the end of the period, axes and daggers were manufactured, a development also found in the Vinča, Tisza, Lengyel, and Cucuteni civilizations. Workshops of flint, copper, gold, Spondylus shell, and pottery imply craft specialization and increasing division of labour. Gold was obtained from Transylvania and copper from the Carpathian sources. Evidence of linear writing is found on ceramics throughout the duration of Boian–Gumelniţa civilization.

Hamangia. The Hamangia settlements and cemeteries are found along the coastal strip between northern Bulgaria and the western Ukraine. Most information comes from the cemetery at Cernavodă, where some 350 graves have been excavated by D. Berciu (1957). Examination of the skeletons has revealed that the population was predominantly Mediterranean; but also found was a distinct local type with a short, broad skull.

The Hamangians practiced mixed farming; they cultivated wheat and vetch and herded sheep and goats, cattle and pigs. Deep-sea fishing appears to have been important to early subsistence.

Figurine art characteristically shows a standing female, breasts and buttocks well-developed, with a columnar head and neck, lacking facial features. Less schematic are the exceptional pair of seated figures from Cernavodă—a male "Thinker" and a comfortably relaxed female, both masked. The earliest ceramics were impressed with cardium shell. Ornaments, found abundantly as grave goods, include huge bracelets and beads of Spondylus shell. The Hamangian complex was superseded by the east Balkan Gumelniţa civilization.

FROM THE BRONZE AGE TO THE COMING OF THE SLAVS

Indo-Europeanization during the Bronze Age. Old Europe was developing into an urban culture, but its power was cut short by a steadily increasing infiltration of the semi-nomadic pastoralists from the Russian steppes in about 3500 BC. Their culture is called Kurgan because of their burial in tumuli (*kurgan* means "barrow" in Turkic and Russian), covering graves in deep shafts. It reveals elements of the hypothetical mother culture of the Indo-European speakers as reconstructed with the help of common words. In the period *c.* 3500–2300 BC, their presence is traced in Danubian Europe, and, after 2300 BC, their arrival is documented in the Aegean and Adriatic regions. Changes in social structure, economy, and religion show that people of different background imposed their ways of life. Centres of the new power are witnessed by strongholds such as Vučedol (at Vukovar in northern Yugoslavia) and Nagyárpád (at Pécs) in southern Hungary). Cultural uniformity is evidenced over the Danubian plain down to Macedonia before 3000 BC. The new culture with persisting elements of Old European substratum is known by the names of Cernavoda in Dobruja, Ezero in Bulgaria, Coţofeni in Transylvania, and Baden in the middle Danube region.

Kurgan culture

By 2000 BC, nuclear groups parent to Indo-European Illyrian-, Armenian-, Venetan-, Phrygian-, Mysian-, Dacian-, Thracian-, and Greek-speaking units were formed; their cultures were typified by military aristocracies, hill forts, small villages, horses, and vehicles. Their archaeological complexes are dubbed Otomani and Wietenberg in Transylvania; Monteoru in Moldavia; Tei in the lower Danube region and Thrace; and Incrusted Pottery in Pannonia and northwestern Yugoslavia.

The centres of military power and of bronze metallurgy in the mid-2nd millennium were in central Europe and Greece: the Unětice-Tumulus in the upper and middle Danube region; the Otomani-Wietenberg in the Carpathian Basin; and the Mycenaean culture in Greece (see AEGEAN CIVILIZATIONS). Trading of transcontinental Baltic amber and Transylvanian and Mycenaean gold and bronze is richly evidenced from the 17th through the 15th century BC.

Around 1400 BC the Tumulus people extended their influence over the whole middle Danube Basin and to the Adriatic coast. As Urnfield peoples, they caused an upheaval in the Balkans in Anatolia and in the east Mediterranean region before and around 1200 BC. Their graves

and pots are found in Macedonia, while their weapons are found in Greece, Crete, Cyprus, Syria, Palestine, and Egypt. The destruction of the Mycenaean culture is attributed to the middle Danubian Urnfield invasion through the Balkan Peninsula and the Adriatic Sea. The Urnfield people were Indo-European speakers, parent to Venetan and perhaps Phrygian, Mysian, and Armenian. In past research they were assumed to be Illyrians, but this name can be used only as a broad cover name and should not be confused with the Illyrians proper, known since the time of Herodotus in Dalmatia, Bosnia, and Hercegovina. This period of raids and migrations, instigated by central European Urnfield peoples (misleadingly called an "Aegean migration"), resulted in a Phrygian, Mysian, and Armenian exodus to Anatolia, the destruction of the Hittite Empire, the Dorian infiltration into Greece, and the introduction of iron technology and a new ethnic configuration in the Balkans.

Illyrians and Thracians during the 1st millennium BC. Illyrians proper occupied western Yugoslavia, Albania, and Epirus, with the Morava-Vardar line as their approximate eastern limit; in the northeast they extended into southwest Romania. Tribes north of the Sava River, Venetans and Pannonians, are usually treated as northern Illyrians. The general name for the period from 750 to 450 BC is "Illyrian Hallstatt." The exploitation of iron ores, particularly in northwestern Bosnia, increased commerce. From the 7th to the 5th century BC, strong tribal centres emerged, together with a powerful aristocracy, as witnessed by extravagantly furnished royal burials. Autochthonous traditions were maintained even after an exchange with Greek merchants and colonizers of the Adriatic coast started around 600 BC. Glasinac, an Illyrian centre near Sarajevo, displayed an uninterrupted development until the period of Celtic influence in the 4th century BC.

The Iron Age in Illyria

From the late 6th and early 5th centuries, the Greek influence is evidenced by masterpieces of Greek workmanship, which arrived through southern Italy. To this period belongs the treasure of Trebeništa at Lake Ohrid, which includes a large bronze krater, or mixing bowl. Greek influence increased markedly in the 4th century BC after the foundation of towns on the islands and coast of the Adriatic, among them Issa (Vis), Pharos (Hvar), Corcyra Melaina (Korčula), Salomae (Solin), Epidaurus (Cavtat), Iader (Zadar), and Tragurion (Trogir).

Written evidence of an Illyrian kingdom on the borders of the Greek world appeared at the end of the 5th century BC; its capital was at Skodra (Skoútari). Before Illyria was annexed to the Roman Empire in 168 BC, it was ruled by 15 monarchs carrying Illyrian names.

The Thracian culture formed of the Bronze Age substratum mixed with that of the semi-nomadic peoples from the Ukraine. The ethnic infiltration is shown by the appearance of an eastern archaeological complex called Noua, dated to the 12th century BC. The cultural melange that produced the Basarabi–Babadag complex of the Hallstatt era is held to be the basic component of the Daco-Thracian culture. Repeated incursion of groups of warriors from the northeast occurred at the end of the 8th century BC, bringing with them Caucasian elements in art and weaponry. This period is called Thraco-Cimmerian; it coincides with the conquest of the Cimmerians north of the Black Sea by the Scythians (*q.v.*), but the actual ethnic infiltration of the Cimmerians into the Balkans is not evidenced. The Scythians proper appeared in the middle of the 6th and the beginning of the 5th century BC; but it was not until the beginning of the second half of the 4th century BC that the Scythians tried a violent and massive penetration into the territory. Contacts with the Scythian world contributed to the wider diffusion of iron metallurgy and the formation of a distinctive Thraco-Scythian style of art.

At the end of the 7th and in the 6th century BC, the Greeks founded several colonies along the western shore of the Black Sea: Apollonia Pontica (Sozopol), Mesembria (Nesebŭr), Odessus, Callatis (Mangalia), Tomis (Constanţa), and Istrus (Hystria). The Greek cultural influx was of fundamental importance. It included the potter's wheel, which was in use by the local inhabitants at the end of the 6th century BC, and the striking of coins, which the Greeks themselves borrowed from the Lydians in Anatolia. The Greeks were interested in timber and wheat from Moldavia and the Ukraine and in silver and gold from the mines of southwestern Thrace. Greek influence steadily increased. At Panagyurishte, in central Bulgaria, a gold service weighing more than six kilograms (13 pounds) included vessels of superb workmanship by Greek artists of the late 4th century BC. Dionysian festivals and worship of Bendis (Artemis) and Ares are ascribed to Greek influence.

Greek colonies on the Black Sea

Dacians, the northern Thracians, are known from Greek sources of the 4th and 3rd centuries BC. A good deal about southern Thracian life and manners is revealed by the Greek authors Homer, Herodotus, Thucydides, and Xenophon. Thracian nobles are shown on Greek vases wearing caps of fur, decorated cloaks, and leather boots. The society was classified into chieftains, warriors, commoners, and slaves. In the mid-5th century BC, the first known Thracian state emerged in the Maritsa Valley, ruled over by the Odrysian king Teres. According to Thucydides this state enjoyed a large revenue and general prosperity. The Odrysian king Sitalces allied himself with Athens and in 429 overran Macedonia. The 4th century marks the final flourishing of the Thracian civilization. Before mid-century Philip II of Macedon brought unity to the area between southern Thrace and Albania and a new power—Macedonia—emerged. Philip's son Alexander III the Great crossed the lower Danube in 335 BC and attacked the Getae.

By about 300 BC, the Getae of Dobruja and the lower Danube, a people akin to the Dacians of Transylvania and Moldavia, combined to form what was virtually a state, an evolved type of military democracy having as its leader the *basileus* Dromichaites, who in 292 conquered Lysimachus. Greek cities on the coast of the Black Sea found protectors and allies in the Getian *basileis*.

About the beginning of the 3rd century BC, bands of Celtic warriors appeared east of the Tisza and advanced toward Oltenia and Transylvania. This movement can be traced by the evidence of their warriors' graves. The military superiority of the Celts, which had at its disposal a developed iron technology, insured their success. The Celtic element, which was present in Romania for several centuries, contributed to the development of iron metallurgy and to the adoption of the iron plowshare, which led to a great expansion in agriculture. Celtic influence left imprints in a novel Getian–Dacian silverwork and coinage. The southern wave of Celtic invasion—toward Thrace, Macedonia, Thessaly, and Greece—was broken at Delphi in the winter of 279–278 BC (see CELTS, ANCIENT).

Arrival of the Celts

The Dacian state extended between the Tisza River and the lower Danube in the period between 200 and 31 BC, reaching its peak during the reign of Burebistas (60–44 BC). For further information see EUROPE, ANCIENT.

Roman conquest and barbarian colonization. After Macedonia crumbled, Thracians, Illyrians, and Dacians were left to their own devices until a major Roman irruption at the end of the 3rd century BC. Thracians and Illyrians combined in self-defense, and there followed a struggle for 150 years. The southern Illyrians were conquered and annexed in 168 BC, but hostilities continued until AD 9, when all Illyrians were entirely subjugated by the Roman general Tiberius (later emperor). The territory along the Adriatic, Dalmatia, Iapodes, and Liburnia, was united as the Roman province of Illyricum. In 29 BC the area between the Danube and the Balkan Mountains was conquered by Crassus and became the Roman province of Moesia, while the area further south was incorporated into the province of Thrace. The emperor Domitian's expedition against the Dacians in AD 85 consolidated his position north of the Danube. In 89 the king of Dacia, Decebalus, was forced into a collaboration with Rome against the Germanic Marcomanni, the Quadi, and the Iranian Sarmatians, the latter coming from the Pontic steppe. The complete conquest of Dacia in 106 followed two vigorous campaigns by the emperor Trajan in 101–

103 and 105–106. The Dacian capital of Sarmizegethusa was captured and the province of Dacia was established.

The process of Romanization was swift. In Romania, Latin completely supplanted the Dacian and Thracian languages. Trade and handicrafts penetrated deep into the provinces. Trade routes connected the growing cities: the northern road led to Sirmium (Sremska Mitrovica on the Sava), Singidunum, Viminacium, and Ratiaria on the Danube; the southern led to Stobi (south of Tiov Veles) and Scupi (now Skopje); and the eastern ran from Sardica (Sofia) via Naissus (Niš) and Margum to Viminacium on the Danube; in the west, the road crossed from Salonae and Narona on the Adriatic through Bosnia to Sirmium. Commercial expansion reached its culmination in the course of the 2nd century. Rome became decentralized and its art cosmopolitan. At the end of the 3rd century, Sirmium became a capital of the empire. Diocletian built his palace near Salonae (Split), and official art was created in Dalmatia, Pannonia, and Moesia. From the time of Marcus Aurelius, the empire lived under the protection of an Illyrian–Thracian army. When Constantine I established Constantinople as a second Rome, in AD 330, the eastern part of the empire split in two: Thrace was included in the prefecture of the Orient, whereas the rest of the peninsula was in the prefecture of Illyrium.

In the 3rd century there began an infiltration of the Goths into the peninsula. About AD 214 the Goths clashed with the Romans at the Dacian frontier and conquered Dacia. The Goths then attacked Thrace continually through the middle and late 3rd century. The Gothic state flourished for nearly 200 years until the Huns invaded and swept across the country.

The coming of the Slavs

The devastation of the empire by the Huns, followed by that of the Bulgars at the end of the 5th century and that of the Avars in late 6th, prepared the ground for a wide Slavic dissemination from their homeland north of the Carpathians in the Ukraine. (M.G.)

THE BALKANS IN THE MIDDLE AGES

The repopulation of the Balkans. With the settlement of the Germanic Ostrogoths in Italy and the disintegration of the Western Roman Empire in the final decades of the 5th century, the Balkans again came under the direct control of the Eastern Empire at Constantinople, at least in theory. But the barbarian invasions of that century had left the Balkan Peninsula a sparsely populated wasteland destined to attract new and more permanent settlers, the Slavs. An Indo-European people, the Slavs had settled in the valley of the upper Vistula in central Poland during the great prehistoric migration that brought most of the ancestors of the modern Europeans from Asia. Basically a sedentary people with a highly developed tradition of agriculture and animal husbandry as well as of hunting and fishing, the Slavs were easy prey to the highly mobile and often strongly united nomadic hordes that swept across eastern Europe. These barbarian tribes easily superimposed their rule over the native Slavic political system of democratically run clans, owing only loose allegiance to a chief, or *župan*. It was as vassals of such nomadic tribes that the Slavs spread into south Russia and Pannonia, a former Roman province that occupied parts of modern Austria, Hungary, and northern Yugoslavia. From these two areas beyond imperial control, the Slavs began their steady movement into the Balkans.

Slav invasion of the Balkans

By AD 517 groups of Slavs were crossing the lower Danube to raid in Thrace, Macedonia, Thessaly, and even Epirus. Almost simultaneously, others were moving into the northwest corner of the Balkans toward the Dalmatian coast. The massive military endeavours of the Byzantine emperor Justinian I elsewhere in the empire left no troops free to man the fortresses protecting the Danube River boundary of his realm. Imperial officials in Constantinople attempted to neutralize the northeastern arm of the Slavic pincers encircling the Balkans by offering the Slavs on the lower Danube the status of *foederati*, paid border guards for the empire. The Slavs, however, became vassals of the Avars, a new Turkic horde of nomads who appeared on the south Russian steppe.

The Avars, akin to the Huns, attempted to follow the route of their predecessors. Driving the remaining Germanic tribes from the Pannonian plain, they established there a new state that loosely controlled the Slavs settled in modern Czechoslovakia, the Balkans, and even north of the lower Danube in south Russia. Both the Avars and their Slavic tributaries undertook plundering expeditions throughout the Balkans. Eventually, the invaders took effective control of many areas as far south as the tip of the Peloponnesus and as far east as the great wall of Thrace, the first line of defense of Constantinople. Usually, the ravaging of territory was followed by permanent Slavic settlements being made in the devastated villages. It was only after 591 that the Byzantine army, finally withdrawn from other theatres of war, attacked the Avar–Slavic armies in the Balkan Peninsula; but by then the situation was hopeless. The Byzantine troops, committed to driving the Avars back beyond the Danube, revolted in 602, and soon the emperor was paying tribute to the Avars while the latter's Slavic vassals moved unimpeded through the Balkan Peninsula, populating it anew. Remnants of the local agricultural population probably remained among the Slavic settlers, but many of the Greek-speaking and Latin-speaking inhabitants took refuge in and around fortified Byzantine cities, particularly along the coasts. The safety of these cities, however, was in no way guaranteed. Thessalonica, the leading city of Macedonia and the second-richest city in the Byzantine Empire, several times came close to falling to the Slavs. In the interior, no imperial islands remained. Singidunum (Belgrade), Naissus (now Niš in Yugoslavia), Sardica (Sofia), Sirmium (now Srem in Yugoslavia)—the mighty fortresses guarding the middle Danube frontier—fell, with the surrounding countryside, to the Avars and Slavs. Contemporary sources often speak of the Balkans simply as Sclavinia (Slavdom). Byzantine control of this region had been so completely disrupted that the emperor Heraclius (*q.v.*) omitted the Balkan provinces from the administrative reorganization of the rest of his realm. Rather, to preserve his capital, Constantinople, Heraclius paid massive tribute to the Avar khan (chief) before he dared move against the equally serious Persian threat on the empire's eastern border.

In 626, however, the strength of the fabled walls of Constantinople and Byzantine naval superiority gave the Byzantines a wholly unexpected but decisive victory against a joint attack by the Avars and Persians. From this point on, both the Persian and Avar states began their rapid decline. Already weakened by a revolt of their Slavic subjects in what is now Czechoslovakia, the Avars were further threatened when Emperor Heraclius invited two strong tribes from beyond the Carpathian Mountains, the Croats and the Serbs, to settle in land held by the Avars in the northwest Balkans. Aided by the Byzantine navy and the few fortified imperial cities left on the Dalmatian coast, these tribes took control first of Dalmatia and then of the territory that, after their occupation, bore their names, the Croatia and Serbia of modern Yugoslavia. Once settled, theoretically as vassals of Byzantium, they mingled with the earlier Slavic settlers.

Return of Byzantine power

Greece from the 7th to the 15th century. The period from the mid-7th century on saw the slow reintegration of the southern Balkans (*i.e.*, Thrace, southern Macedonia, and Greece) not only into the Byzantine administrative and cultural system but also into the Greek-speaking world. The most important factor in reintegrating this territory into the empire was the gradual radiation of Byzantine culture from the Greek coastal enclaves into the surrounding countryside and the subsequent conversion of the new settlers to Greek Christianity. With the institutions of the Byzantine Church came the officials of the Byzantine state. The presence of Greek-speaking natives in the villages taken over by the Slavs served to facilitate the acculturation of the newcomers in many places. Military campaigns against recalcitrant Slavs and occasional resettlement of populations drawn from Asia Minor aided the process in other areas. The progress of Byzantine reabsorption of the southern Balkans is most readily seen in the establishment of "themes"—regular Byzantine administrative units—on the territory of Scla-

vinia. By the end of the 7th century, themes existed in Thrace and in the southeast region of mainland Greece; by the end of the 8th century, themes had been established in the Peloponnesus and Macedonia, and the next century saw the completion of a network of imperial themes covering the southern Balkans (Cephalonia for the Ionian Islands, Dyrrhachium [now Durrës] for Albania and northern Epirus, and Nicopolis for southern Epirus). The history of what is today the territory of Greece becomes a regional history of the Byzantine Empire marked by two peculiarities: occasional uprisings of isolated Slavic tribes in mountainous areas and vulnerability to attack from Bulgaria and Serbia as well as from western Europe.

The northern areas of Greece were often subjected to pillaging and territorial losses at the hands of the Bulgarians and, later, the Serbs; of a more lasting character, however, was the danger to Greece from the West. The Norman rulers of Sicily coveted Greece as a stepping-stone to the Byzantine capital, and, three times in the century and a quarter before the Fourth Crusade, which began in 1202, they launched successful campaigns in Greece. The Norman invasions, however, were but a foretaste of the fate awaiting Greece after the Latin conquest of Constantinople in 1204. The French and Italian leaders of the Fourth Crusade had little difficulty establishing control over most of the territory forming modern Greece. Boniface of Montferrat was crowned king of Thessalonica, and he, in turn, set up his own vassal states, the most important being the duchy of Athens, which included Attica and Boetia, and the principality of Achaea, which encompassed the Peloponnesus. Venice was awarded most of the Greek islands and many of the major ports. Yet the elaborate feudal superstructure erected by the crusaders effected little social change. Generally, Greek landlords were retained in their holdings after swearing fealty to the new Latin suzerain, and the masses continued to pay their taxes and rents as before.

Advent of Western rule

Simultaneously with the advent of Latin rule in eastern Greece and the Peloponnesus, an exile Greek government took control of western Greece from the neighbourhood of Dyrrhachium south to the Gulf of Corinth and established the state of Epirus. Taking advantage of the weakness of the neighbouring Latin states, the Greek government of Epirus quickly spread its power east to Thessaly, northeast to much of Macedonia, and, in 1224, to the city of Thessalonica. The hold of the Epirote government on its new possessions was tenuous, however, and much territory was soon lost both to the Bulgarians and to the exiled Byzantine emperor at Nicaea in Asia Minor, who took Thessalonica. By 1264 the Epirote rulers of Epirus and Thessaly had been forced into an uncomfortable vassal relationship with the restored Byzantine Empire and later were completely incorporated. Eighty years later, there were no longer any independent Greek states in the Balkans, and even the Albanian tribes that had been spreading across Greece for the previous half century recognized the rule of Constantinople. Byzantine territory reached to the borders of the duchy of Athens, which had fallen to a roving band of Catalan mercenaries in 1311. Farther south, the Byzantines, who had retaken Constantinople in 1261, used their foothold on the southern tip of the Peloponnesus to spread their influence outward to the Greek population dissatisfied with the administration of their Latin rulers. By 1340 Byzantium controlled all of the Peloponnesus except for some Venetian-held ports, and the area became the "Despotate of Morea," ruled by viceroys appointed from Constantinople. Within a generation, Morea had also forced the Latin duchy of Athens, together with most of Attica, into submission.

The period of the 1340s saw the beginnings of far-reaching changes in the Balkans. The Byzantine Empire was plunged into civil war and social strife, complicated by a Serbian invasion that occupied most of Macedonia and Greece as far south as the Gulf of Corinth. During the 1340s the Ottoman Turks also established their first permanent settlement in Europe. Yet, although in the next 100 years Greece was absorbed into the Turkish Empire, it was a century of almost frantic cultural activity in Greece. The capital of the Morea at Mistra became a vital centre of Greek studies and Renaissance Humanism, in spite of almost constant Turkish raiding. But eventually, one by one, the Greek cities of Europe fell to the Ottomans (Thessalonica in 1430, Constantinople itself in 1453, Athens in 1456). In 1460, when the Morea fell, all of Greece had become part of the new Turkish Empire.

From *Grosser Historischer Weltatlas*, vol. 2, *Zweiter* (1970); Bayerischer Schulbuch-Verlag, Munich



The Bulgarian Empire under the Asenids about 1241.

Bulgaria. As Avar power was waning in the 7th century, a new threat to the peace of the Balkans appeared —the Bulgars, a Turkish tribe that, after casting off Avar suzerainty, began to cross the Danube. After successfully repulsing a combined land and sea attack by Byzantium in 680, the Bulgars, under their khan Asparukh (Isperich), took control of the area between the Dniester River and the Balkan mountains and quickly subjugated the Slavic inhabitants. This new Balkan state, comprised of an aristocracy of Turkic–Bulgar nobles ruling over a Slavic population, was soon strong enough to demand the payment of tribute by the Byzantine Empire; Bulgaria was to remain thereafter a permanent factor in Balkan history.

After the final defeat of the Avars by Charlemagne in 796, the Bulgarian state spread even farther, taking the Byzantine fortress of Sardica (later Sofia) and besieging the key Byzantine city of Adrianople (modern Edirne, Turkey). The Byzantines retaliated in force, but a Bulgarian ambush took the life of the Byzantine emperor Nicephorus in counter-retaliation. Two years later, the Bulgarian khan Krum was ravaging the suburbs of Constantinople. By 864, however, the Byzantine Empire had recouped enough of its armed strength to force the new ruler, Boris I of Bulgaria (*q.v.*), to accept Baptism from Byzantine missionaries and to establish a branch of the Byzantine church as the state religion of Bulgaria. The gates were opened to a flood of Byzantine culture. When the old Turkic aristocracy rose up against this wave of Byzantinization, the khan Boris, now christened Michael, backed by the Slavic population, crushed the opposition. The entire country quickly became Slavicized, and Bulgaria permanently joined the Byzantine cultural sphere.

Arrival of the Bulgars

The reign of Boris-Michael's son and successor, Symeon (893–927), witnessed the high point of medieval Bulgarian history. The seemingly invincible Bulgarian army conquered most of the Balkans, until the Bulgarian state stretched from the Black Sea almost to the Adriatic and from the middle Danube into northern Greece. The Byzantine Empire in Europe was reduced to Thrace, southern Macedonia, Greece, the islands, and a narrow stretch of territory along the Adriatic coast; and even

these areas were subjected to Bulgarian raids. In the year 913 the Bulgarian army invested Constantinople itself and forced the Byzantine government to crown Symeon as tsar. Symeon dreamed of becoming the unique emperor in Byzantium, but the enthronement of a brilliant general in Constantinople in 920 frustrated his dream.

Tsar Symeon's Byzantine orientation is particularly apparent in the internal reforms he introduced into the structure of Bulgaria. The oligarchic state of Symeon's ancestral Turks became the autocratic centralized state of a neo-Byzantine tsar. The Bulgarian government imitated the Byzantine administrative system down to the very titles of officials and their ceremonial dress and soon even forced the free peasantry into the traditional forms of Byzantine rural servitude on the lands of the magnate boyars. As guide and protector of the church in his realm, Symeon introduced important changes. He encouraged the work of the Slavic-speaking clergy and began to make the Slavic language of his subjects the liturgical language of the church in Bulgaria, which allowed him to replace Greek clergy with his own subjects. Moreover, he made the Bulgarian Church an independent jurisdiction with its own patriarch, as befitted the church of a state ruled by a tsar. The Bulgarian people now had a religious as well as a political symbol of national unity. The new Bulgarian capital at Preslav became a veritable centre of Byzantine culture patronized by the Tsar. Monumental churches and palaces on the Byzantine model made the city a source of wonder to visitors and a young rival to Constantinople itself. The ruler supported translations of Byzantine liturgical, literary, and legal works, as well as the writing of new works in Slavic. From Preslav the Tsar disseminated a syncretic Slavo-Byzantine culture across the Balkans. Large numbers of Bulgarian subjects, however, rejected the state, the church, and the culture that Symeon had fostered and united in a puritanical sect called the Bogomils, which rejected all that was material as evil. This protest movement, which spread into Serbia and particularly Bosnia, significantly weakened the internal cohesion of the Bulgarian state and, coupled with the weakness of Tsar Symeon's successors, left Bulgaria unable to resist a Byzantine attack in 971. Emperor John I (Tzimisces) annexed Bulgaria, abolished the Bulgarian church patriarchate, and carried the Tsar back to Constantinople as a prisoner.

Barely five years after Tzimisces' victory, however, a revolt broke out in Macedonia that eventually recreated a Bulgaria almost as large as that ruled by Symeon. The revolt was led by the four sons of a provincial governor, the youngest of whom, Samuel, became tsar, locating his capital at Ohrid, where he also established the resurrected patriarchate of the national church. From Ohrid, Tsar Samuel was able to conquer the Serbian principalities, Thessaly, Epirus, and most of Albania, as well as all of Macedonia except the imperial stronghold of Thessalonica. In 1014, however, the long war waged against Bulgaria by the Byzantine emperor Basil II (see BASIL II BULGAROCTONUS), the "Bulgar Slayer," bore fruit. Basil captured and systematically blinded almost 14,000 Bulgarian soldiers. The First Bulgarian Empire quickly disintegrated.

For a brief moment, the Balkans once again belonged to the Byzantine Empire. The core of Samuel's state was divided into themes. The parts of Samuel's domain that lay in modern Yugoslavia were returned to their local princes, who ruled once more as Byzantine vassals. The concessions made by Emperor Basil to local customs in the newly conquered territory, however, were gradually abrogated by subsequent emperors. Finally, the discontent engendered by the growing harshness of Byzantine rule, the chaos of the period of the Crusades, and the disappearance of a powerful centralized administration in Constantinople caused a minor uprising that led to the foundation of the Second Bulgarian Empire.

The leaders of the uprising of 1185, two brothers named Peter and Asen, landed magnates of the Byzantine theme of Bulgaria, were of Vlach origin—that is, descendants of the old inhabitants of Roman Dacia. Their revolt quickly spread across the eastern Balkans, until the Byzantines were forced to recognize the rebels as an independent Bulgarian empire ruled by "Asen, tsar of the Bulgars" and to allow the appointment of an independent archbishop for the new Bulgarian capital at Tŭrnovo as a mark of the state's imperial status. The Bulgarian leaders attempted to re-create the imperial aura of Symeon's days by making Tŭrnovo a new cultural centre for the Balkan's, with magnificent churches and monasteries and flourishing literary groups in close touch with cultural developments in Constantinople. Neither military campaigns nor the fostering of internal opposition to the Asenids could restore Byzantine power in Bulgaria nor even arrest the new state's westward expansion at Byzantine expense. By the time the crusaders captured Constantinople in 1204, Bulgaria was the strongest power in southeast Europe and was still expanding into Thessaly, Macedonia, and Albania.

Second Bulgarian Empire

The death knell for the Second Bulgarian Empire as the decisive element in Balkan politics came most unexpectedly, following the death, in 1241, of one of its most successful rulers, Tsar Ivan Asen II. In the midst of the ensuing succession crisis, a Mongol army slashed its way through the country, wrecking its economy and forcing the payment of massive tribute to the Mongol khans for the next 59 years. Bulgaria was broken as a major power. Serbia and the various Greek states gradually reduced the extent of its territory, while internal dissension produced a chaotic succession of tsars of various political persuasions, including briefly a Turkic Cuman, a reminder of the Bulgarian penchant for calling in nomads from beyond the Danube to strengthen their army. In the end, the combination of external threats and internal disequilibrium opened the way for Turkish conquest. In the years 1393–96 Bulgaria became the first European state to be absorbed into the Ottoman Empire.

From C. Previte-Orton and Z. Brooke (eds.), *The Cambridge Medieval History;* Cambridge University Press



The Serbian kingdom of Stefan Dušan, 1340.

Serbs and Croats. After the settlement of the Serbs and Croats in the territory of present-day Yugoslavia in the 7th century, missionaries from the Western Church were called in to restore the lands of Illyricum to Christianity. After the missionaries, at least in Dalmatia and

Croatia, came the margraves of the Frankish Empire. When Charlemagne's knights moved to assert their claim to the western Balkans, the Western cultural orientation that was to distinguish Croatia and Dalmatia from the Byzantine-oriented bulk of the Balkan Peninsula was established. The Frankish domination of Dalmatia and Croatia, however, was by no means uncontested. The Byzantine navy guarded the Adriatic littoral, and the growing Bulgarian Empire's sway extended even west of Belgrade. The Pannonian Croats revolted, albeit unsuccessfully, while the Dalmatian Croats, allied with the Byzantine coastal cities, sought a counterpoise to Frankish domination in the Emperor at Constantinople.

In 879, however, Zdeslav, the last pro-Byzantine ruler of the Croats, was overthrown by the Frankish vassal Branimir. But soon it appeared that the Holy Roman emperors of the Franks were no more able to guarantee the safety of Dalmatia than the distant Byzantine emperors, and in 925 Tomislav, duke of the Dalmatian Croats, turned to Rome for a royal crown, drew the Frankish vassal Croats of Pannonia into his kingdom, and successfully organized local defenses against Symeon's Bulgarian Empire. The resulting kingdom of Croatia prospered through the 11th century.

Kingdom of Hungary and Croatia

But, when Zvonimir, the last native Croatian king, was assassinated in 1089, the nobles invited Zvonimir's brother-in-law, King Ladislas I of Hungary, to take the Croatian throne; Ladislas' brother and successor as king of Hungary, Kálmán (Koloman), made this arrangement permanent by the Pacta Coventa of 1102. Henceforth, the king of Hungary was also king of Croatia and was represented there by a ban, or viceroy. The coastal cities in Croatian hands retained their special privileges, however, and local Croatian traditions remained intact. The countryside continued to be ruled by the local *župani*, or chiefs, probably descendants of the original Croats called in during the 7th century. The leading *župani* continued to form a royal council, and national assemblies were called to settle important matters. In some ways, then, the basic traditions of the Croatian people were preserved, even though their political state was henceforth drawn into the affairs of central eastern Europe rather than of the Balkans.

The history of the Serbs has proved very different from that of the Croats, largely because they settled farther south and east when they came to the Balkans in the 7th century. The Serbs seem to have been absorbed rather completely by the Slavic population of what was to become Serbia. The political organization of the area, however, remained unclear until the mid-9th century, when the grand *župan* Vlastimir united several Serbian tribes against a threatened Bulgarian invasion. Vlastimir also sought Byzantine support, and it was probably in this connection that the Byzantines initiated large-scale missionary work among the Serbs, drawing Serbia into the Byzantine cultural sphere almost simultaneously with the spread of Latin Christianity among the Croats.

In the interior of the Serb lands, Byzantine support seems to have been restricted largely to Christianization, for most of this area was quickly absorbed into the Bulgarian Empire of Symeon early in the 10th century. Once established, however, Bulgarian rule seems not to have been particularly oppressive. Local *župani* continued to rule, though as vassals of the Bulgarian tsar rather than of the Byzantine emperor. Similarly, the fledgling Christian church in Serbia was subordinated to the Bulgarian patriarchate. When, after the death of Tsar Symeon, the Bulgarian Empire disintegrated, an interlude of Serbian unity under Byzantine protection followed; soon, however, it gave way to petty feuding ended only by the conquering forces of Tsar Samuel's new Bulgarian state. The Byzantine defeat of Samuel in 1014 brought the Serbian lands once again under the sway of Byzantium as exercised through local princes. Around the year 1036, however, Stephen Vojislav, the prince of Zeta (modern Montenegro), renounced Byzantine suzerainty and began to absorb neighbouring principalities. By 1077 Zeta had grown important enough to merit a royal crown; under King Constantine Bodin it succeeded in absorbing most of the Serbian principalities. At Bodin's death around 1101, however, the Serbian state fell once more into civil war, and the centre of political gravity shifted northeast to Rascia (Raška). Sometime after 1165, Stephen Nemanja took the throne as grand *župan* of Rascia and founded a dynasty that was to rule more than two centuries.

The Serbian Empire

The signal for the birth of the Serbian Empire of the Nemanjids was the death of the Byzantine emperor Manuel I Comnenus in 1180, after which Byzantium was never again in a position to reassert its rule over Serbia. Before his retirement to a monastery in 1196, Stephen Nemanja had absorbed the kingdom of Zeta as well as considerable territory along the Adriatic and had obtained Byzantine recognition as the ruler of an autonomous state. Nemanja's son and successor, the able Stephen II (1196–1228), did much to solidify the unity of the new Serbian state. He obtained a royal crown from the papacy, and, from the Byzantine emperor in exile at Nicaea, he received an independent archbishop for the Serbian Church in the person of his younger brother, St. Sava, destined to become the patron of the national church of Serbia. The following 100 years of Serbian history were highly complex: the elected *župani* of the clans and tribes became gradually transformed into hereditary castes of lords and nobles, and the drift of free peasants toward serfdom began, a movement that was to accelerate in the following century. In addition, the period saw a tangle of short-term alliances and of wars both with foreign powers and within the ruling Nemanjid family. Through all this, however, the Serbian state continued its steady accumulation of territory, a movement capped by Stefan Dušan.

Stefan Dušan

Magnates bent on acquiring new land for their estates were responsible for placing Dušan on the Serbian throne in 1331, and they were in no way disappointed. A brilliant statesman and military leader, the young king moved quickly to take advantage of civil war in Byzantium. By diplomacy and war he moved his border south, until all of Macedonia except the area around Thessalonica was subject to him, as was most of Albania. In 1346 he was solemnly crowned tsar by the newly created patriarch of the Serbian Church, assisted by the Bulgarian patriarch, for Bulgaria had now become little more than a dependency of the new Serbian Empire. Soon Epirus and Thessaly also fell to the Serbs. Stefan Dušan was also a brilliant administrator of his realm, fostering trade, promulgating a comprehensive legal code, establishing an efficient civil service on the Byzantine model, and, in general, giving his subjects peace and prosperity. Like his Nemanjid predecessors, Dušan endowed the country with monasteries and churches and patronized the arts. The writers of his time continued the tradition of literary productivity inaugurated by St. Sava. At his death Dušan could describe himself as "emperor of the Serbs and Greeks, Bulgars and Albanians," for he had doubled the territory of Serbia and had taken half of the territory remaining to the Byzantine Empire. His rule at Skopje was recognized from the Danube to the Gulf of Corinth and from the Adriatic to the Aegean. There remained for him but one unattained goal, the conquest of Constantinople, which he was preparing at the time of his death.

Within a few years of his death, Serbia, the power and unity of which largely depended on the personality of the monarch, was again an amalgam of petty kingdoms, principalities, and despotates. The enemy was no longer a weakened Byzantine Empire, however, but a unified and growing Ottoman state. The Serbian kingdom of Bosnia was a potential source of hope, but it was compromised as a possible leader of the Balkans by its heretical, radically ascetic Bogomil Church and its proximity to Hungary. Macedonia had fallen into Turkish vassalage by 1371, and, when, on June 15, 1389, the Ottoman sultan Murad I smashed the Christian forces led by the Serbian prince Lazar at the Battle of Kosovo in Serbia, the end of the Balkan Christian states had come. One by one, the Slavic leaders of the Balkans accepted Turkish suzerainty, which, in time, gave way to direct Turkish rule. By 1463 Bosnia, the last major Serbian stronghold, had been taken by the Turks, and there remained free only the tiny

Serbian enclave of Montenegro hidden in the almost impenetrable mountains along the Adriatic coast, a perpetual reminder of Serbia's past independence.

Romania and Albania. The last two ethnic groups to attain statehood in the medieval Balkans represent, curiously enough, two of the oldest ethnic lines inhabiting the peninsula. The Romanians, or Vlachs, as they were usually called in the Middle Ages, represent the descendants of the Latinized Thracian population of the old Roman province of Dacia. Originally inhabitants of the Danubian plain, many of them seem to have migrated into the mountainous region of Transylvania after the withdrawal of the Roman army from Dacia in 271. There they lived the life of nomadic sheepherders, comparatively safe from the barbarian migrations that followed, until the late 12th century, when large numbers of them seem to have moved into the plains of Macedonia, Thrace, and Dobrudja—possibly as a result of tightened Hungarian control over their Transylvanian refuge. There they played a major part in the Second Bulgarian Empire, the leader of which, Ivan Asen I, was often called the tsar of the Bulgars and the Vlachs.

The appearance of the first specifically Romanian state, Walachia (Land of the Vlachs), can be dated to the early 14th century, when large numbers of Vlachs, who had moved from the Carpathian Mountains to the great plain of the lower Danube, united under Prince Besaraba with some of the Vlach population of the Second Bulgarian Empire. In 1349 Moldavia, a second Romanian principality, was founded east of the Carpathians in the Pruth Valley. Its *vojvoda* (military governor) was Bogdan, a former official of the Hungarian government. Like its sister state Walachia, Moldavia was a product of Vlach emigration from Transylvania, where the Hungarian government was encouraging Roman Catholic missionary work among the Eastern Orthodox Romanian population. The Byzantine government eventually recognized the existence of the Romanian principalities by appointing bishops for Walachia and Moldavia. The states adopted forms of government imitative of the Byzantine except for the fact that the *vojvode* were elected from the ruling families by the boyar aristocracy of each state.

Walachia and Moldavia

It was the misfortune of the Romanians to move toward statehood at a most unpropitious time. Pressed from the rear by their Hungarian former rulers, they also had to meet the growing Turkish power head on. Occasional superb leadership slowed the Turkish advance into the Romanian principalities and ameliorated its results. Finally, in 1393, however, the Walachian *vojvoda* Mircea the Old was forced into vassalage by the Turkish army. Alliances, participation in Hungarian-inspired crusades against the Turks, and cooperation with the sister principality of Moldavia were all vitiated by the military might of the Turks as well as by dynastic civil war. Walachia was rarely more than temporarily free from the burden of tribute to the sultan. With Walachia as a buffer between it and the Turks, Moldavia maintained its independence somewhat longer, although it had to contend not only with Hungary as a neighbour but also with Poland. But in 1455 Moldavia, too, began paying tribute to the Turks. Like his Walachian prototype Mircea, Stephen the Great of Moldavia became a national hero by temporarily breaking with his Ottoman suzerain. In his case, too, the victories were Pyrrhic. Walachia and Moldavia were to remain Turkish dependencies until the age of 19th-century nationalism.

The history of the Albanians in some ways parallels that of the Romanians. As the Romanians were a remnant of the native Thracian people who retreated into the mountains during the earlier barbarian invasions, the Albanians in the western Balkans were remnants of the old Illyrian population who took refuge in the mountain fastnesses along the Adriatic during the great Slavic migration into the Balkans in the 6th century. There, again, like the Romanians, they lived largely as shepherds, having little to do with the larger world of politics for several centuries. Although they were theoretically ruled by the Byzantine governors of Dyrrhachium, Byzantine interference in their mountain territory was minimal. From the 10th to the 14th century, however, almost every state with an interest in the Balkans contended for control of Albania and its strategic coastline. Bulgaria, Serbia, the various rulers of south Italy, Venice, the Greek state of Epirus, and, of course, Byzantium were all involved in the struggle. The contest for domination over Albania had a threefold effect on the population: many Albanians migrated east and south, especially to Thessaly and the Peloponnesus; a distinction developed between the Albanians of the north, who were predominantly Roman Catholic, and those in the south, who were primarily Eastern Orthodox; many of the Greek landowners fled, leaving local government in the hands of minor native chieftains. With the gradual withdrawal of the foreign powers, the country fell into such anarchy and feuding that the land was easily taken by the Turks in 1385. The foremost Albanian leaders were forced to accept Ottoman suzerainty, to pay tribute to the conquerors, and to send their sons as hostages to the sultans.

Contest for Albania

The Albanian national hero, George Castriota, better known under his Turkish name, Skanderbeg, began his career as an Albanian hostage at the Ottoman court. There he was converted to Islām and rose to be an important general, only to desert in 1443, revert to Christianity, and return to his native land to lead a successful revolt against Ottoman rule. His charismatic personality attracted a large following, and his political acumen won him the support of the Kingdom of Naples, the papacy, and even the pragmatic republic of Venice. But, most important, the military ability he had developed in the Ottoman army held back the Turks for 24 years, making him a hero not only in Albania but, indeed, in much of Christendom. His death in 1468 opened the way for Turkish reconquest and large-scale Islāmization of what had previously been a loosely controlled Ottoman dependency, an ironic conclusion to a heroic resistance (see further BYZANTINE EMPIRE). (G.P.M.)

Skanderbeg's revolt

THE BALKANS UNDER OTTOMAN RULE

The Ottoman conquest. The Ottoman Empire had its origins in a small Turkish emirate established in the second half of the 13th century in northwest Anatolia, near the Sea of Marmara and close to the borders of the Byzantine Empire. The name Ottoman was derived from Osman I, a Turkish chieftain who ruled the emirate from 1281 to 1324. An active and aggressive military organization of dedicated frontier warriors, this state first established itself on the European mainland in 1354 on the Gallipoli Peninsula. In 1362 the Ottoman armies took Adrianople, which soon became the Ottoman capital. From this centre the Ottoman armies moved up the Maritsa and Vardar valleys against the Christian states of the Balkans.

The Ottoman successes were made possible by superior military power based on a mobile light cavalry and by excellent leadership and organization. The Christian powers were hampered by their failure either to cooperate among themselves or to secure effective European assistance. Pope Boniface IX's army, sent in an attempt to aid the Hungarians, was crushed by the Ottoman army at Nicopolis in 1396. A second effort in the next century by Pope Eugenius IV collapsed with the defeat of the Christian armies led by János Hunyadi and King Wladyslaw III Warneńezuk of Poland and Hungary at Varna in 1444.

The victorious march of the Ottoman forces into Europe was checked briefly, at the beginning of the 14th century, by internal problems within the empire and by Timur's threat in the east. Under the leadership of Mehmed I (1413–21), who won the struggle for succession, and his successor Murad II (1421–51), the march forward was resumed. The empire at this time also had to develop a navy to deal especially with Venice. Until the end of the 18th century this city-state was a major power in the eastern Mediterranean and waged with the Ottoman Empire a continual duel for control of the islands and the peninsulas of this sea—in particular, Crete, Cyprus, and the Morea.

The greatest single Ottoman military achievement occurred in the reign of Mehmed II (*q.v.*) the Conqueror

(1451–81). Although Ottoman territories now reached far into Europe, the Byzantine Empire, reduced to the city of Constantinople and a small surrounding area, still stood.

Fall of Constantinople

After long and careful preparation, Mehmed took the city in 1453, a date that compares in Greek history to that of Kosovo in the Serbian national memory. The fall of Constantinople marks also the completion of the process of the subjugation of the Balkan Peninsula to Ottoman rule. The empire now held most of Serbia, Bosnia, Bulgaria, and Greece. Although some isolated areas still resisted, no major centre of opposition remained. With the capture of the imperial city, the Ottoman conquerors made it their capital and developed an administrative system that was to dominate the Balkan region for the next five centuries.

The Ottoman Empire reached its height in power and prestige during the reign of Süleyman I, known in the West as "the Magnificent" and in Ottoman history as "the Lawgiver" (1520–66). A great military leader and an able administrator, Süleyman extended Ottoman territory far into the centre of Europe. Belgrade fell in 1521; in 1526 Süleyman defeated a Hungarian army at Mohács; and in 1529 the Ottoman army reached Vienna and lay siege to that imperial city. With the failure of this attempt, the limits were set to Ottoman westward expansion, but Süleyman's campaigns had left his country with control over most of Hungary and Transylvania. Transylvania, like the principalities of Walachia and Moldavia, taken in the 15th and 16th centuries, became an autonomous tributary province of the empire. See further OTTOMAN EMPIRE AND TURKEY, HISTORY OF THE.

Although some territorial changes occurred in the Balkans during the next century and a half, the Ottoman Empire lost no significant territory to Christian Europe until after 1683. Moreover, it kept at least titular control of the majority of the Balkan lands until 1878. Thus, for almost five centuries the greater portion of the Balkan people lived under an alien Muslim rule. It is the common experience of the peninsula and one that has profoundly shaped its present condition.

The character of Ottoman rule. When the Balkan people first fell under Ottoman control, they became a part of one of the great empires of world history, a worthy successor to the Roman and Byzantine states, which, too, once had their centres at Constantinople. The negative opinion often held of Ottoman civilization is usually based on judgments made of conditions in the 18th and 19th centuries, when the state was in a period of obvious decline. In the 15th and 16th centuries, however, Ottoman institutions may have offered the Balkan Christian a better life than he had led previously under his native feudal governments.

To understand the political conditions in the Balkans in the Ottoman era, it is first necessary to emphasize that the conquering power was a Muslim and not a Turkish national state. The Ottoman leaders regarded their peoples as divided by religious faith rather than by nationality. Any individual could join the ruling group by converting to Islām. Basing the political structure on this concept, the non-Muslim peoples were divided into five religious communities called millets: Orthodox, Gregorian Armenian, Roman Catholic, Jewish, and Protestant. Each group was under the direction of its religious head. Thus for the Balkan people, the vast majority of whom were Orthodox, the titular leader was the patriarch of Constantinople. But in practice, during the years of Ottoman rule, the Balkan Orthodox church organization became divided into its national components. Constantinople became a Greek centre; the Serbians had their own patriarchate at Peć, and the Bulgarians had a metropolitanate at Ohrid. The Romanians had similar national institutions. Thus national separateness and local tradition were preserved through the ecclesiastical organizations. The Ottoman government expected the church authorities to assume many civil functions, in particular judicial and tax-collecting duties. The Christian churches thus became, in a sense, part of the Ottoman state system.

The Ottoman government had a regularly organized administrative system with its centre in Constantinople and extending over the entire empire. The Balkan lands were part of the region (*beylerbeylik*) of Rumelia; the territories under direct administration were subdivided into provinces and then into lower administrative units. The principal interest of the Ottoman officials was the collection of taxes, which were to pay for the military power and the administration of the empire. Soldiers, policemen, and judges were also present in the administrative centres to maintain conditions beneficial to the proper collection of revenue. Although most of the Balkan lands were administered by Ottoman officials, some areas enjoyed unusual rights of self-government. The Danubian principalities of Walachia and Moldavia, for instance, as well as Transylvania, being autonomous tributary regions, were governed by their own aristocracy. The merchant city of Dubrovnik was an autonomous republic. Certain other cities and regions either had won special rights or were too remote and primitive to arouse Ottoman concern. They enjoyed almost complete self-government. Even among the areas under direct administration, the Ottoman authorities remained chiefly concerned with taxation and the maintenance of public order. They did not attempt to regulate other details of Balkan Christian life. These fell under the jurisdiction of the local village authorities or the church.

Although only Muslims could hold office in the empire or serve in its military forces, converted Balkan Christians came to occupy the highest positions at this time. In fact, these converts came to form the main basis of the military and administrative apparatus of the state. By the time of Süleyman, the highest state offices were held by slave administrators. These were either purchased, were prisoners of war, or were acquired through the *devşirme* system: from the 14th century to about the middle of the 17th, approximately every five years one out of four boys in the Balkans, aged between 10 and 20, was taken from his parents, brought to Constantinople, and converted to Islām. The most able of these were sent to the palace school, where they were trained as administrators, and then assigned posts throughout the empire. Other children recruited in this manner became members of the Janissary Corps—an elite infantry unit armed with muskets—which became the most effective fighting arm of the Ottoman army. Forbidden at first to marry, this body of new converts proved to be fanatic and dedicated soldiers.

The *devşirme* system

In the first period of Ottoman rule, conditions for the Balkan peasant on the land were probably not overly onerous; in fact, he may have enjoyed a better position than his counterpart in western Europe. In theory, the sultan held all of the land taken in war; he was free to dispose of it at will. In practice, the Balkan lands were used to support cavalry units of the army. The members (spahis) were given grants of land (timars) in return for which they had to provide military service. At first, the amount that the peasant had to contribute in dues and services was carefully regulated. Later, however, the system tended to break down. With the introduction of the wide use of firearms the effectiveness of the cavalry was reduced. The government became more interested in increasing the tax load to pay for new weapons.

Despite the fact that the Ottoman government showed a high degree of toleration for non-Muslim faiths, there was no question of equality between religious groups. To rise in Ottoman society the Balkan Christian had to abandon his faith. There were few attempts at forced conversion, but the subordinate position of the Christian was constantly emphasized. He could not wear conspicuous or rich clothes, for instance, or the colour green, sacred to Islām. If when on horseback he passed a Muslim, he was forced to dismount. Old churches could be repaired, but new ones could not be built. They could not have bells or bell towers or be constructed in a place or manner likely to "offend the eyes of the faithful."

Even more important, the Balkan Christian was forced to carry an unequal share of the tax burden of the empire. Although he was not subject to military conscription, he was assessed a special tax as a replacement. He was also liable to other services and payments, particularly in time

of war, which were connected to his secondary status within the state. Despite these and other severe disabilities, it is interesting to note that there were relatively few examples of mass conversions among the Balkan people, with the exception of those that occurred in Bosnia, Albania, and Crete, where local conditions were unusual.

Despite the fact that the Ottoman system of government served to maintain the separation of the national groups, great similarities nevertheless existed in the conditions of life of all of the people who lived under direct Ottoman rule. The vast majority of the Balkan population lived as farmers or herdsmen, either on estates or in the less economically valuable hills and mountains on virtually independent family farms. They inhabited small primitive houses, clustered in villages, where their local and personal affairs were managed by the village elders and the church. Larger towns and cities served principally as trading and administrative centres. Ottoman officials usually resided in the largest cities, as did those Muslims who held large estates. Handicrafts and trades, with their centres in these cities, were organized in a guild system.

The average Balkan peasant in these circumstances lived a primitive existence, cut off from the rest of the world. Since education was controlled by the church and was extremely limited, illiteracy was nearly universal. The life of the individual was limited to his village. Nevertheless, in each region a unique peasant culture was retained. Each area and national group had its own style of houses, decorative patterns, and, most important, its traditional songs and stories that substituted for a written literature.

While the smaller Balkan communities reflected the local peasant culture, the Balkan cities, as the military and administrative centres for the Ottoman authorities, were built according to the architectural preferences of the ruling group. Typical of Ottoman architecture were the massive stone bridges, fortresses, mosques, baths, covered markets, and caravansaries.

Although most of the Balkan peoples tended to prefer to remain on the lands of their ancestors, they could, of course, as citizens of one state, move about within the empire. During the four centuries of Ottoman domination, there was not only much internal migration, due to war and similar causes, but certain nationalities tended to specialize in occupations that drew them out of their national area. For instance, Greek merchants could be found in most major Balkan cities; Vlach (Romanian) shepherds roamed the mountains of the entire peninsula. Some Turks moved from Anatolia to areas such as the Black Sea coast, Macedonia, Bosnia, and Thrace. Albanians served as guards or police in many localities. This intermixture of population caused complications later when it became necessary to draw the boundaries of the national states.

The decline of the Ottoman Empire

Despite the long existence of the empire, signs of weakness appeared in its structure as early as the reign of Süleyman. The Ottoman system could not function well without strong direction from the top. Up through Süleyman the state had been remarkably fortunate in its rulers, but among the 17 sultans who followed, few were men of ability. With the lack of firm direction from above, the administration based on slave officials who rose by merit inevitably changed. Born Muslims and sons of officials naturally sought to enjoy the rewards of public office. The *devşirme* system came to an end by the middle of the 17th century. Parallel with these developments was the increasing corruption of all aspects of Ottoman political life. High offices were regularly sold; the officials who purchased their positions were primarily concerned with recovering their expenses and enriching themselves. The breakdown of the Ottoman administration, which commenced at the centre in Constantinople, spread out to encompass the entire governmental network.

Most dangerous for the future of the empire was the decline of the Janissary Corps. Though its members had gained the right to marry and to enlist their children in its ranks, they were soon forced to earn extra money and many became craftsmen on the side. Thus, this once-powerful military unit degenerated into a weak organization whose chief role in the state was to exert influence on the government. The corps, like the rest of the Ottoman military forces, failed to keep up with Western advances in military technology and organization. The military pre-eminence and the qualities of leadership that had won for the Ottoman Empire its vast domain were now lost.

The Christian, together with his Muslim neighbour, was directly affected by the decline of Ottoman administration. His chief grievances became the willful and arbitrary manner of the local officials and the grave abuses that were now associated with the collection of taxes. The central government relied for this service on tax farmers who competed yearly in bidding for the position. The tax farmer was chiefly interested in making a profit from his investment. Forceful methods and unfair standards of assessment were regularly used to extract high payments. The system was economically ruinous for the peasant and the central government alike.

The Balkan peasant found his standard of life also deeply affected by a change in the landholding system. As mentioned previously, under the timar system, strictly controlled nonhereditary land grants had been made in return for military service. With the relative decline in importance of the cavalry and the reduction in the number of military campaigns, large estates tended to pass into the hands of those who wished to work them at a profit and who now held them as hereditary rights. The peasant who worked the fields of the estates (now called *chifliks*) was reduced to the position of a sharecropper; his dues in kind and in labour were greatly increased.

Among the most severe consequences of the breakdown of the Ottoman government was the rise of lawlessness throughout the peninsula. As the central authority grew weaker, local Muslim leaders throughout the empire established what were in effect small principalities from which they were able to defy Constantinople and war among themselves. The most important of these for Balkan affairs were Ali Pasha of Janina and Pasvanoglu of Vidin. Christian bands of robbers (haiduks or klephts) also existed in large numbers. Their activities, together with the destruction caused by the Ottoman wars of the 18th century, made certain areas uninhabitable for long periods. These conditions were the direct cause of both the Serbian revolt of 1804 and the Greek revolution in the Morea in 1821.

Although all of the inhabitants of the empire suffered from this situation, certain Balkan Christians nevertheless did benefit from the growing weaknesses in the Ottoman system. Of the non-Muslim people, certainly the Greeks associated with either the Ottoman administration or the commercial world were in a privileged position. By the 18th century certain Greeks residing in Constantinople, known as Phanariotes, had come to secure regularly certain high appointments in the Ottoman service. Most important were the offices of grand dragoman (secretary of state) and the governorships (*hospodar*) of Walachia and Moldavia. At the same time, through the patriarchate at Constantinople the Greek hierarchy was able to dominate the other national churches. In the middle of the 18th century, it secured control of the Bulgarian and Serbian ecclesiastical centres at Ohrid and Peć. In addition, Greek merchants were established in the cities of the peninsula, and they controlled a large part of the trade of the Black and the Mediterranean seas. These merchant communities were the most susceptible to the influence of Western political and social ideas.

The position of the Greeks in the empire

After the Greek, the Romanian landowner of the privileged provinces of Walachia and Moldavia held a superior position within the empire. These principalities were never under direct Ottoman administration, although they did pay tribute and they were subject to regulation from Constantinople. In both, a native aristocracy held control over an enserfed peasantry. After the Russian invasions of the early 18th century, during which Romanian leaders cooperated with the enemy, the Ottoman government appointed Greek rather than Romanian governors for the provinces. The Phanariote Greek regime, which lasted until 1821, was extremely corrupt and was strongly disliked by the Romanian aristocracy.

In comparison with these groups, the Serbs and Bulgarians—without a native aristocracy and largely peasants—were at a distinct disadvantage. Both nations suffered severely from the corruption and lawlessness of the time. Their lands were the scene of repeated struggles between local Muslim leaders and of battles with the invading great powers. In the 18th century, both found their national churches taken over by Greeks.

A special word must be said concerning the Albanians and the Montenegrins. As the only Balkan people among whom a majority (70 percent) converted to Islām, the Albanians were not subject to many of the problems of their Christian neighbours. Many rose high in the Ottoman state and in military service. Moreover, the central government made little effort to control closely this distant, backward, mountainous region. Relatively content with their status, the Albanians were the last to establish an independent state. Montenegro, too, although a Christian state under a prince-bishop, because of its poverty and the difficulty of access to its lands, was usually able to govern itself.

The increased role of the European powers. Although the internal problems had become acute long before, in international affairs the empire was able to preserve its territories almost intact until the end of the 17th century. In the second half of that century, a last great offensive was begun and additional lands in the Ukraine were acquired. The symbolic turning point for the Ottoman offensive against Europe was the second siege of Vienna, in 1683. The failure to take the city marked a reversal in Ottoman fortunes. Henceforth, the European states moved against Ottoman possessions. Russia, Venice, Poland, and Austria now joined in a victorious coalition. In 1699 the empire was forced to sign the Treaty of Carlowitz, the first time that the Ottoman government appeared on the diplomatic stage as a clearly defeated power. By this treaty, Austria gained control of most of Hungary, Transylvania, Croatia, and Slavonia; Venice took Dalmatia and the Morea; Poland gained Podolia. Thus large numbers of Balkan people passed from Ottoman to Austrian and Venetian control. The agreement established what was to be until 1878 a relatively stable Austrian–Ottoman frontier along the Danube and Sava rivers. The main territorial gains at the expense of the Ottoman Empire in the next century were to be made by Russia.

Russian invasions

The Russian advances against Ottoman lands began under Peter I the Great at the end of the 17th century. The major acquisitions were made, however, by Catherine II the Great (*q.v.*), who by the end of her reign had won the lands north of the Black Sea to the Dniester River. For the future of the Balkan Peninsula the most important agreement between Russia and the Ottoman Empire was the Treaty of Küçük Kaynarca (1774). By its terms, Russia gained territory in the Black Sea region and certain trade rights, but, most important, because of the ambiguous wording of the treaty, Russia was later to claim that it had won certain rights to speak in behalf of the Orthodox Christians of the empire. In 1781 Catherine joined with Joseph II of Austria (*q.v.*) in what was to be the first of numerous partition agreements among the great powers. According to this agreement, both Austria and Russia were to receive territory, but the majority of the lands of the Ottoman Empire in Europe were to be organized into a Kingdom of Dacia, under Russian influence, and a Greek kingdom with its capital at Constantinople and with Catherine's grandson as its ruler. General European events prevented the success of this venture, but, by the end of the century, it was apparent that Russia was both the chief threat to the empire and the major hope of the Balkan people for outside aid.

The French Revolution and the Napoleonic era. The wars of this period were to have a deep effect on the Balkans, both from the international aspect and because of the response that French Revolutionary ideology evoked within the peninsula. This period also witnessed the entrance of France and Britain actively into the diplomatic and military struggle over the control of the Ottoman Empire. That state, in the midst of another grave internal crisis, now found its lands under attack from France. In 1797 France took the Ionian Islands from Venice. A year later, Napoleon launched an invasion of Egypt and Syria. As a consequence of this action, the empire was in alliance with Britain and Russia from 1798 to 1802. After a period of peace from 1802 to 1806, the Ottoman government shifted sides and joined France against Russia. This alliance proved disastrous, because in the next year, by the Treaty of Tilsit, France came to terms with Russia. It is interesting to note that at this time these two governments discussed a partition of the Ottoman Empire, but they were unable to agree on the fate of Constantinople. In 1812 Russia, faced with an impending French invasion, signed the Treaty of Bucharest with the Ottoman Empire, which ceded the Romanian territory of Bessarabia to Russia.

The Balkan peoples were affected not only by the wars that were waged in the area, but also by the political arrangements made by the several powers. Most influential for the future was the introduction into the Balkans of French Revolutionary institutions and ideology through the Napoleonic conquests. For example, in the Ionian Islands, which France held from 1797 to 1799 and again from 1807 to 1814, a constitutional government on Western patterns was established. The Septinsular Republic was the first Greek national government in modern times. French political experiments also involved South Slav lands. In 1809 Austria was forced to cede to France its Balkan territories. From Dalmatia, Slovenia, Istria, Trieste, and parts of Croatia a new political entity, the Illyrian Provinces, was formed and became a part of the French Empire. As in the Ionian Islands, the government was based on revolutionary principles. The Illyrian Provinces has been described as the first Yugoslav state, as it included within its boundaries Serbs, Croats, and Slovenes.

This period also brought about important changes in Moldavia and Walachia and in Serbia. The Ottoman Empire was forced to agree to an increase of Russian direct influence in the Danubian principalities; henceforth, no governor of either province could be dismissed without Russian consent. Russia thus acquired what was, in effect, a protectorate over the country. This power also played a major role in the events of the first Serbian revolution (see below).

The defeat of Napoleon in 1815 resulted in the restoration of the political and territorial status quo in the Balkans, with only a few exceptions—Russia, of course, retained Bessarabia; the Ionian Islands were placed under British protection; the Illyrian Provinces disappeared as a political entity; and Austria received back its former land and, in addition, acquired Dalmatia. The Ottoman Empire, which did not attend the Congress of Vienna in the aftermath of the Napoleonic Wars, retained most of its territories intact. It was also to be aided by the conservative reaction that followed in a Europe tired of war and revolutionary upheaval.

In Balkan history the French Revolution and the wars of Napoleon mark the shift from the long period of Ottoman domination into the era of the national revolutions of the 19th century. To some extent, the stage had already been set in the 18th century, when both Russia and Austria appealed to the Balkan subject populations for assistance against the Ottoman Empire. These wars, as well as those of Napoleon, had shown how weak the Ottoman military forces really were. Moreover, Balkan nationals had fought in these wars. They had learned modern military methods and they were armed. Equally important, the period of war and internal upset had opened the area to outside influences. The national and liberal ideology of Revolutionary France provided a program that would allow Balkan leaders to combat not only Ottoman political control but also the stifling cultural influence of their Christian church hierarchies.

Although the Balkans were now ready for revolt, the international situation was to change to the detriment of such actions. After the period of war, which had wasted the resources and energy of Europe before 1815, the powers desired, above all, peace with political and social

stability. There occurred in all the European states a conservative reaction directed against revolutionary methods and liberal–national programs. Subject Christian peoples could thus not expect aid or sympathy from abroad. At the same time, it was also apparent to the European powers that the Ottoman Empire was in a dangerous condition of internal decay and military weakness. The question of the fate of the Ottoman territories and of the control of that government became perhaps the most important single diplomatic problem for Europe in the century after the Congress of Vienna. This issue, the so-called Eastern Question, was the direct cause of the two great wars of that period—the Crimean War and World War I—and the occasion of repeated less serious conflicts among the powers. See EUROPEAN DIPLOMACY AND WARS (C. 1500–1914): *Revision of the settlement of Vienna: The Eastern Question again: The Crimean War.*

The basic problem in the disposition of the Ottoman lands was their strategic position across three continents. Because of its past history and its close links with the Balkan people through the Orthodox Church, Russia was in the best position to gain predominant influence in the area. Russia's chief rivals were Austria and Great Britain. The Habsburg monarchy could not afford to allow Russia more political control in lands along its frontiers. The British feared for their communications with their empire, which ran through the eastern Mediterranean, and their control over India. Moreover, despite its favourable relations with the Balkan people, the Russian government, too, because of its abhorrence of revolutionary activity and liberal reform programs, stood after 1815 against national revolt. Thus, although the era of the French wars had prepared the Balkan people for revolution, the international situation was not favourable. Nevertheless, the next 65 years were to witness the establishment of independent, or autonomous, governments for almost all of the Balkan national groups. (C.J./B.Je.)

II. The Balkans from 1815 to 1914

The era of national liberation

The 19th century saw the political and social development of the Balkans and the formation of independent states. During this period the Balkans were drawn, economically and culturally, into the orbit of contemporary Europe. While it is somewhat arbitrary to divide this exceedingly complex and dynamic era into segments, it is possible to delineate four general periods: (1) the beginning of the process of national liberation during the first Balkan revolutions, 1800–30; (2) a period of political and social development, 1830–78; (3) the inclusion of the Balkans in Europe during the age of imperialism, 1878–1903; and (4) the Balkans in European crisis on the eve of World War I, 1903–14.

NATIONAL REVOLUTIONS (1804–30)

At the beginning of the 19th century the Habsburgs ruled the northwestern part of the Balkans, while the Ottomans dominated the main central and southern regions of the peninsula. The development of a national renaissance of the Balkan peoples varied according to the different political and social conditions prevailing in these areas: political development occurred under the Habsburgs, and revolutionary upheavals predominated under the Ottomans.

Three main factors influenced the national revolutions in the areas under the Ottomans: the decline of the Ottoman Empire, provoked by a general crisis of Ottoman feudalism; the gradual shaping of a new Balkan society, as a result of the economic traffic with Europe and the development of local autonomy; and the influence of the "outer" Balkan world on the "inner" (*e.g.*, the influence of the Greek, Bulgarian, and Albanian trading colonies in the Mediterranean and Black Sea areas, of the Serbs in southern Hungary, etc.). The Balkan revolutions had two general aspects: nationalistic and agrarian. The national aspect was expressed in the drive for national liberation, the creation of national economies and cultures, and the political organization of national states. The agrarian aspect was marked by the endeavours of the peasantry to get rid of the Ottoman landlords and take possession of the land.

The Yugoslavs. The Yugoslavs were divided under the Ottomans and the Habsburgs. The Serbs in the south and east were part of the Ottoman Empire and the Slovenes and Croats in the north and west were ruled by the Habsburgs.

The rise of the Serbian principality (1804–30). The creation of the autonomous Serbian state went through three stages: an uprising in 1804–13, the restoration of Ottoman power in 1813–15, and the uprising that led to the achievement of autonomy in 1830. The peripheral position of the Serbs in the Ottoman Empire facilitated the outbreak of the 1804 uprising in the pashalic of Beograd (Belgrade). The movement began as a revolt against the local Ottoman authorities, who themselves had rebelled against the Sultan. After overthrowing the local rulers, the Serbian insurgents, led by the former peasant, Karageorge, went on to fight the imperial army in 1805 and 1806. Through a number of victories in 1806, they liberated all of Serbia. The outbreak of the Russo-Turkish war in 1806 encouraged the insurgents; in the hope that they might win their independence, they asked for help and protection from Russia, Austria, and even Napoleon. But the Napoleonic Wars created an unfavourable international situation and, except for some passing assistance from Russia, the Serbs had to rely on their own forces. The revolution was in motion, however: the Serbian insurgents broadened their political program to include complete independence and planned to join forces with the Montenegrins, Bosnians, and Bulgarians. But the movement collapsed in 1813, after Russia concluded the Treaty of Bucharest with Turkey (1812), enabling the much superior Ottoman army to crush the rebels. Karageorge, accompanied by other leaders and a part of the population, escaped to Austria.

The Serbian struggle for freedom

During the 1804–13 revolution the Ottoman social, economic, and political system in Serbia was completely destroyed and the nucleus of the future Serbia was established. The peasantry carried out an agrarian revolution; a class of small peasant proprietors developed and was destined to become important in the future. The basis of a national economy emerged in the liberated areas. Serbian national culture was shaped through the work of Dositej Obradović, the educational reformer and translator, and Vuk Karadžić, a reformer of the literary language and collector of folk poetry. The effect of the 1804–13 uprising was to make the complete restoration of Ottoman power impossible. This became obvious when another uprising occurred in 1815 under the leadership of Miloš Obrenović. After defeating the Ottoman garrisons in Serbia, Miloš reached a compromise with the Turks that enabled him gradually to widen the bounds of Serbian autonomy. Over the years from 1815 to 1833 he was able to take advantage of growing Serbian strength and declining Ottoman power. Developments elsewhere in the Balkans helped this evolution: rebellions of the Turkish pashas in Bosnia and Albania; an uprising in the Danubian principalities; the Greek revolution; and the Russo-Turkish war of 1828–29. A Russo-Turkish convention, signed at Akkerman in 1826, required Turkey to fulfill its obligations under the Bucharest treaty of 1812 with respect to Serbian autonomy. These stipulations were reiterated in the Russo-Turkish Treaty of Adrianople (1829), together with the requirement that Turkey restore to Serbia the areas in the east that had been liberated in the first rising but later recovered by Turkey. When the Turks were slow to comply, Miloš organized an uprising in those districts and joined them to Serbia in 1833. He was now recognized as the hereditary prince of an autonomous principality under the suzerainty of the sultan.

The Slovenes and Croats. The development of the Slovenes was closely linked with that of central Europe. Placed as they were between the Austrians and the Italians, divided into six political units (Carniola, Styria, Carinthia, Görtz, Trieste, Istria), and exposed to the centralistic policies and Germanizing influence of the Habsburg Empire, the Slovenes encountered many obstacles

The Habsburg domain

on their way to national emancipation. In order to isolate Austria from the Adriatic, Napoleon in 1809 organized the Illyrian Provinces (Slovenia, Dalmatia, a part of Croatia, and the Military Frontier [Vojna Krajina]) with their capital in Ljubljana. His rule introduced French ideas of administration and education, but after 1813 Austria restored the old system. The Slovenian national renaissance found its expression in cultural and literary life, rather than in politics.

The Croatian lands were divided in the 19th century: civil Croatia and Slavonia were ruled by Hungary, while the Military Frontier (a zone of defense against Turkey) and Dalmatia were under Austria. The upper layers of society were either foreign or had been Germanized, the middle class was rudimentary, and the peasantry lived in serfdom. The Croatian national renaissance passed through three phases: the struggle for the use of the popular language in administration and education; the demand for cultural and territorial autonomy; and, finally, the movement for national independence. Croatian politics evolved within the Zagreb-Budapest-Vienna web of Habsburg power. The first Croatian-Hungarian clash occurred in the 1820s when Hungarian nationalism sought to develop a Magyar state from the Carpathians to the Adriatic. This meant Magyarization for all the peoples under Hungarian rule. The Croats responded by stressing their own language and culture.

The Greek revolution. The Greek revolution was even more successful than the Serbian and, from the historian's point of view, represented another large step toward the affirmation of the nationality principle in the Balkans. The Greek achievement was the consequence of Greek social development and of European interference in Greek affairs. The Greeks had created under the Ottomans two societies: one on the mainland and another in the colonies established by their traders all over the Balkans, the Mediterranean, and western Europe. The Greek merchant became the nucleus of the Balkan middle class, the Greek language a lingua franca for the Balkans. To this was added the cultural and political power of the Greek Church in Constantinople and of the Greek Phanariote families who governed Turkey's Danubian principalities from 1711 to 1821. During Napoleon's continental blockade, the Greek merchant fleet dominated trade in the eastern Mediterranean; in 1813 it was composed of 615 ships and 37,500 sailors. The Ionian Islands, occupied by the French in 1797–99 and 1807–14, served as a base for the spread of revolutionary ideas. Various secret organizations cherished the idea of the liberation of Greece. The most effective of them was the Philikí Etairía, formed in Odessa in 1814 to organize a general Balkan uprising for the liberation of Greece. A new impulse was given to its activities when the Phanariote Greek Alexandros Ypsilantis, a Russian army general, accepted the leadership in 1820. Planning to start an uprising in Turkey's Danubian principalities that would spread all over the Balkans to embrace Greece, Ypsilantis invaded Moldavia in March and entered Bucharest. But the attempt ended in failure: the Russian tsar Alexander I denied him the expected support; the general Balkan uprising did not occur; and the Greek movement in the principalities clashed with a similar Romanian movement. Defeated by Ottoman troops at Drăgăşani, Ypsilantis was forced to flee.

Reasons for Greek success

The activities of the Philikí Etairía and the simultaneous rebellion of Ali Pasha in Epirus opened the way for the outbreak of the uprising on the Greek mainland. On March 25 Bishop Germanos raised the banner of revolt at the monastery of Aghia Lavra in the Peloponnesus. The ensuing Greek revolution went through three phases: local successes in 1821–25, the crisis caused by the intervention of Muḥammed ʿAlī of Egypt in 1826–28, and a period of European intervention ending in Turkish recognition of Greek independence in 1832 (see also MUHAMMED ALI PASHA). At the beginning the revolution spread in a flash, embracing the Peloponnesus, central Greece, and the island of Crete. By the summer of 1822 the revolutionaries had captured Missolonghi, Athens, and Thebes. At that point the revolution was checked by the Ottomans and by internal conflicts among the Greeks. The Ottomans hanged the patriarch Gregorios in Constantinople and suppressed the uprising in Thessaly, Macedonia, and Mt. Athos. A stalemate prevailed until the beginning of 1825. There were dissensions among the Greek insurgents, and the aspirations of different classes and regions were too various to be reconciled. In 1822 two Greek governments existed: one on the mainland dominated by Theodóros Kolokotrónis, and the other at Ídhra led by Geórgios Kountouriótis and Aléxandros Mavrokordátos. By 1824 open civil war prevailed in Greece. A turning point came with the intervention of Egyptian forces under Muḥammed ʿAlī's son Ibrāhīm, who landed on the Peloponnesus with a large army in February 1825. By June 1827, burning and plundering, he seemed about to reconquer Greece. Missolonghi, the key fortress of Corinth, fell in April 1826 after a heroic defense, marked by the death of the English poet Lord Byron. The Greek revolution was rescued by the powers of Europe, who, motivated by complex political factors, offered mediation to the belligerents in 1827. An accidental clash between the Ottoman and European fleets at Navarino Bay on October 20, 1827, resulted in the total annihilation of the Turko-Egyptian fleet. In 1828 Britain and France took advantage of the Russo-Turkish war to conclude an agreement with Muḥammed ʿAlī, providing for the evacuation of Egyptian forces from Greece. In March 1829 England, France, and Russia agreed on autonomy for Greece, but a final settlement of Greek affairs was not achieved until February 1830, when the three powers declared Greece an independent monarchy under their protection.

The Greek success furthered the idea of national emancipation throughout the Balkans. The Serbian and Greek revolutions together highlighted and aggravated the European dilemma that was becoming known as the Eastern question: *i.e.*, whether the Ottoman Empire should be preserved for the sake of the European balance of power or should be divided among its successor states. The appearance of the first Balkan states gradually changed the role of the Balkans in European politics: from being objects of European policy, the Balkan peoples were becoming its subjects.

The Romanians at the beginning of the 19th century. Nineteenth-century Romanian development was dominated by the facts of geography and power. The Romanians had as neighbours the Russian, Habsburg, and Ottoman empires. They began the century politically divided: Walachia and Moldavia under the suzerainty of the sultan; Bessarabia under the Russians (from 1812); and Transylvania, the banat of Timisoara, and Bukovina under the Habsburgs. Most Romanians were peasants, and their masters were the Phanariotes, boyars, and the church with its landed estates. Romanian prosperity was based upon exports of grain, stimulating traffic down the Danube and through the Black Sea ports of Galaţi and Brăila. The rising middle class became the agent of the Romanian renaissance. A new national intelligentsia, educated in western Europe, replaced the traditional Old Slavonic, Greek, and Latin with a specific Romanian culture.

The struggle for national unity

A popular national movement in 1821, led by Tudor Vladimirescu, paralleled that of the Greek Etairísts. Although himself an Etairíst and a former lieutenant of the Russian army, Vladimirescu was interested in the Romanian rather than the Greek cause. His program envisaged the election of a national assembly, organization of a national army, and tax reform. But the peasant insurrection was inevitably transforming itself into an agrarian revolution, directed against the boyars. Alienated from the Greek Etairísts by his Romanian goals and from the boyars by the spectre of revolution, Vladimirescu found himself isolated. He tried to negotiate with the advancing Ottoman troops but was captured by the Etairísts and executed in the summer of 1821. His movement was suppressed by the Ottomans. The events of 1821 brought the end of the Phanariote Greek rule in Walachia and Moldavia; the Turks replaced them with Romanian boyars. The Russo-Turkish War of 1828 led to

Russian occupation of the two principalities and the promulgation of the Règlement Organique, which recognized the boyars' landed estates and imposed further restrictions on the peasantry. Exposed to both the old feudal relations and to a developing market economy, the condition of the Romanian peasant worsened during this transitional period.

Obstacles to Bulgarian nationalism

The Bulgarians in the early 19th century. The Bulgarians were at the centre of the Ottoman Empire, exposed to Turkish colonization, economically part of the Ottoman market, and culturally influenced by the Greeks. Nationalist impulses arose from internal economic and social change and from the Bulgarian merchant colonies in the Black Sea area. At the same time the decline of the Ottoman system aggravated the situation of the Bulgarian peasant. In the period from 1792 until 1815 Bulgaria was plagued by armed bands known as Kurdzhali. The most famous rebel was the feudal lord, Osman Pazvanoglu, who in 1794 captured the fortress of Vidin and ruled independently of the Turks. Bulgarian volunteers took an active part in the Serbian and Greek revolutions, as well as in the Russo-Turkish wars. During the war of 1828–29 a Bulgarian committee in Bucharest sought to obtain from the Russian high command the same rights accorded to Serbia, Greece, and the Danubian principalities. Georgi Mamarchev led an unsuccessful uprising in Sliven in 1829. Another attempt by a Türnovo merchant, Velcho Atanasov, in 1835 ended with his arrest and execution.

The Bulgarian cultural renaissance got under way at the beginning of the century. Until the 1830s basic education was developed in schools run by Greek monks. These schools were slowly replaced by Helleno-Bulgarian schools, in which the teaching was modernized and Bulgarian was taught along with Greek. The impulse toward the Bulgarization of education was supplied mainly by emigrants from Russia and the Danubian principalities.

THE DEVELOPMENT OF BALKAN STATES AND SOCIETIES (1830–78)

In the half century after 1830 the principle of national rights made further progress. Three new Balkan states appeared: Romania, Montenegro, and Bulgaria. The peoples living under the Habsburgs took an active part in the 1848 revolution and in the 1867 reorganization of the monarchy. The continuing decline of Ottoman power, and the failure of intended reforms in 1839, 1856, and 1876, gave further impetus to Balkan national revolutionary movements, which reached their zenith in the Eastern crisis of 1875–78.

Slovenian and Croatian nationalism

The Yugoslavs. *Slovenes and Croats under the Habsburgs.* The middle class of Slovenia took advantage of the revolutionary upsurge in 1848 to stress their national program: unification of the Slovenian lands in a federalized Austria, the use of the vernacular language, and a closer cooperation with the other South Slavs. These aims were reinforced by the liberal movement of the 1860s, which took over leadership from the old conservative groups. In 1871 a Yugoslav congress was held in Ljubljana, stressing the need for cooperation among all Yugoslavs living under the Habsburg monarchy.

The Croatian national renaissance was manifest in the Illyrian movement of the 1830s. Its basic idea was to link all Croatian and other Yugoslav lands in a single cultural unit bound by a common literary language and national consciousness. Starting primarily as a cultural movement, the Illyrians gradually became a political party opposing Hungarian domination and the centralistic policy of Vienna. The 1848 revolution that shook the Habsburg monarchy provided the Croats an opportunity to struggle for their national program: unification of all Croatian lands (civil Croatia, Slavonia, Dalmatia, and the Military Frontier); autonomy within a federalized monarchy; and use of the vernacular language in public life and education. In 1848 a similar program was accepted in southern Hungary by the Serbs, who proclaimed an autonomous Serbian Vojvodina. When nationalistic Hungarians rejected their demands, the Croats helped Vienna against the Hungarians. There were also clashes between Serbs and Hungarians in the Vojvodina. After the Revolution of 1848 had been suppressed, the Yugoslavs were subjected to new centralizing policies from Vienna. In the years preceding the reorganization of the Habsburg Empire in 1867, Croatian politics was divided between the three poles. The majority Popular Party called for a unification of Croatian lands in a federalized monarchy based on a Yugoslav program. Another group (called Unionists) desired a compromise with the Hungarians. The young Croatian nationalists espoused a program of Croatian national and political individuality under the Habsburgs. The establishment of the Austro-Hungarian dual monarchy in 1867 was followed by a compromise between Hungary and Croatia in 1868, whereby the kingdom of Croatia was given some kind of autonomy under Hungary. This failed to satisfy the nationalistic impulse in Croatia, which found expression in Ante Starčević's Party of Rights.

Formation of Serbian state

Serbia and Montenegro (1830–78). The autonomy achieved by the Serbs in 1830 was followed by internal dissension. Opposition to the autocratic rule of Miloš Obrenović compelled him to abdicate in 1839. Under the rule of Karageorge's son, Alexander, from 1842 to 1858, the government was administered by an oligarchical State Council. The agrarian revolution that had previously occurred was reflected in a law of 1833. Civic rights and efficient government were secured by the civil code of 1844 and by a series of laws concerned with public affairs and education. The aspirations of the Serbs found expression in 1844 in a National Program that called for the unification of all Serbs and access to the Adriatic Sea; efforts were made to achieve this with the help of the western European powers and of other Yugoslavs. The young liberals brought Miloš Obrenović back in 1858. On his death in 1860 he was succeeded by his son, Prince Michael Obrenović, who ruled until 1868. Influenced by the rising tide of nationalism in the Europe of the 1860s, Serbian politics under Michael took a revolutionary course. Michael wanted a general Balkan uprising, a Balkan alliance, and a common war against the Ottomans. In these aims he was seconded by Hungarian and Polish émigrés, Italian national revolutionaries, and Russian Pan-Slavs. Even such statesmen as Cavour and Bismarck encouraged the Serbs in the hope of weakening the Habsburg Empire. After an incident in Beograd, the Ottomans bombarded the city in 1862. War was avoided through the mediation of the European powers, and Ottoman garrisons were withdrawn from the Serbian cities in 1867. The rest of the Turkish population also left Serbia.

One of the problems encountered by the Ottomans in other territories under their rule was the difficulty of reforming the administration of their empire against the opposition of local beys, or governors. This opposition was especially strong among the conservative landlords and local governors in Albania and Bosnia-Hercegovina; it contributed to the chronic unrest among those populations. The Ottomans sent in their armies several times in the period 1830–50 to put down uprisings.

Montenegro was also involved in conflict with the Ottomans in 1852, 1858, and 1862. The tribal Montenegrins had never been subjugated by the Turks. The formation of a Montenegrin state progressed through these wars, and a central state authority was gradually established. In 1860 Montenegro obtained international recognition.

Greek politics from 1830 to 1878. The development of government administration in the new Greek state was accompanied by conflicts between centralizing and decentralizing tendencies at various levels of Greek society, as well as by foreign intervention. The first president of Greece, Count Ioánnis Kapodístrias, was assassinated in 1831. In 1832 the European powers imposed Prince Otto of Bavaria, who became King Otho. In the first decade of his rule, he introduced a system of centralized, absolute monarchy run mainly by Bavarians. In 1833 a Greek Orthodox Church, independent of the patriarchate in Constantinople, was established. Byzantine law was introduced in 1835, and the University of Athens was created in 1837. But general dissatisfaction with the

rule of foreigners led to a revolt in 1843, forcing the king to dismiss his Bavarians and agree to a new constitution.

Greek irredentism

Greek nationalism was ascendant, both within Greece and throughout the eastern Mediterranean. The consequence was a program of irredentism known as the Megali Idea (Great Idea) that envisaged a Great Greece bounded by the Adriatic, the Black Sea, and the Mediterranean, with its capital in Constantinople. Secret national societies devoted to this goal sprang up in Epirus, Thessaly, Macedonia, and Crete. The revolutionary activities of Greek guerrilla bands in Thessaly and Epirus during the Crimean War (1853–56) strained relations with the Ottomans and also with the French and British. The spread of liberal national ideas throughout the Balkans in the 1860s was reflected in renewed Greek attempts to provoke a Balkan uprising, a rapprochement between Greece and Serbia, the deposition of the Bavarian dynasty in 1862, the unification of the Ionian Islands with the mainland in 1864, and a large, bloody uprising on the island of Crete in 1866. After the ouster of King Otho in 1862, Prince William of the Danish Glücksborg dynasty was elected "George I of the Hellenes." A new, democratic constitution was adopted in 1864, and in the 1870s a two-party system began to emerge. In the years from 1864 to 1880 Greece went through nine elections and had 31 governments.

The birth of Romania. For half a century the Danubian principalities had offered hospitality to revolutionary émigrés from the Balkans, and as a result a number of Balkan uprisings in the 1840s were planned in the principalities. In the meantime the idea of Romanian national unity evolved and was especially cherished by the young Romanian intelligentsia. The radical Philharmonic Society, established in 1833, stressed a program of unification and political freedom but failed to realize these goals during the Eastern crisis of 1839–41. The 1848 revolution offered another opportunity to those favouring unification. During the revolutionary days in Paris, Romanian students hoisted the Romanian flag at the Hôtel de Ville. A secret revolutionary organization, the Fratria (Brotherhood), aroused revolutionary fervour at home. An uprising in Moldavia proved a fiasco, but another in Walachia was successful. In Transylvania 40,000 peasants gathered on the "Field of Liberty" and demanded liberal reforms. A civil war broke out between Romanians and Hungarians in Transylvania, joining social to national conflicts. A joint Russo-Turkish military invasion put an end to the liberal nationalist movement, and many of the leaders had to flee the country. During the Crimean War (1853–56) the Danubian principalities were occupied first by the Russians and then by the Austrians. At the Paris peace conference in 1856, France, Prussia, and Sardinia supported the Romanian cause; England, Russia, Austria, and the Ottomans opposed unification. The Treaty of Paris provided for elections in both principalities to ascertain wishes of the population, and these demonstrated an overwhelming wish for unification. After much deliberation, it was decided in 1858 that the principalities would remain separate, with two princes and two assemblies. In 1859 the Moldavian and Walachian parliaments separately elected a single prince in the person of Alexandru Cuza. After further negotiations the union was formally proclaimed on December 23, 1861. The new state took the name of Romania, with Bucharest as its capital.

The union of Moldavia and Walachia

The young state faced many national, political, and social problems. Many Romanians still lived under foreign rule: in 1867 Transylvania was incorporated in Hungary; Romanians in Bukovina were under Austrian rule; and Bessarabia was part of Russia. Prince Cuza introduced two big reforms: the confiscation of the land of dedicated monasteries (ruled by Greek monks) in 1863, and an agrarian law (1864) that tried unsuccessfully to solve the important peasant problem. During his reign the Code Napoléon was adopted, the judicial system was reorganized, and progress was made in the field of education. Having alienated the boyars and the clergy, Cuza was forced to abdicate in 1866. His successor was Prince Charles of Hohenzollern-Sigmaringen, who took his oath in 1866, became King Carol I in 1881, and ruled until 1914. A new constitution was also adopted in 1866. But the peasant problem remained unsolved.

The emergence of Bulgaria. Bulgarian national emancipation required first a social and cultural emancipation. After 1835 the Bulgarian schools gradually introduced teaching in the native tongue, replacing the previous emphasis on Greek. The first Bulgarian-language journals and periodicals, published abroad, appeared in the 1840s. The movement for a Bulgarian church, which began in the 1820s, met opposition from the Greek patriarchate and from the Ottomans. In 1870 the Sultan established a Bulgarian exarchate with 15 dioceses, and the first exarch was elected in 1872. The patriarch immediately excommunicated him and his followers, declaring them schismatics.

The Bulgarian national movement

Peasant uprisings in the 1830s and 1840s had done much to undermine Ottoman power. In 1850 a peasant revolt in the Vidin district dramatized to Europe the existence of the Bulgarian question. During the Crimean War (1853–56) Bulgarian detachments were formed in the Russian army in the Danubian principalities. An organized revolutionary movement developed, primarily among Bulgarian merchant colonies abroad. A plan for Bulgarian liberation was elaborated by Bulgarians living in Serbia in 1861, and they were allowed to form a Bulgarian military unit in 1862. When the Serbian government hesitated to start a war against the Ottomans, the émigrés moved to Romania, where they obtained the support of Romanian liberals. In 1866 they reached an agreement for a Bulgaro-Romanian coalition. In 1867 the Serbian government and a secret Bulgarian committee in Bucharest signed an agreement for the creation of a common Serbo-Bulgarian state. A number of such secret Bulgarian national revolutionary committees were formed in Romania during the 1860s and 1870s, representing various political views among the émigrés; the movement became more radical and began to put down roots within the country. Among the Bulgarian leaders of the 1860s and 1870s were Georgi Rakovski, who organized national revolution within Bulgaria; Lyuben Karavelov, who fostered the idea of a liberal Balkan federation; Vasil Levsky, who died establishing an internal organization; and Hristo Botev, who linked the national with the social revolution. By the time of the Eastern crisis of 1875, the Bulgarians were prepared for the creation of a state.

International politics. *The first Balkan alliance (1866–68).* During the wars of the 1860s that brought the unification of Italy and of Germany, the main goal of European diplomacy was to preserve the status quo in the Balkans. But the leaders of the emerging states and national societies wanted a general Balkan uprising and a final settlement with the Ottomans. To achieve this, they needed to unify their forces. The Greeks took the initiative in 1860 and tried to reach an agreement with Serbia. After long deliberations a treaty (1867) and a military convention (1868) were signed in which Greece and Serbia agreed on the preparation of a Balkan uprising, a common war against the Ottomans, and the application of national self-determination in the Balkans. Serbia and Montenegro had made an agreement in 1866. Romania joined the group by a treaty of friendship and alliance signed with Serbia in 1868. The Balkan alliance was supplemented by an agreement that the Serbian government reached with Bishop Strossmayer in Croatia (1867) for a common Yugoslav state and, by an accord between Serbia and the Bulgarian committee in Bucharest (1867), for a Serbo-Bulgarian union. The alliance disintegrated after the assassination of Prince Michael of Serbia (1868), but the fundamental idea was to be applied later in the Balkan War of 1912: localization of the war without the involvement of the European powers and the concentration of effective military force against the Ottomans.

Unity against the Ottomans

The Eastern crisis (1875–78). The revolutionary activities of the 1860s exploded in 1875. The uprising in Hercegovina and Bosnia resounded throughout the Balkans, involving the European powers and reopening the Eastern question. In April 1876 an uprising in central

Bulgaria occurred. Serbia and Montenegro declared war on the Ottomans in June 1876, but the Serbian army was defeated. Russia declared war on the Ottomans in April 1877. Overcoming a stubborn Ottoman defense at Petroúsa, the Russians pushed southward until they reached the approaches to Constantinople. Bulgaria was in turmoil, and Serbia reopened hostilities. By the Treaty of San Stefano, imposed on the Ottomans on March 3, 1878, a "great Bulgaria" was created as a Russian satellite, reaching from the shores of the Adriatic to the Aegean and the Black Sea; Serbia, Montenegro, and Romania were to be independent. The Bulgarian arrangement was unacceptable to the European powers, and their reaction led to the Congress of Berlin (June 13–July 13, 1878). Russia was ejected from the Balkans and Austria-Hungary was given a mandate to occupy Bosnia and Hercegovina, thus becoming a Balkan state. Bulgaria was confined to the north of the Balkan range and made an autonomous principality under Turkish suzerainty. Macedonia remained under the Ottomans. Serbia, Montenegro, and Romania obtained recognition of their independence together with territorial changes. None of the Balkan states participated in the Berlin decisions. Greece, through other negotiations, obtained Thessaly and a part of Epirus in 1881. The process of national liberation had been given further impetus, but the Eastern question had not been solved. Instead, the Berlin settlement gave birth to European conflicts that led to World War I.

THE AGE OF IMPERIALISM (1878–1903)

Political and economic influence of the European powers

In the last two decades of the 19th century the Balkans became part of the European political and economic system. Balkan economic, political, and social development was largely a product of European influence. Russian predominance gave way after 1856 to Anglo-Austro-Russian rivalry in the peninsula. European dynasties were introduced in the Balkans: the Bavarian and Danish in Greece, those of Battenberg and Saxe-Coburg in Bulgaria, and the Hohenzollern in Romania. European intervention was a consequence of imperialistic trends in the eastern Mediterranean, the decline of the Ottoman Empire, the opening of the Suez Canal (1869), increased traffic on the Danube, and the extension of railroads. The Balkans gradually became a European crossroads.

A period of economic growth began. Industrialization, initiated by Balkan capitalists, was helped along by European banking and industrial interests through state loans and the opening of railroads and mines. Further social differentiation took place. Though most of the population was still peasant, a middle class was developing along European lines. Modern political parties appeared, and liberals contended with conservatives. The growing strength of the Balkan states found expression in nationalistic policies directed not only against the Ottomans but also against each other.

The Macedonian question. The nature of the area called Macedonia and the ethnic and political character of its inhabitants have been matters of intense dispute in the Balkans. The Greeks have denied the existence of a separate Macedonian nationality, reducing Macedonia to a geographical notion. Bulgarians have claimed the Macedonians as Bulgarians; the Serbs before World War II claimed them as Serbs. After World War II, some Yugoslav historians began to write of a Macedonian nation existing since the Middle Ages and embracing not only the present Yugoslav People's Republic of Macedonia but also parts of Bulgaria and Greece.

Geographically, Macedonia covers the central part of the Balkans. The Vardar valley, with its exit to the Aegean, is the strategic backbone of the peninsula. During the 19th century, as more and more Balkan areas were liberated from the Ottomans, Macedonia was exposed to more intensive Turkish colonization. To the primarily Slavic–Macedonian peasant majority was added a growing urban population in which the Ottoman garrisons and the Greek, Armenian, and Jewish trading middle class prevailed. Economic traffic along the Vardar line and the migration of peasants to the cities led to the gradual formation of a Slav–Macedonian middle class. Educated members of Macedonian society were heavily influenced by Greek culture or, alternatively, by fellow Slavs in Bulgaria and Serbia. There were four large peasant uprisings in the 1880s, influenced by events in Bulgaria, Serbia, and Greece. As Bulgarian activity in Macedonia grew, Greece and Serbia countered with their own national propaganda. During the two last decades of the century Austria encouraged the Serbs in this, while England backed the Greeks. The struggle over Macedonia developed until it was carried on by armed bands.

In the latter part of the 19th century nationalist ideas began to appear. There were demands for the use of the vernacular in schools and for the establishment of a Macedonian church. Revolutionary activities began in the 1880s. The Internal Macedonian Revolutionary Organization (Vatreshna Makedonska Revolutsionna Organizatsia) was created in 1893. Macedonian and Bulgarian wings emerged within it. Its program was "Macedonia for the Macedonians," but as a partner in a future Balkan confederation. Throughout this period Macedonia was torn by conflicts. A bloody uprising in 1903 was repressed by the Turks with heavy reprisals against the noncombatant population.

The Ghegs and Tosks

The Albanian national awakening. The Albanian nation in the Balkans was divided between tribes in the mountainous north (the Ghegs), a feudal society in the southern plains (the Tosks), and a dispersed settlement in the regions of Kosovo, western Macedonia, and Epirus. The north was Roman Catholic, while the south was Muslim and—in Epirus—Greek Orthodox. Albanian emigrants, leaving their poor and overpopulated country, had established national enclaves in other Balkan countries, the eastern Mediterranean, and Italy (in 1901 there were 200,000 Albanians in Italy).

Complex economic, social, and cultural factors governed the Albanian national renaissance. The southern landlords, linked to the Ottomans by common religious and political interests, found in the empire a guarantee of their ruling position and a defense against the growing pressure of the neighbouring Balkan states. The northern tribal society, rejecting the centralistic tendencies of the Ottoman authorities, sought autonomy. Throughout the 19th century the Albanian lands were the scene of local revolts against the Turks. Nationalism, however, flourished mainly among Albanians abroad. Three great Albanian writers promoted Albanian nationalism in Italy: Girolamo De Rada, Demetrio Camarda, and Giuseppe Schiro. The first expression of an organized Albanian national movement in Albania proper was the Albanian League for the Defense of the Rights of the Albanian Nation, established in 1878. Created by Albanians in the Kosovo region as a reaction to the decisions taken by the Congress of Berlin, the League fought against the cession of Albanian territory to Montenegro. It was crushed by the Turks in 1881. The League had nevertheless demonstrated the existence of an Albanian national movement. Further progress was made at the end of the century through the formation of Albanian societies abroad, and the publication by them of Albanian books and periodicals. A national symbol was found in Skanderbeg, the 15th-century fighter against the Turks.

There was no unified Albanian alphabet and writing was in local makeshift Italianate, Hellenic, or Turco-Arabic characters and systems. Political differences prevented the adoption of a single alphabet until 1909, when the Latin was chosen. Turkish schools were established in the 1860s, and subsequently Greek schools in the south; there were no Albanian schools until the 1890s.

After a national uprising in 1912, the European powers accepted the principle of an independent Albania. But no viable national government was formed until after World War I.

Trends among the South Slavs

The Yugoslavs. *Slovenes, Croats, and Serbs under the Habsburgs.* Under the dual Austro-Hungarian monarchy established in 1867, the Yugoslavs had been divided: Slovenia, Istria, and Dalmatia had been assigned to Austria, while Croatia, Slavonia, and the Vojvodina were under the rule of Hungary. The Military Frontier was dissolved and joined to Croatia in 1881.

In Slovenia, where the Austrians imposed a harsh policy of Germanization, the liberals joined with the conservatives in seeking autonomy for the separative Slovenian regions.

Croatia became increasingly linked with the Hungarian economy. A growing nationalism found expression in Ante Starčević's Party of Rights, which demanded the union of all Croatian provinces of the empire, including Dalmatia and Bosnia-Hercegovina. Budapest tried to restrain Croatian opposition by appointing Count Károly Khuen-Héderváry as governor, but his policy of Magyarization only made matters worse. He sought to play off Croats against Serbs in disputes over the future status of Bosnia-Hercegovina. In the 1890s the younger generation began to abandon these futile chauvinistic struggles and to lay the basis for a new course of Yugoslav cooperation.

Bosnia and Hercegovina, under Habsburg rule after 1878, made progress with industrialization and the construction of roads and railways, but the national and agrarian problems remained unsolved. The Habsburg authorities attempted to foster a Bosnian national consciousness. In 1910 it was estimated that 100,000 families still lived under serfdom. A Serb–Muslim uprising in 1882 demonstrated peasant dissatisfaction with the social and political order.

Serbia and Montenegro. The latter part of the century brought economic growth for Serbia, but a decline of its political fortunes at home and abroad. The cattle and grain trade were stimulated by commercial treaties with Austria. A national bank was established in 1883. Government budgets rose from 19,500,000 dinars in 1880 to 76,000,000 in 1900. A Danubian shipping company was started and Serbia was linked with Austria and the Ottoman Empire by rail. European financial capital penetrated the country. Light industry began—breweries, mills, slaughterhouses, sugar refineries, textiles mills.

Serbia was closely tied to Austria-Hungary in 1881 through a secret convention (not made public until 1893) that sanctioned Serbian expansion toward the south. Domestic politics in the 1880s was divided among conservatives, liberals, and a middle class left. The latter succeeded in involving large peasant masses in politics. Serbia was proclaimed a kingdom in 1882. Supported by a conservative minority, King Milan Obrenović IV led an unsuccessful war against Bulgaria's union with Eastern Rumelia in 1885. Defeated by the Bulgarians, King Milan lost his popularity and in 1889 was forced to abdicate in favour of his son Alexander. A liberal Swiss-type constitution was introduced in 1888, but Serbian political life was tumultuous. The behaviour of Alexander and his father Milan, who resumed a personal role in politics, destroyed the prestige of the dynasty.

Montenegro became independent in 1878 through the Congress of Berlin, and its territories were doubled in size. There was an inflow of Italian capital. A modern government was introduced in 1879, with five departments, and the civil law was codified in 1888.

Bulgaria. During the 1877–78 war an important change in social structure occurred in Bulgaria: Ottoman landlords, followed by a large number of Turkish peasants, left the country. The Bulgarian peasantry seized their lands (approximately a quarter of the arable land) and established small and medium-sized farms. This social structure strongly influenced Bulgarian economic and political development.

Bulgaria entered a phase of state-sponsored urbanization and industrialization. At the end of the century there were 1,129 kilometres (700 miles) of railway lines. Foreign capital was imported, mostly through state loans; Bulgaria's foreign indebtedness at the end of the century reached 250,000,000 francs.

Bulgaria's political instability

A liberal constitution was adopted in 1879, and the national assembly elected Alexander of Battenberg as prince of Bulgaria. In Eastern Rumelia, separated from Bulgaria in 1878, the unionist movement, supported by the entire population and all the political parties, organized an army coup and a popular uprising that resulted in the proclamation of the union of Eastern Rumelia with Bulgaria in September 1885. Serbia's attempt forcibly to prevent the unification ended in military defeat.

A number of political parties appeared after 1885, resulting from a split among the Liberals. The role of the army and of the Macedonian émigrés grew. A conflict with Russia deprived Battenberg of Russian support, and a pro-Russian army conspiracy forced him to abdicate in 1886. After a period of crisis, the national assembly elected Prince Ferdinand of Saxe-Coburg-Gotha as his successor in June 1887. The government formed by Stefan Stambolov (1887–94) took a pro-Austrian attitude; in the face of Russian hostility Stambolov had to rule almost as a dictator. The Bulgarian peasantry entered politics at the end of the century, mobilized by the Bulgarian Agrarian Union. But, as in Serbia, political life was unstable at the end of the 19th century: there were plots against the government, political assassinations, a growing underground activity by Macedonian émigrés, and several peasant uprisings.

Greece. The period after 1878 was one of steady economic growth in Greece: the Corinth Canal was completed in 1893; a number of paved highways were begun in the 1880s; and by 1905 Greece had 1,000 kilometres (620 miles) of railroads. Piraeus was becoming the Mediterranean's fourth-largest port. Shipping and light industry flourished, and foreign investment grew. On the other hand there was poverty in the countryside, resulting from overpopulation and from the French tariff on currants; this induced large numbers of Greeks to emigrate in the 1890s.

In foreign affairs the Greeks sought to incorporate Macedonia to the north and the island of Crete to the south. Many Cretans had long nourished the idea of *énosis* (union) of the island with Greece. A Cretan uprising against the Turks in 1866 had been put down. A new uprising in 1896 led to a proclamation of union with Greece in February 1897. Supporting the Cretans, the Greek government sent an armed force to annex the island, but intervention by the European powers followed. In April 1897 Greek troops attacked the Turks in Thessaly and suffered a defeat. The peace settlement required Greece to pay a war indemnity that led to bankruptcy. Crete, however, received its autonomy, and Prince George, the second son of the Greek king, was named commissioner of the island. At the turn of the century Greek politics was characterized by frequent cabinet changes and petty party strife.

Romania. The Romanian economy also made progress after 1878. An oil industry was developed at Ploeşti with financing by German, British, and Dutch capital. By 1906 there were 2,000 kilometres (1,240 miles) of railroads. Romanian trade was closely bound up with Austria-Hungary: 50 percent of Romanian imports were from the Habsburg Empire, and 32 percent of its exports went there in the period 1875–82. The Austro-Romanian tariff war (1886–1893) emancipated Romanian trade. An export surplus enabled the government to borrow abroad to the extent of 1,700,000,000 francs by 1914. Romanian foreign policy turned toward Austria after 1878, and in 1883 the government signed a secret alliance with Austria, Hungary, and Germany (renewed in 1896, 1902, and 1913).

Romania's ties with Austria-Hungary

The central Romanian social and political problem was the peasantry. Although Bucharest was known as "the Paris of the Balkans," the peasantry lived under appalling conditions. About 85 percent of the peasants either had no land or were forced to work part of the time for landlords. Five thousand large estates covered about half of the total arable land.

Romanian politics was divided between the Conservative and Liberal parties. The leader of the Liberals, Ion Brătianu, was until his retirement in 1888 the most influential figure in Romanian politics. His three sons all became leaders of the Liberals. The Romanians had irredentist ambitions in Transylvania, where large numbers of Romanians lived under Hungarian rule. Romania also took part in the Macedonian controversy by claiming the rights of a protector over the Vlachs.

THE BALKANS BEFORE WORLD WAR I (1903–14)

Imperialism, nationalism, and Balkan unity

At the beginning of the century the Balkans were increasingly the scene of international conflict. The European powers had clashing military, political, and economic interests in the peninsula. The Baghdad railway project at the turn of the century symbolized German ambitions to push eastward, challenging French financial dominance in the Ottoman Empire. The growing weakness of Russia upset the Austro-Russian balance in the Balkans, while growing Italian strength in the Adriatic stimulated efforts by Austria-Hungary to reach the port of Thessaloníki and the Aegean.

Among the Balkan countries themselves, developing national strength resulted in a general movement toward political and economic emancipation. In the Yugoslav lands the year 1903 marked a turning point. Croatian political forces began to organize against Hungarian repression, and a similar restlessness was apparent in Bosnia. The assassination of King Alexander Obrenović in Serbia prepared the way for a dynamic, nationalistic foreign policy (under the rule of Peter I Karageorgević). In Macedonia in 1903, the Ilinden uprising displayed to Europe the weakness of Turkey, the "sick man on the Bosphorus." In Croatia and Dalmatia, a coalition of political parties put forward a program calling for Yugoslav unity and social and political reform. The movement received support from Serbia and Montenegro and began to gather impetus. The Balkan states started to arm, buying artillery and ammunition in Europe.

Two tendencies were at work in the Balkans: rivalries among the states for territorial aggrandizement, and a common hostility toward interference from outside powers. A growing inter-Balkan struggle next occurred over Macedonia and was expressed in the activities of armed bands from neighbouring states that brought much suffering to the local populations. The fear of European intervention and the needs of common defense imposed a political rapprochement on Serbia and Bulgaria in 1904, joined later by Montenegro. A peasant revolt in Romania in 1907, put down by the military at the cost of thousands of lives, demonstrated that Romania's agrarian problem was far from solved.

The crisis of 1908–09. A diplomatic struggle among the European powers began in 1908 over railway projects in the Balkans. These projects expressed the political tendencies of the states involved: Austria's push toward the Aegean (the Novi-Pazar railway project), Russia's and Serbia's toward the Adriatic (the Danube–Adriatic project), Italy's penetration of southern Albania (the Vlöre–Munushtir railway), and the effort of Greece and Bulgaria to absorb central Macedonia. In July 1908 the Ottoman garrison in Thessaloníki rebelled, as a result of the revolutionary activity of the Young Turks. The officers (among them Mustafa Kemal, the future leader of the Turkish republic), forced the Sultan to proclaim a constitutional era (see also ATATURK, KEMAL). The Young Turks wanted a regime that would give liberty and equality to all the nationalities within the empire. On October 5 the Bulgarians took advantage of the confusion to proclaim their full independence. On October 6 Austria-Hungary announced its annexation of Bosnia-Hercegovina. Two days later the Cretans proclaimed their union with Greece. The system created by the Congress of Berlin in 1878 had collapsed. The major crisis was over Bosnia-Hercegovina: the Serbs protested vehemently against the annexation; England opposed Austria; Russia backed Serbia and Bulgaria. The Austro-Turkish conflict was settled by an indemnity paid by Austria to the Turks, but the Austro-Serbian conflict brought Europe to the edge of war. Pressure from Berlin in March 1909 forced Russia to yield. Serbia had to follow the Russian example and accept Bosnia's annexation by Austria-Hungary. The crisis of 1908–9 foreshadowed coming events.

Revolt of the Young Turks

The Balkan Wars of 1912–13. The situation in the Balkans remained uneasy. The Albanians, after a short-lived collaboration with the Young Turks, revolted against them in 1909–12. The Cretan *énosis* with Greece failed because of European opposition. This caused general indignation in Greece, opening the way to a military coup and the premiership of the Cretan political leader Eleuthérios Venizélos (*q.v.*). The Venizélos government introduced reforms, reorganized the army, and revised the constitution. Macedonia continued to be a magnet for the nationalist ambitions of the Serbs, Bulgars, and Greeks. The Young Turks banned political parties and national organizations in Macedonia in 1909, and, consequently, the armed bands of various nationalities reopened their guerrilla warfare. An opportunity for an effective cooperation against Turkey presented itself in September 1911, when war broke out between Turkey and Italy in North Africa.

The Balkan alliance against the Ottomans

War against the Ottomans required a Balkan alliance. Serbo-Bulgarian negotiations started in the fall of 1911 but were troubled by differences in regard to the future delimitation of Macedonia. Finally an agreement was reached under Russian auspices in March 1912, providing for the division of Macedonia. A Greek–Bulgarian agreement was reached in May 1912, without touching the delimitation problem. Montenegro then joined the alliance, which disposed of armed forces totalling 750,000 men. The war began in October 1912. The Balkan allies were soon victorious: the Serbs defeated the Ottomans at Kumanovo, joined forces with the Montenegrins to enter Skopje, achieved another victory at Munushtir, and reached the Adriatic at Durrës. The Bulgarians defeated the main Ottoman forces at Kirklareli and Lüleburgaz, advancing to the Çatalca lines in front of Constantinople. The Greeks seized Thessaloníki and laid siege to Ioánnina. The Ottomans lost all their territories in Europe, except a small strip around Constantinople. On December 3 an armistice was concluded. Peace negotiations opened in London on December 16.

The victories of the Balkan allies affected many Austo-Hungarian interests. Vienna reacted vehemently, demanding the withdrawal of Serbian troops from the Albanian coast where an Albanian state had been proclaimed on November 28, 1912. The Austro-Serbian conflict automatically developed into an Austro-Russian one.

The crisis was settled at a conference of ambassadors in London in December 1912 that recognized the new Albanian state and obliged Serbia to withdraw its troops from the Adriatic. But at the end of January, 1913, after a coup d'etat in Constantinople by the nationalistic Young Turks, the war with the Ottomans was resumed. The allies were again victorious: Ioánnina fell to the Greeks, and Adrianople to the Bulgarians. Another crisis arose when the Montenegrins refused to leave Shkodër, which the London ambassadors' conference had given to Albania. The Montenegrins were forced to yield under the threat of a European naval blockade of their coast. A peace treaty, signed in London on May 30, 1913, gave all the territory west of the Enez–Midye line to the Balkan allies. Crete was united with Greece.

Serbia and Greece against Bulgaria

The territorial settlement produced discords among the Balkan allies. Serbia refused to give up the parts of Macedonia assigned by the 1912 treaty to Bulgaria on the ground that it had already been forced to withdraw from the Adriatic. A bitter struggle then ensued between the Greeks and the Bulgarians over Thessaloníki and Thrace.

Romania, as a price for remaining neutral, demanded from the Bulgarians a part of the Dobruja. Both Serbia and Bulgaria were reluctant to accept Russian arbitration. Austro-Hungarian diplomacy sought to undermine the Balkan alliance. Serbia and Greece allied themselves against Bulgaria on June 1, 1913. King Ferdinand of Bulgaria, backed by army circles, ordered his troops to attack Serbia and Greece in Macedonia on June 30. But the Bulgarian armies were defeated by the Serbs and Greeks. At the same time the Romanians entered the Dobruja and the Ottomans recaptured Adrianople. An armistice was concluded on July 31 and a peace treaty signed in Bucharest on August 10, 1913.

Results of the wars. As a result of the Balkan Wars, Greece obtained Thessaloníki, Kaválla, and a large coastal part of Macedonia; Serbia gained the northern

and central part of Macedonia; Montenegro acquired a portion of the sanjak of Novi-Pazar, establishing a common frontier with Serbia; Bulgaria retained a part of eastern Macedonia; and Romania procured its part of Dobruja. The long decline of Ottoman rule in the Balkans had ended.

The political consequences of the Balkan Wars were considerable. Apart from Turkey, the real loser was Austria-Hungary. The partitioning of the sanjak of Novi-Pazar between Serbia and Montenegro made it impossible, in the subsequent crisis of June–July 1914, for Austria-Hungary to intervene in the Balkans by occupying the sanjak. The success of Serbia and Montenegro stimulated the Yugoslav movement for union in the Habsburg Empire. The wars similarly altered the structure of alliances in the Balkans. Dissatisfied, Bulgaria henceforth looked to Austria-Hungary for support, while Romania tended to move away from its allies and toward Russia.

The most alarming result was the growth of tension between Austria-Hungary and Serbia. Serbia had extensive claims upon Albanian territory. Having obtained an assurance of German support, Austria-Hungary delivered an ultimatum in October 1913 to compel Serbia to withdraw from the Albanian borderlands. This, however, did not solve the Southern Slav question for Austria-Hungary, and it emerged once again in an acute form with the assassination of the archduke Francis Ferdinand on June 28, 1914, in Sarajevo, Bosnia. This event was followed by Austria-Hungary's ultimatum and by its declaration of war on Serbia (July 28, 1914) and the outbreak of general war in Europe in August (see WORLD WARS: *Origins*). (D.V.D.)

III. The Balkans after 1914

WORLD WAR I AND THE PEACE SETTLEMENTS, 1914–23

The role of the Balkan states in the war. The World War of 1914–18 was triggered by Balkan revolutionary nationalism. Gavrilo Princip, who shot and killed Archduke Ferdinand, heir apparent to the Habsburg throne, was one of many young Serbs who pinned so much faith upon the advantages of national unification as to risk their lives and flout existing political authorities, both in Serbia itself and in the adjacent lands of Bosnia, Hercegovina, Dalmatia, and southern Hungary, where Serbian populations lived under Habsburg rule.

Breakup of traditional life-styles

Nationalist movements. The breakup of traditional peasant styles of life among the South Slav peoples—Slovenians, Croats, Serbs, and Macedonians—fuelled this revolutionary movement. Population growth made subdivision of peasant holdings necessary; but, when a father had to divide his land among several sons, the new families could not hope to live as their parents had. This pushed innumerable young men off the land and into revolutionary activities. The pattern was as follows: as life on the farm became impossible, ambitious persons tried to get enough formal education to qualify for a desk job in town. But even after several years of secondary schooling, desk jobs were hard to come by, and many young men who had gone that far by desperate effort were not willing to wait patiently until something turned up. Instead, they listened eagerly to those who preached extreme revolutionary action.

According to Serbian nationalist agitators, justice, freedom, and a decent regard for Serbian dignity required that all speakers of the South Slav tongue who were also of the Orthodox faith should belong to the same state. The fact that the end of Habsburg rule over Serbian populations would inevitably mean more government jobs for educated Serbs was an attractive additional advantage.

Some thought that Muslim and even Roman Catholic speakers of the South Slav language should join their Orthodox brethren in a new South Slav state. But this Yugoslav (*Yug*, "South") ideal had little appeal for most Serbs. It attracted support mainly in Dalmatia, where rural Serbs (Orthodox South Slavs) and Croats (Roman Catholic South Slavs) found it easy to cooperate against the Italians, who had long dominated town life along the Adriatic.

Despite Serbia's tiny size compared to the vastness of the Habsburg monarchy, many high Austrian officials concluded that intransigent Serbian nationalism constituted a serious threat to their state. The power of the nationalist ideal in the Balkans had been demonstrated in 1912 by the First Balkan War, which all but drove the Ottoman Empire from Europe and added substantial new territories to Serbia. This success fanned the Serbian nationalist ambition which was already at white heat, to complete the task of liberation by disrupting the Habsburg state.

What made such a program so frightening to the Austrians was that other nationalities within the Habsburg Empire shared in some degree the ambition to win greater control over their own affairs; a few even dreamed of achieving complete national independence. But in 1914 nationalist movements among the Czechs, Germans, Italians, Poles, Magyars, Slovenes, and Croats were less emotionally intense and, therefore, less immediately threatening to constituted authority than was the case among the Serbs.

Austrian reaction to the assassination of Archduke Ferdinand

Outbreak of war. Austrian officials, therefore, felt that a showdown was desirable. They decided to seize upon Archduke Ferdinand's assassination to settle accounts with Serbia. Preliminary investigation failed to turn up definite evidence that the Serbian government had been connected with the assassination; but historians have since shown that the Serbian premier, Nikola Pašić (*q.v.*) knew what was going on and tried indirectly to warn Austrian authorities of the plot. Col. Dragutin Dimitrijević, chief of intelligence for the Serbian general staff and leader of the Black Hand—a secret society that planned the assassination and armed Princip (with several others)—was one of Pašić's political enemies and rivals. In hinting to the Austrians of what was afoot, Pašić did as much as he dared to thwart Dimitrijević's risky, revolutionary plans. But the Austrian authorities failed to take the hint. The Archduke was assassinated, official Europe was horrified, and on July 23, 1914, the Habsburg government delivered an ultimatum to the Serbs requiring suppression of patriotic societies and the establishment of a joint commission to investigate and punish those persons who were responsible for organizing the assassination.

The ultimatum was designed to be unacceptable. Despite a conciliatory reply, the Austrians declared war against Serbia on July 28, exactly a month after the assassination itself. Within a week Europe's alliance system swung creakily into action, pitting the Triple Entente—Russia, France, and Great Britain—on Serbia's side against the Central Powers—Germany and Austria-Hungary.

Russia's entry into the war upset Austrian plans for crushing Serbia. Austrian troops had to be diverted to the Russian front, and, when the Austrians were finally ready to attack the Serbs, on August 13, they were repulsed. A Serbian counteroffensive soon petered out, however, and by the end of 1914 the battle line stood very close to the prewar frontiers.

The outbreak of hostilities provoked intense diplomatic–military activity elsewhere in the Balkans. On August 11, 1914, two German cruisers that had been trapped in the Mediterranean Sea at the beginning of the war arrived in Constantinople. A fictitious sale transferred them to the Turks, and in October these vessels sailed into the Black Sea to bombard Russian coastal towns. The Allies (Entente) promptly declared war against Turkey (November 4–5, 1914).

Both the Entente and the Central Powers sought to rally support for their cause among the three remaining uncommitted Balkan states—Bulgaria, Romania, and Greece (Albania was in chaos and lacked a central government). Bulgaria wanted the parts of Thrace and Macedonia that had been annexed by the Serbs and Greeks after the Second Balkan War in 1913; Romania wanted the Habsburg territories of Transylvania and ad-

jacent regions in Bukovina and the Banat; Greece had ambitions in Anatolia and, above all, desired to possess Constantinople (later Istanbul).

The Central Powers were able to promise the Bulgars most of what they desired, because Serbia would be the main loser. Similarly, the Entente was able to appease the Romanian appetite at the expense of the Habsburgs. Greece's ambitions were more difficult, however, because Russia, too, aspired to possess Constantinople, and so did the Bulgars.

Courtship of the Balkan states

The work of diplomats and intelligence agents in lining up allies in the Balkans for one side and the other came slowly to fruition in 1915–17. The Bulgars were the first to commit themselves, by allying with the Central Powers in September 1915. By then the strategic situation in the Balkans had altered greatly. The first important move was British. In February 1915 the Royal Navy bombarded Turkish forts in the Dardanelles, and a month later British warships tried unsuccessfully to force their way through the straits. Then, on April 25, 1915, British troops went ashore on the Aegean side of the Gallipoli Peninsula in order to take the Turkish defenses of the Dardanelles in the rear and open a way to Constantinople for the Royal Navy; but once again the Turks were able to stop the British advance before it could achieve strategic success.

The next major move came from Italy. Despite a defensive alliance with Germany and Austria-Hungary, the Italian government remained neutral in 1914, arguing that, because Austria had attacked Serbia, the terms of the alliance had not been fulfilled. Italian nationalists wished to add Dalmatia and other Austrian provinces to their country; despite Serbian objections, the Allies agreed to most of the Italian demands in the secret Treaty of London (signed in April 1915). Accordingly, in May the Italians declared war and opened a new front against Austria in the Alps. Italian troops also crossed the Adriatic to occupy Albania.

The intervention of Italy, plus Austrian commitments on the Russian front, forced the Austrians to postpone large-scale action against the Serbs until fall; but by the first week in October everything was ready for a major assault. The Bulgars saw in this their opportunity to take Macedonia from Serbia, and they prepared to join the attack.

In a last-minute effort to aid the Serbs, French and British troops landed at the Greek city of Thessaloníki on October 3. This touched off a major crisis inside Greece, pitting Prime Minister Eleutherios Venizelos, champion of the Allies, against King Constantine, who favoured continued neutrality, at least until it was clearer which side would win the war. (In the end, by bombarding the royal palace in Athens the Allied powers were able to compel Constantine to abdicate in June 1917; and Venizelos, who took command of Greek affairs, promptly caused Greece to declare war against the Central Powers.)

A combined Austrian and German force attacked Serbia on October 6, 1915, with overwhelming strength. Allied troops in Salonika were too few to check the simultaneous Bulgar advance from the east. In November, the Serbian Army began a painful retreat through the mountain passes of Albania. A remnant 125,000 strong found refuge on the Greek island of Corfu, where the Serbian government, still being led by the aged and ailing King Peter I, with Nikola Pašić as prime minister, established a temporary headquarters (January 1916) in what had formerly been the German kaiser's holiday palace.

Meanwhile, the British withdrew their troops from Gallipoli, transferring most of them to the Thessaloníki front, which soon extended westward to link up with the Italians in Albania. Trench warfare then set in along a battle line extending all the way from the Adriatic coast to Kaválla on the Aegean. Neither side could break through the other's prepared defenses.

Stalemate of 1916

This stalemate was demonstrated in 1916, when the last Balkan neutral, Romania, entered the war on the Allied side on August 27. This move was planned to coincide both with a Russian offensive against Austria and with a major push against Bulgaria along the Thessaloníki front. Despite substantial reinforcement by rested and re-equipped Serbian troops transferred to Thessaloníki from Corfu, the offensive failed. The Romanian Army proved ineffective, and by January 1917 most of Romania was in the hands of the Central Powers.

The year 1917 saw no large-scale military action in the Balkans, but revolutions in Russia changed the realities of power profoundly. The Bolsheviks publicly renounced Russian claims on Constantinople; instead, Lenin appealed over the heads of all governments to the peoples of Europe to rise against their exploiters and inaugurate Socialism. Balkan response to Lenin's revolutionary summons was slight, partly because the Western Allies forestalled the Bolshevik appeal by endorsing a different revolutionary formula—national self-determination. The war thus became far more ideological than before, pitting nationalist revolutionary ideals against Socialist revolutionary ideals through the whole of eastern Europe.

The most complex nationality issue in the Balkans centred around Serbia's relationship to the other South Slav peoples of the Habsburg Empire. On July 20, 1917, the Serb prime minister Pašić signed a document declaring that the Serbs, Croats, and Slovenes should form a single state after the war, under the Serbian Karageorgević dynasty but with appropriate local autonomies. Exiled Dalmatian politicians, claiming to represent the Croats and Slovenes, also signed this Pact of Corfu. In April 1918 a Congress of Oppressed Nationalities met in Rome and, with Italian blessing, reiterated the idea that Serbs, Croats, and Slovenes belonged together, despite the fact that the three nationalities distrusted one another profoundly and that most Croats remained loyal to the Habsburg cause.

End of the war. The long military stalemate in the Balkans ended on September 30, 1918, when the Bulgars decided to sue for an armistice. From the Thessaloníki front, Allied troops under French command marched northward to the Danube. As the Allied forces approached, the Romanians, who had signed a peace with the Central Powers in May 1918, re-entered the war on November 8, just in time to count as an Allied and victorious power at the peace conference.

The Ottoman government followed the example of the Bulgarians by suing for an armistice on October 30, 1918. The armistice terms allowed a British force to advance through the Dardanelles and to occupy Constantinople on November 13. Amid the crash of falling empires, the Habsburg monarch also signed an armistice on November 3; but events had stripped the Emperor of his power to influence affairs. The Romanians (already in possession of the former tsarist province of Bessarabia) hastened to take possession of as much of Transylvania, Banat, and Bukovina as possible; meanwhile, Serbian troops were moving into the former Habsburg territory from the south, while the victorious Italians also sought to get control of as much of the Adriatic coastlands as they could.

Under these circumstances the Slovenes and Croats had little room for manoeuvre. On October 29, 1918, a national council meeting in Zagreb proclaimed the independence of "Yugoslavia" (meaning the former Habsburg lands in which the South Slavs lived). Before a stable settlement with the Serbs could be achieved—a preliminary agreement reached at Geneva on November 9 was later repudiated—the Zagreb National Council broke apart.

A faction opted for immediate union with Serbia; and, accordingly, on December 1, 1918, the Serbian monarch formally proclaimed a new Kingdom of Serbs, Croats, and Slovenes from his capital, Belgrade. A political coup unseated the Prince of Montenegro, whose state was also merged with the new South Slav state. But whether the new state would be federal or unitary and the question of how Serbs and Croats would come to an understanding with one another remained completely unset-

Kingdom of Serbs, Croats, and Slovenes

tled. (For additional information see WORLD WARS: *World War I.*)

Results of the peace conferences. The Paris Peace Conference that began in January 1919 drew up separate treaties for each defeated enemy state. By the Treaty of St. Germain (signed September 10, 1919) Austria ceded Slovenian and Dalmatian territory to the new Kingdom of Serbs, Croats, and Slovenes; by the Treaty of Neuilly (signed November 27, 1919) Bulgaria lost a strip of the Aegean coast (acquired in 1912) to Greece and surrendered small border territories to Serbia; by the Treaty of Trianon (signed June 4, 1920) Hungary transferred Transylvania and part of the Banat to Romania and surrendered Croatia, Slavonia, and the rest of the Banat to the new South Slav state; lastly, by the Treaty of Sèvres (signed August 10, 1920), Turkey assigned most of Thrace as well as the hinterland of Smyrna (Izmir) in Asia Minor, to Greece.

From L. Stavrianos, *The Balkans Since 1453;* Holt, Rinehart and Winston, Inc.



The Balkans after World War I.

These treaties settled many of the territorial questions that had long distracted Balkan politics and in a comparatively enduring way. But "national self-determination" proved a difficult formula to apply in lands where mixtures of nationalities were the rule rather than the exception. Some issues, such as the delineation of the Italian–Yugoslav boundary, were never resolved at the peace conference; instead, bilateral negotiation eventually (1924) defined a frontier that satisfied neither side.

Albania's borders were not fixed until 1926, when an international commission, established in 1912, finally concluded its labours. This settlement, which left large Albanian populations inside the new South Slav state, made little difference politically until after World War II, because Albanian national self-consciousness was weakened by traditional loyalties to rival kindred groupings, on the one hand, and by religious (Muslim, Orthodox, Roman Catholic) and linguistic (Gheg, Tosk) differences on the other.

Turkish and Tartar minorities along the Black Sea coast in Bulgaria and Romania shared the Albanians' prepolitical status; and Jews, important mainly in Romania and at Thessaloníki offered no systematic resistance to the new masters of the Balkans. The case was far different, however, with other Balkan nationalities. Those who had formerly enjoyed a leading position in society and government found it all but impossible to accept willingly the loss of former privileges. The victors often retaliated bv subjecting German, Magyar, Turkish, and (in Dalmatia) Italian minorities to flagrant administrative discrimination.

Hence, the territorial settlement of 1919–20 had the effect of transferring to Romania and the new Kingdom of Serbs, Croats, and Slovenes many of the nationality problems that had plagued the Habsburg and Ottoman empires before 1914. In Romania, Magyar, German, Jewish, Ukrainian, Bulgarian, and Turko-Tartar minorities amounted to at least 4,500,000 in a total population of about 18,000,000; but the Romanian majority did give a solid core to the new state. The situation in the Kingdom of Serbs, Croats, and Slovenes was less stable, for, although the Serbs constituted the largest single nationality, they were still a minority in the state considered as a whole.

Continuing nationality problems

German, Magyar, Albanian, and Romanian minorities totalled over 1,600,000; about 1,300,000 Muslims (mostly speaking Serbo-Croatian) constituted another distinct bloc; Slovenes (about 1,100,000) formed a self-conscious, distinct nationality, too; but the really critical matter was the relation between the Croats (about 3,500,000) and the Serbs (about 5,500,000). A new constitution went into effect on January 1, 1921, establishing a unitary state on democratic lines. Croats felt that the Corfu Declaration of 1917 had committed the Serbs to federalism, and they refused to accept the new arrangement. This was especially dangerous for the new state, because the Italian government was dissatisfied with the peace settlement in the Adriatic and set out actively to encourage disruptive forces. The death of King Peter in 1921 made small difference; his heir, King Alexander I (reigned 1921–34), had already exercised the royal powers for several years.

The southern Balkans were even more distracted in the first postwar years. A Turkish nationalist movement, headed by Kemal Atatürk, refused to accept the terms of the Treaty of Sèvres. The Greeks, seeking to make good their claim to even more extensive territories in Asia Minor, invaded the interior in hope of forcing the Turks to yield; in 1921 they met defeat, and the angry Turks forced all Greeks and other Christian inhabitants from the land. About 1,500,000 survivors fled across the Aegean in 1921–22.

A new peace, agreed to at Lausanne, Switzerland, in 1923, provided for the systematic exchange of populations between Greece and Turkey under League of Nations supervision. Greek and other Christian inhabitants of Constantinople were exempted from this exchange; in return, Greece promised to allow Turkish peasants in western Thrace to remain on their land. This treaty provided that the Greeks relinquish their claim to Asia Minor entirely and retroceded eastern Thrace to Turkey.

Exchange of population with Bulgaria was also arranged by a separate agreement. The result, by about 1927, when major transfers ceased, was the sorting out of the populations of Greece, Bulgaria, and the western parts of Turkey by nationality.

No politically important national minorities remained in the southern and eastern Balkans. This radical surgery allowed nationality frictions to subside slowly in later decades. In the northern Balkans, however, such frictions continued to constitute a major axis of politics during the interwar years.

INTERWAR DEVELOPMENTS, 1923–39

The upheavals of World War I did little to solve the underlying problems of Balkan society. It is not strange, therefore, that revolutionary discontent found new chan-

nels of expression after the war. Success for certain nationalities (Serbs, Romanians) automatically meant frustration for others (Bulgars, Magyars, Croats, Macedonians), so that in some regions of the peninsula old-fashioned nationalist conspiracy and agitation continued as before, but now directed against the new masters of the land.

But national self-determination lost much of its glamour as it became clear that old problems were not really relieved by the changes in political boundaries and shifts in dominating nationalities that had occurred in 1918–19. Two new revolutionary movements, therefore, surged to the fore: peasantism and Communism.

New revolutionary movements. The three main bearers of the peasantist idea were the Peasant Party of Bulgaria, led by Aleksandŭr Stamboliyski; the Croatian Peasant Party, led by Stjepan Radić; and the Romanian Peasant Party, led by Iuliu Maniu. The latter, based mainly in Transylvania, opted for parliamentary and peaceable agitation and, even when Maniu briefly became prime minister (served 1928–30), accomplished little to reform rural conditions. The Croatian Peasant Party quickly became identified with Croat nationalism, thanks to Radić's unbending opposition to Serbian preponderance in the new Kingdom of Serbs, Croats, and Slovenes; by boycotting the parliament, Radić showed how shaky the new state really was, but his policy, too, accomplished no positive ends. Stamboliyski, on the other hand, came to power in 1919 on a wave of revolutionary feeling; but his efforts to end bureaucratic oppression and overthrow the parasitic classes that fattened on peasant labour only succeeded in putting crude former peasants and party men into administrative roles they were poorly equipped to handle; in 1923 a coup d'etat led to Stamboliyski's assassination and to the establishment of a shaky parliamentary regime, closely controlled from behind the scenes by royal, military, and semi-military manipulators.

Peasantism

The basic reason for the failure of peasant parties to achieve their ends lay in their programs. Men who wished to destroy the so-called social parasites (*i.e.*, everyone who did not work with his hands and raise his own food) could not take power without themselves becoming that which they wished to abolish: bureaucrats and paper shufflers.

Opposition to those who ruled was the only role such parties and movements could accept comfortably. Only in this way, in fact, could they hope to remain faithful to the perennial distrust their peasant constituencies felt toward government in any form and toward city people generally.

The other new revolutionary movement, Communism, was better equipped ideologically and organizationally. Communist parties had arisen in each Balkan state by 1921. In Bulgaria and Yugoslavia, the new parties met with rapid initial success but subsided into small, quarrelsome groups of hunted revolutionaries when the two governments officially outlawed Communist agitation, in 1921 (Yugoslavia) and 1923 (Bulgaria). In Romania, Greece, and Albania, Communist organizers met with only slight response in the 1920s, but they did succeed in creating revolutionary cadres ready and willing to operate outside the law.

Rise of Communist parties

The frustration of peasantist aspiration and the prevalance of nationalist revolutionary sentiment among such groups as the Macedonians seemed in 1924 to offer Communists a chance to unite all the disaffected elements of Balkan society into a grand alliance. The Comintern (Communist International, formed in 1919 to help in spreading Leninism around the world) approved the formula of Balkan federation as a solution for the peninsula's political and economic ills and instructed each national party to form "popular fronts" with any and all available groups. Radić flirted openly with Moscow; so did leaders of the Internal Macedonian Revolutionary Organization (IMRO).

But these incompatible bedfellows soon parted. IMRO leaders who cooperated with the Communists were killed by rivals within the organization. Radić became a cabinet minister for a brief period (1925–26), but in 1928, during a session of the parliament, he was shot and killed by a Montenegrin Serb. This assassination provoked a strong reaction among the Croats against any kind of cooperation with the Serbs. King Alexander, therefore, decided to scrap the constitution, which had done so little to heal the fissures within his kingdom; he proclaimed a dictatorship, dissolved the political parties, and officially renamed the state Yugoslavia.

Government reactions. By resorting to authoritarian government, Alexander openly admitted the breakdown of effective government by consent. The same thing had also happened in Bulgaria, Romania, and Albania; but in those countries the pretense of parliamentary elections and the rituals of party coalitions were preserved, and they sometimes registered real adjustments in public mood.

This was the case, for instance, when Maniu forged a coalition of peasant parties in Romania and emerged as premier in 1928. In Greece, too, the parliamentary elections that returned Venizelos to power in 1928 registered public feeling quite accurately; but Greek electoral politics were frequently punctuated by coups d'etat (1922, 1925, 1933, 1935), that were sometimes successful, sometimes not.

Official efforts to cope with the problems of Balkan society and to meet the new revolutionary thrusts directed against existing governments were not entirely fruitless. They took three forms: land reform, industrialization, and police repression.

Efforts to solve social problems

Land reform. Widespread redistribution of land shifted ownership toward the peasant families who actually worked the soil. Such reforms went fastest and farthest when land could be taken from owners of a different nationality. Thus, Magyar estates in the north and Turkish estates in the south disappeared at once. Gradual reapportionment took place elsewhere, too, so that by 1939 large estates had almost disappeared from the Balkans, surviving only in those parts of Romania where the landlords were politically powerful Romanian nationals.

Industrialization. The new owners of small peasant plots were often unable to cultivate the soil as efficiently as had the large-scale operators. Redistribution of land was no solution to the ills of Balkan overpopulation and underdevelopment. Each of the Balkan states recognized this fact by attempting to forward industrialization. Romania, building upon an oil industry that had arisen in the 19th century, made by far the most substantial progress in this direction; Bulgaria made almost none. Tariff protection, subsidy, and state enterprise were the devices used to develop industry. But shortage of capital and a generally low level of skills made progress painfully slow.

The only important policy difference in the interwar period with respect to industrialization was about how to treat foreign capital. Romania tried to finance industrial development from internal sources and actively discouraged further foreign investment in the oil industry, fearing that control and the real benefits of such expansion would pass exclusively to the foreign capitalists. Albania, under Ahmed Zogu (president, 1925–28; king, 1928–39), on the other hand, depended wholly on Italian capital for whatever modernization was achieved. Greece and Yugoslavia gave an ambivalent welcome to foreign investment, but political instability in these countries kept the flow of foreign capital to modest proportions. Bulgaria was more xenophobic and attracted almost no foreign funds.

Repression. The third major official response to the problems of Balkan society was repression. High-handed police action was common; elections were often "made" to suit the interests of those in power by overt use of army and police personnel; and sharp legal restriction, if not outright prohibition, was generally imposed on revolutionary political organizations and propaganda. Such measures were generally successful, even when the revolutionaries received systematic encouragement from abroad.

Alliances. The Soviet Union provided at least moral support for Communists throughout the period. Fascist Italy spun a web of intrigue aimed against Yugoslavia; by the late 1920s Mussolini's agents had entangled the Bulgaria-based IMRO, extreme Croat nationalists (Ustachi) based in Rome, and some elements of the Hungarian government in plots to dismember King Alexander's state.

In Romania, too, a Fascist movement, founded in 1927 (renamed the Iron Guard in 1928) by Corneliu Codreanu, rose quickly to prominence; it owed little to foreign patronage, however, because Codreanu based his appeal mainly upon harsh anti-Semitism, for which the ground was already well prepared.

The Balkan governments took steps to counter the foreign threat. A diplomatic alliance of Czechoslovakia, Romania, and Yugoslavia—the so-called Little Entente—dated from 1921. In 1934 a new Balkan Pact allied Greece, Yugoslavia, Turkey, and Romania. Such treaties were aimed mainly against Hungarian (Little Entente) and Bulgarian (Balkan Pact) aspirations for frontier revision. In the background, France played the role of great-power patron vis-à-vis Romania and Yugoslavia, rivalling Italy, the patron of Bulgaria and Hungary; Britain, the patron of Greece; and the Soviet Union, the pan-Balkan patron of Communists.

The Balkans in the 1930s. The economic depression that settled upon world trade in the early 1930s put the Balkan countries at a great disadvantage. Farm prices plummeted, making it all but impossible for high-cost Balkan peasant producers to compete on world markets. Hardship and a pervasive sense of failure and confusion in official quarters encouraged revolutionary sentiment, especially among the young. But the enhanced power of revolution, especially in Communist guise, provoked more ruthless repression. Democratic and parliamentary government seemed to have failed everywhere, especially after 1933, when the rising influence of Nazi Germany began to make itself felt in the Balkans.

The effects of the world depression

German trade policy, as a matter of fact, brought an effective solution to the economic crisis that had paralyzed the Balkan markets since 1930. The Nazis offered to buy agricultural products from the Balkans, at prices fixed through bilateral negotiation, and offered manufactured goods in exchange, again at prices fixed by interstate bargaining. Germany often drove a hard bargain in these trade negotiations, but the fact was that only in Germany could high-cost Balkan farm products find any kind of sale. Hence, these exchanges—administered by state trading agencies and kept track of through blocked accounts managed by state banks—benefitted all parties. Romania and Bulgaria, in particular, began to market substantial surpluses in Germany.

By the late 1930s the resulting economic improvement allowed precarious political stabilization in both countries. Thus, for example, King Boris III of Bulgaria pulled IMRO's teeth by condoning, if not instigating, a coup d'etat in 1936, which brought to power a military group that dispersed IMRO's gunmen and drove its leaders from the country. Similarly, in 1938 King Carol II of Romania was able to have Codreanu killed and to repress the Iron Guard through unscrupulous and high-handed police methods; the king owed his success to the fact that, despite the strong appeal of its anti-Semitism, trade revival had taken the cutting edge from the Iron Guard's revolutionary agitation.

Greece followed a similar route. The early 1930s saw a rising political crisis, with abortive coups d'etat in 1933 and 1935. A "managed" plebescite in 1936 led to the recall of King George II from exile; but new elections resulted in a parliamentary deadlock between royalists and republicans. A handful of Communist deputies held the balance of power, being in a position to make or break any parliamentary majority. King George reacted to this situation by entrusting the government to Gen. John Metaxas, who ruled without a parliament until his death in 1941.

In Albania, however, King Zog's regime collapsed in 1939, and Italians took over direct administration of the country. This allowed Mussolini to threaten both Yugoslavia and Greece across a new frontier. Such an advance of Italian power was profoundly disturbing to both Balkan governments. Ever since 1934, when King Alexander of Yugoslavia had been assassinated in Marseilles by an IMRO gunman (supported by Hungary and Italy), Yugoslavia's internal problems had offered the Italians a promising field of action. Prince Paul, brother of the assassinated Alexander, took over the reins of government as regent for the heir, King Peter II, who was still too young to rule. Paul, like other Balkan monarchs, experimented with authoritarian rule and worked diligently for diplomatic rapprochement with both Bulgaria and Italy.

Italian control of Albania

In the late 1930s Paul decided that the kingdom could survive only by coming to a basic understanding with the Croats. Prolonged negotiations led to an agreement, in August 1939, by which a generously defined Croatia would enjoy extensive autonomy; the Croatian Peasant Party accepted this arrangement, and only certain extremists (the Ustachi) remained irreconcilable. This drastically weakened the Italian leverage inside Yugoslavia; but it did not solve the state's problems, because most Serbs balked at the prospect of giving the Croats control over territory in which large numbers of Serbs lived. Yet nothing less would satisfy Croatian ambitions. As a result, the federal structure for Yugoslavia as promised in the 1939 agreement had not been fully implemented when the flood tide of World War II overwhelmed the Balkan Peninsula.

WORLD WAR II, 1939–45

Axis victories and occupation. Stalin's cooperation with Hitler against Poland (September 1939) turned previously proferred Anglo-French guarantees of Polish and Romanian territorial integrity into worthless scraps of paper. The Romanian government had to yield Bessarabia and Bukovina to the Soviet Union (June 1940), part of Transylvania to Hungary (August 1940), and the southern Dobruja to Bulgaria (September 1940). King Carol, discredited by such losses, fled the country, and Gen. Ion Antonescu became dictator. Antonescu invited German troops to enter Romania in October 1940. The Romanian government remained a loyal and relatively enthusiastic ally of the Nazis in their struggle against the Soviet Union until 1944.

The Italian government, jealous of Germany's successes in Poland, Scandinavia, and France, decided to make Greece into a dependency. When the Greek dictator Metaxas refused to yield to Mussolini's ultimatum (October 1940), Italian troops crossed the Albanian frontier, expecting to meet only token opposition. To their surprise and discomfiture, the politically divided Greeks joined ranks against the Italians and drove the attacking forces back across the Albanian border. Fearing the provocation of German military intervention, the Greek government reacted coolly to the British offers of air and naval support; these fears assumed new plausibility when Nazi troops moved secretly into Bulgaria in March 1941.

German advance into Bulgaria

The German move attempted to forestall Soviet influence in Bulgaria and to compel Greece and Yugoslavia to repudiate ties with Great Britain and align themselves with the Nazi cause. As a diplomatic prelude to their planned attack on the Soviet Union, the Germans demanded that each Balkan state adhere to the Tripartite Pact (concluded initially by Germany, Italy, and Japan in September 1940). Hungary and Romania did so in November 1940; Bulgaria followed suit on March 1, 1941; and the Yugoslavs reluctantly did the same a few weeks later (March 25). News of this act led Serbian radicals, already bitterly opposed to the deal the government had concluded with the Croats, to seize power in Belgrade. This act of defiance outraged Hitler, then at the summit of his diplomatic success; to safeguard his southern flank for the thrust against the Soviet Union, he determined to crush the Yugoslavs and Greeks in a lightning campaign.

Accordingly, on April 6, 1941, German forces attacked

and speedily overran Yugoslavia. A hastily assembled British expeditionary force scarcely reached Greece from North Africa when headlong retreat began. In May 1941 German parachutists invaded Crete, and by the end of the month they had won a costly victory there; German arms had thus chalked up yet another brilliant success, but it forced them to postpone by a few weeks the start of the campaign against the Soviet Union. Whether this delay actually helped the Soviets survive Hitler's assault can never be definitively decided; but Greeks and Yugoslavs easily convinced themselves that their nations' defeat was an essential precondition for the later Soviet victory.

Greece and Yugoslavia remained under Axis occupation until 1944–45. Bulgarian, Italian, and German troops carved out separate zones of occupation—most of Macedonia was incorporated into Bulgaria; an independent Croatia, ruled by the Ustachi and under Italian patronage, was proclaimed. Romania, too, staked out its territorial claims in Bessarabia and adjacent regions of the Ukraine.

The resistance movements. The first people who actively opposed these arrangements were Serbs, who had everything to lose and whose national tradition of heroic outlawry against the Turks inspired guerrilla action. Bands of demobilized Serbian soldiers—"Chetniks"—formed under the leadership of Col. Draža Mihailović and engaged in acts of sabotage. When the Germans sent fresh troops into Serbia and retaliated brutally against communities suspected of harbouring guerrillas, the Chetniks abandoned active operations, husbanding their strength against a future day of liberation when Germany's defeat would permit them to restore the Serbian nation to its accustomed and proper position in the Balkans.

After the Nazi attack on the Soviet Union (June 22, 1941), the Communist parties of the Balkans, which had previously cooperated with the Germans, made an abrupt about-face. Communist activity in Bulgaria remained marginal until 1944; in Romania the party was of trifling importance, being firmly identified with Jewish and other minority groups.

But in Yugoslavia, Albania, and Greece the Communists rapidly built up powerful resistance organizations. The Communist Party line was to cooperate with all anti-Fascist elements in the population; hence, popular front, uniting Socialists, peasants, nationalists, and anyone else willing to cooperate with Communists, sprang into existence. Communist policy systematically played down Marxism and disguised the extent of party control over the network of political and military resistance organizations they created.

The popular fronts

An essential strength of the Communists in this situation was the availability in each Balkan state of underground party cadres already accustomed to survival in face of police harassment. In addition, the Communists emphasized political organization in towns and countryside, thus providing the armed guerrilla bands responsive to their leadership with far firmer support than any rival organizations could offer. Finally, Communist policy aimed at helping the common cause of Socialist revolution by fighting Germans wherever they could be found; this simple policy had the effect of attracting the support of restless and active men throughout the western Balkans. Hence, by degrees the Communists were able to live down their traitorously antinational past and to far outdistance all rivals.

In Yugoslavia, when the Communist-led Partisans first took the field in the summer of 1941, they tried, in accord with popular front tactics, to cooperate with Mihailović's Chetniks. But violent quarrels soon broke out. Two points were at issue: whether to persist in active operations against the Germans and Italians, despite the cost to civilian populations, as Tito (Josip Broz), the Communist leader, desired, (Mihailović opposed this); and whether to welcome all Yugoslavs into the resistance, as Tito assumed, or accept only Serbs, as Mihailović felt was necessary.

Behind these disagreements lay rival visions of the future: Mihailović desired above all to protect Serbdom against Croat, Bulgar, and other threats, whereas Tito (*q.v.*) envisioned a revolutionary brotherhood of all Balkan nationalities, as projected by the Comintern ever since the 1920s. In the long-drawn-out struggle that ensued, advantage lay overwhelmingly with Tito; his policy of conducting active operations against the occupiers of his country won the backing of the British, American, and Soviet allies (Tehrán Conference, November 1943). In addition, as the war continued to bite into their daily lives the peoples of Yugoslavia more and more rallied to the only active transnational Yugoslav organizations in sight: Tito's Partisans and the Anti-Fascist Peoples' National Council (AVNOJ), the political voice and arm of the movement.

In Greece, the Communist resistance also took the form of an armed guerrilla organization, known by the acronym ELAS, and of a political organization, called National Liberation Front (EAM). British policy, however, never abandoned support for the Greek government in exile, headed by King George II. In Greece itself, anti-Communist guerrilla groups that were supported by British agents and supplies continued to divide the ground with ELAS.

The situation in Albania was similarly confused; rival Albanian resistance groups achieved effective organization only after 1943, when the surrender of the Italian government to the Allies in September put large stocks of arms into Albanian hands. The Italian surrender also allowed the resistance movements of Yugoslavia and Greece to acquire large quantities of Italian weapons and briefly to take possession of territories formerly policed by Italian troops.

Acquisition of Italian weapons

The Germans were able to reoccupy major cities and lines of communication in the former Italian zones of occupation; but by 1944 the continuing retreat of German armies in the Soviet Union made clear the ultimate outcome of the war. On August 23, 1944, when the advancing Soviet forces were close to the Romanian border, King Michael kidnapped General Antonescu and adroitly changed from the Axis to the Allied side. As a result, Soviet troops were able to advance rapidly toward the Danube, and Romanian units, formerly Hitler's allies, switched allegiances, turning their arms against the Germans.

The Bulgarians followed suit on September 9, as the Soviet advance guard neared their border. The Soviets then turned westward, passing through a corner of Yugoslavia (Belgrade was liberated, October 20, 1944) on their way to Budapest.

This abrupt military reversal compelled the Germans to withdraw what forces they could from the western Balkans. Isolated German garrisons, such as that in Crete, were left behind; but the main forces moved northward in good order, and at the end of the war (May 1945) much of Croatia still remained under German occupation.

The last Germans left Greece in October 1944, whereupon the royal Greek government returned to Athens (October 18) with the protection of a handful of British troops; quarrels soon broke out, however, and, during six bitter weeks (December 1944 to January 1945), fighting flared in Athens between British forces and the resistance guerrilla army, ELAS. (For further information see WORLD WARS: *World War II.*)

WORLD WAR II TO THE PRESENT

Postwar settlement, 1945–49. The armed collision between British- and Communist-led forces in the streets of Athens showed how hard it was, as the prospect of final victory over Germany came closer, for the great Allied powers to keep on cooperating. In May 1944 the United States had reluctantly approved a division of the Balkans into British and Soviet spheres of influence; this was confirmed and adjusted in favour of the Soviet Union when the British prime minister, Winston Churchill, visited Moscow that September to settle details of armistice arrangements with Romania and Bulgaria and to clear the way for the British landing in Greece.

British and Soviet spheres of influence

The Soviets were not unhappy to see the British resort to high-handed means to enforce their will in Greece. They hoped for a similar free hand in Romania, where the absence of any strong native Communist party made it particularly difficult for them to establish a government they could depend on. The Bulgarian Communist Party, though, was relatively strong and, operating through a "Fatherland Front," gave the Russians no cause for concern.

In Yugloslavia, however, where Tito easily took power as the Germans withdrew, the Soviets found it hard to restrain the Partisans' revolutionary enthusiasm. In the first flush of victory, Tito's followers were eager to put the Communist recipe for Balkan federation into effect and saw no reason to compromise with the "effete capitalist imperialists" of Britain and the United States over such an issue as control of Trieste. When the Soviets, in conformity to their deal with Churchill to divide the Balkans into spheres of influence, advised Tito to make conciliatory gestures towards the Royal Yugoslav government (in exile since 1940), the Yugoslav Communists reluctantly complied by allowing King Peter's representative, Ivan Šubašić, briefly to join the Cabinet. But the Partisans had come to power by taking risks and despising compromise; accordingly, many of Tito's followers were appalled at the Soviet Union's unwillingness to act upon revolutionary principles in the immediate postwar period.

The British, American, and Soviet foreign ministers slowly negotiated peace treaties with the former enemy states of the Balkans (1945–47). As long as the Anglo-Americans had not recognized the postwar Communist-dominated regimes of Bulgaria, Romania, and Hungary, the Soviets had a powerful argument for moderation; Stalin accordingly supported collaboration with all anti-Fascists and opposed further revolutionary adventures in the Balkans.

In Romania and Hungary there were strong practical reasons why out-and-out Communist Party dictatorships could not come to power—national feeling was distinctly anti-Soviet Union and, therefore, anti-Communist, and pre-existing Communist Party structures were extremely weak. But even where there were strong Communist parties, as in Bulgaria and Yugoslavia, until 1947 Soviet policy continued to support anti-Fascist popular fronts in which Communist preponderance was at least partially camouflaged.

Changes in Soviet policy toward the Balkans

In 1947 two events altered Soviet policy toward the Balkans. First, the United States and Great Britain signed peace treaties with Bulgaria, Romania, and Hungary in February. These treaties did not much alter interwar boundaries, although Romania did lose Bessarabia and Bukovina to the Soviet Union once again and ceded a small strip of Dobruja to Bulgaria; Greece acquired the Dodecanese Islands in the Aegean from Italy; and Yugoslavia annexed a strip of territory in the Istrian peninsula. The treaties did, however, disband the Allied armistice commissions, which had exercised some control over local affairs in both Bulgaria and Romania following 1944.

The cancellation of British and American legal claims to authority in these countries freed local Communists from what had been a real, if modest, hindrance. At the same time, because the treaties authorized the Soviets to maintain their forces in Romania as "lines of communication" troops to forward garrisons in Austria, the Communists retained what was still an essential guarantee of their hold over Romania.

The second event that altered Soviet policy occurred in March 1947, just a month after the peace treaties came into effect, when Pres. Harry S. Truman asked the United States Congress to authorize military and economic aid to Greece and Turkey, both of which were under Communist pressure. After prolonged debate, Congress voted to approve Truman's request and thereby endorsed the "Truman Doctrine," according to which the United States undertook to defend the rule of law internationally by coming to the aid of any government threatened with Communist subversion.

The U.S. decision to back the Greek government against a renewed Communist guerrilla movement signalized a sharp shift of American policy vis-à-vis the Soviet Union. Quarrels over the peace treaties, over Poland, over the occupation regimes in Germany and Japan, as well as over the rising tide of Communist power in China and several other parts of Asia, all contributed to the change of American public attitudes; but events in the Balkans were also important.

Spread of Communism. In 1945 and 1946, through a series of diplomatic notes, the Soviets tried to browbeat the Turks into giving them military bases on the straits between the Black Sea and the Aegean. In addition, Tito and other Balkan Communists, most notably Georgi Dimitrov of Bulgaria, set about actively implementing their ideal of Balkan confederation. A first step was to federalize Yugoslavia itself. A new constitution, closely modelled on that of the U.S.S.R., was accordingly promulgated in January 1946; it established six federal republics (Serbia, Slovenia, Croatia, Montenegro, Macedonia, Bosnia) and several autonomous regions. King Peter lost his throne, and Communist Party control came fully into the open.

Attempts to form a Balkan confederation

A second step was to bring a reliable Communist regime to power in Albania. The Yugoslavs sent troops and technical advisers to help Albanian revolutionaries seize firm control (January 1946). Soon the entire country began to behave much like another constitutive republic of Tito's emerging Balkan superstate.

The next item on the Communists' agenda was never realized: the creation of a united Macedonia that would combine the portions of that land belonging to Greece, Bulgaria, and Yugoslavia into a single whole. Bulgaria did in fact briefly cede Pirin Province to the new Macedonia, but the Greeks refused to cooperate. Accordingly, Tito sent Greek veterans of the wartime guerrilla force (who had retreated into Yugoslavia at the end of the war) back to Greece, where they formed the core around which fresh bands of guerillas quickly formed in 1946 and 1947.

The Greek government's efforts to repress this renewal of guerrilla activity were ineffective; British resources were too straitened at home to permit fresh involvement in Greece. Hence, for a few months—until the United States committed itself fully to stopping the Communist advance—Tito's revolutionary policy seemed on the verge of paying off.

Such heady prospects encouraged the Yugoslavs to be aggressive along their frontier with Italy, demanding further territory to unite all Slovenes in the new Slovene republic. They also entered into negotiations with the Bulgarians (and perhaps also with Romanian Communists) for merging their countries into the proposed Balkan federal state.

Tito's activity antagonized the United States and was a potent factor in persuading the U.S. Congress to support the embattled Greek government in 1947. Tito's effect on Soviet policy remains a matter for speculation; Stalin probably backed the idea of abandoning the popular-front tactic and the placing of out-and-out Communist regimes in power. Soviet diplomatic agents played the key role in driving King Michael from the Romanian throne (December 1947), thus instituting Communist Party dictatorship in that country. Bulgarian Communists needed no outside help to achieve the same result by the end of 1947. For further information see also INTERNATIONAL RELATIONS (1945 TO C. 1970).

Communist regimes in Romania and Bulgaria

Tito's break with the Soviet Union. Eventually, Tito's efforts to overthrow the royal government in Greece and to press ahead with Balkan federation alienated the Soviets. Perhaps Stalin feared Tito's independence or attributed the United States' involvement in Greece and Turkey to what he saw as Tito's recklessly revolutionary policies.

At any rate, in 1948 the Soviet dictator decided to call Tito to heel. With characteristic guile, Stalin set out to overthrow Tito by stirring up an intrigue within the ranks of the Yugoslav Communist Party. It did not work; Tito's prestige inside his own country was too

great for an outsider—even Stalin—to succeed in unseating him.

But when the quarrel between the Soviet Union and Tito came out into the open (June 1948), the entire strategic situation in the Balkans altered abruptly. Albanian Communists, warmly backed by the Soviet Union, broke away from Tito and unceremoniously evicted all the Yugoslav experts and advisers who had until then been running the country.

Bulgaria and Romania disclaimed all sympathy for Tito and hurried to participate in Stalin's economic blockade and propaganda war against the stubborn Yugoslavs (all the more energetically because of their previous associations with the new heretic).

In Greece, the Tito–Stalin split meant, first of all, the cessation of Yugoslav aid to the Greek guerrillas. This crippled the guerrilla cause. Then, in an effort to get the Macedonians to imitate the Albanians and secede from Tito's dominion, the Cominform (Communist Information Bureau, established in 1947) announced its program of a united Macedonia. Radio broadcasts failed to stir the Yugoslav Macedonians to action; but the news did create consternation in the ranks of the Greek guerrillas, who were unwilling to fight for a cause that, it now appeared, would lead, if successful, to the surrender of Greek territory to a Slav people. Greek Communist morale therefore collapsed, so that Greek government troops—equipped, advised, and assisted by a large American military mission—found it easy to win decisive victory in the summer of 1949.

Changes caused by the war. By August 1949, therefore, active military operations in the Balkans ceased; nearly a decade of war thus came to a conclusion. By that date Communist Party dictatorships were firmly established in all the Balkan countries except Greece and Turkey.

Yet, in spite of this fact, the changes World War II brought to the Balkans were distinctly less drastic than those that came during the 20th century's earlier decade of Balkan fighting, 1912–23. Boundaries shifted only slightly after 1945; and war or postwar population movements wiped out or greatly reduced the numbers of some national minorities—Germans and Jews in particular. The effect was to confirm and enhance the sorting out of Balkan populations into territorially defined national states, according to the patterns of 1918–19.

Confirmation of national states

From a political and social point of view the one-party regimes of the northern Balkans, set up by Communists, proved a little more ruthless and persistent in pursuit of the same goals than did the interwar Balkan governments: economic and political mobilization of the peasant mass—by police hectoring and controlled elections if need be—to hasten the development of industries and cities.

All in all, the major crisis of transition from a traditional peasant and premodern style of life apparently took place in the Balkans before and after World War I, whereas after 1949 somewhat more stable patterns of modernization and mobilization established themselves in Greece as much as in Communist lands. The waning force of revolutionary movements of every kind, a marked feature of the post-World War II Balkan scene, is an index of this basic transformation.

The 1950s and after. Though Balkan politics after 1949 were less tumultuous and also less bloody than was the case earlier, confusion and upheavals were not lacking. Among the Communist states, fluctuating relations with the Soviet Union defined major shifts of government policies.

Political and economic changes. From 1948 to 1953 all Stalin's satellites in eastern Europe purged real or suspected "Titoists"—*i.e.*, Communist leaders suspected of putting national interests ahead of subservience to the Soviet Union. The countries formerly closest to Tito were, not surprisingly, the most energetic in carrying out such purges.

Albania broke away from Yugoslavia and became for a while the Soviet Union's most enthusiastic satellite. The Bulgarian government staged a show trial against Titoists and used the occasion to implicate personnel of the United States embassy, with the result that official U.S.–Bulgarian relations were broken off in 1950, not to be resumed till 1966. Yet, in spite of a trade blockade and various kinds of subversive activity aimed against the Yugoslav regime, Tito's power remained unshaken. By cautiously accepting profferred American aid, the Yugoslavs managed to survive even Stalin's wrath.

After Stalin's death in 1953, the Soviet government tried to mend its fences with Tito. As a result, the Yugoslav government was able to balance itself between the Soviet Union and the United States—playing one off against the other and even getting economic and military aid from both great powers at once. The advantages of such an independent policy were obvious to other Balkan Communist states. But not until after 1961, when a quarrel between Chinese and Soviet Communists came into the open, did first the Albanians and then the Romanians venture to defy the Soviet Union.

Albania's policy was governed mainly by fear of a lasting Yugoslav–Soviet rapprochement; when that seemed likely in 1961, the Albanians threw out their Soviet advisers, as they had earlier ejected Yugoslavs from similar roles, and allowed the distant Chinese to become their new patrons.

Defiance of the Soviet Union

By the early 1970s Albania was expanding and normalizing its relations throughout the world. In 1971 diplomatic relations were either established or elevated to the ambassadorial level in Greece (with which country Albania had been in a technical state of war since 1940), Yugoslavia, and a number of other countries. Chinese influence, however, remained dominant in Albania, with China supporting numerous industrial enterprises under Albania's five-year plan for 1971–75.

A far more serious break in the Communist ranks occurred in 1964, when Romania declined to accept the role assigned to it by the Soviet Union in the Comecon (Council for Mutual Economic Assistance) plan for economic development of eastern Europe. Instead of concentrating on agriculture for export, the Romanian government wished to continue to emphasize industry, even if this meant duplicating Czech or East German factories. Romanian national feeling, always strongly anti-Soviet, surged to the surface in support of the government's stand.

In Greece, American policy played a dominating role after 1947. Until 1956 American funds continued to provide the Greek government with substantial aid for economic rehabilitation; the United States thus was able to control a number of important aspects of government policy.

After 1956, aid tapered off, and American diplomats deliberately tried to pull out of Greek politics. Such aloofness was all the more attractive to the United States because Greece seemed to have attained a relatively stable parliamentary regime under Marshal Alexandros Papagos—hero of the Albanian war of 1940 and of the guerrilla war of 1947–49—and, after his death (1955), Konstantinos Karamanlis, who survived three elections to remain prime minister for an unprecedentedly long period (1955–63).

Yet the fact that Greece had become a member of the North Atlantic Treaty Organization (NATO) in 1952 meant that American influence upon the Greek armed forces remained strong even after other forms of aid had stopped. And, because the political attitudes of the army mattered in Greek politics, American efforts to leave the Greek politicians alone were never very successful. Thus, in April 1967, when a clique of army officers seized power by coup d'etat, American agents were generally suspected of being responsible for what was probably planned in secret by a narrow and purely Greek circle. The policy of the new Greek government lent some colour to the charges, however, because the officers who emerged as rulers of the country combined an enthusiastic anti-Communism with an earnest courtship of private foreign (principally American) investment, which they needed to balance their international accounts.

Military coup in Greece

The authoritarian military dictatorship, headed by Col. Georgios Papadopoulos, lasted till 1974, when intrigues

aimed at bringing Cyprus into union with Greece miscarried and the military rulers of the country were discredited. Karamanlis was recalled from his self-imposed exile in Paris and organized new elections, thus restoring the legal forms of parliamentary democracy. In a national referendum in December 1974 the Greeks rejected the monarchy and became a republic. Relations with Turkey, which had nearly come to war over Cyprus, were not easily healed; and indignation at the failure of NATO commanders to support the Greek cause against Turkey led Karamanlis to cancel most of the agreements that had bound the Greek armed forces to NATO.

These political upheavals and the often noisy public debate that characterized Greece stood in sharp contrast to the strict controls on public discussion that Communist regimes enforced in other Balkan countries. This difference, however, should not disguise some important resemblances in the pattern of development pursued by all Balkan governments in the post-World War II era. Everywhere officials preferred industrial to agricultural investment and subjected most economic activities to elaborate, often cumbersome bureaucratic control. Everywhere, too, cities grew rapidly, offering scope for peasant sons and daughters to pursue new careers more attractive than those their parents had known. This permitted a general relaxation of political and economic dissatisfactions, despite the numerous unsolved problems that persisted within all the Balkan nations.

The only important difference in economic policies was that Greece continued to import private foreign capital and linked internal prices to world markets, whereas Communist countries maintained a command price system at home and regarded all private investment as detestable capitalist exploitation. When importing foreign capital, the Communist countries were willing only to borrow from other governments—a policy that seemed unlikely to make borrowers any less dependent on lenders.

Tangible increases of industrial output occurred in every Balkan state, with the possible exception of Albania. This did not mean, however, that problems of distribution and assortment of products to suit the consumers had been solved. Quite to the contrary, in the Communist lands shoddy goods and insufficient supplies of some commodities were still the rule; and many of the new factories, built as a result of political decisions, were distressingly inefficient when it came to costs. In Greece, unsolved problems were primarily those of income distribution among the different social classes, although inefficient and high-cost factories that sheltered behind tariff and quota protection were not absent.

Recognition of the difficulty of adjusting a command economy to consumer needs led the Yugoslavs to experiment with freer market prices on the one hand and with worker–management partnership in factory administration on the other. A new constitution, effective in 1963, gave enhanced power in economic matters to the federal republics in the hope that this, too, would bring planning and management closer to the people. Yugoslavia met some success in this effort at decentralization and bureaucratic simplification. But divisive tendencies among the various nationalities of the state were perhaps reinforced by allowing greater popular participation in management decisions. Nevertheless, on June 30, 1971, the federal government instituted new amendments that further enhanced the powers of the republics and municipalities, granting the right, for example, to initiate investment projects. The federal government reserved for itself sole power in the areas of defense, foreign policy, and general economic policies.

Political and economic growth

In spite of these and other unsolved problems, the political and economic history of the Balkans after 1949 must be rated an overall success, at least as compared to the experience of earlier decades of the 20th century and to the whole of the 18th and 19th centuries. Peace prevailed, as seldom before. Population and developed resources came more nearly into balance. Administrative and technical skills attained such a level that plans began to have reasonable relation to results. Mass mobilization, both politically and economically, was achieved, along with the enhancement of wealth and power usually expected from such successful modernization.

Effects of the changes. The costs were, of course, very great. As more and more young people left their village homes and managed to establish themselves in the expanding complex of administrative paper work, factory labour, construction work, etc., the fabric of traditional rural life altered profoundly in all the Balkan nations. Old peasant customs decayed; urban styles of dress and, still more, of expectations spread into the remotest villages; handicrafts survived, if at all, mainly on the strength of the tourist trade. The Balkan nations, in short, rapidly transformed themselves into provincial variants of the world-girdling Western style of civilization. Insofar as they were successful, new pains arose from the dissatisfactions of provincialism; but the far sharper psychological strains that had afflicted the generation of Balkan peasants who first witnessed the breakdown of their traditional village ways of life faded into the past. Proportionally, the fierceness of revolutionary aspiration dwindled, fed no longer by the agony of displaced young men who, like Gavrilo Princip in 1914, were glad to give their lives for a cause because only so could they transcend daily frustration and apparent failure.

Regarded as a case study of the interaction between premodern, traditional peasant styles of life and the aspiration for modernity, Balkan history after 1945 has wide significance. The majority of mankind launched itself upon a similar trajectory a good deal later than did the Balkan peoples and could perhaps be expected to follow a somewhat similar curve of historical development in the rest of the 20th century. (For additional information on contemporary affairs in the Balkans see ALBANIA; GREECE; ROMANIA; BULGARIA; and YUGOSLAVIA.)

(W.H.McN.)

BIBLIOGRAPHY

The Balkans to 1815: JOHN ALEXANDER, *Yugoslavia Before the Roman Conquest* (1972), a general summary of archaeology and protohistory of this country with emphasis on the early Iron Age; DUMITRU BERCIU, *Romania* (1967), archaeological survey from the Palaeolithic period to the Geto-Dacian civilization; STANLEY G. EVANS, *A Short History of Bulgaria* (1960), includes an outline of the early historical and cultural setting for the state; JOSEPH WIESNER, *Die Thraker* (1963), a detailed account of the Thracians to AD 600; MARIJA GIMBUTAS, "The Neolithic Cultures of the Balkan Peninsula," in *Aspects of the Balkans* (1972), a concise account of the Neolithic and Chalcolithic civilizations, 6500–3500 BC; *The Bronze Age Cultures in Central and Eastern Europe* (1965), a monograph that includes the Danubian region and Urnfield migrations; *The Slavs* (1971), an account of the beginning and early history of the Slavic peoples and their migrations to the Balkan peninsula; and the *Symposium* of the Centre d'Études Balkaniques, Sarajevo (1964), a collective work dedicated to the distribution and chronology of the prehistoric Illyrians. (*The Balkans in the Middle Ages*): National histories of the individual Balkan states are not always dependable for the medieval period. Two books on the Balkans as a whole, however, are quite reliable and up to date: GEORG STADTMULLER, *Geschichte Südosteuropas* (1950); and DIMITRY OBOLENSKY, *The Byzantine Commonwealth: Eastern Europe, 500–1453* (1971). The Byzantine context of the Balkan peninsula in the medieval period is well treated in the relevant chapters of *The Cambridge Medieval History*, 2nd ed., vol. 4 (1966); and on a more modest scale in GEORGE OSTROGORSKY, *Geschichte des byzantinischen Staates* (1965; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968). On the Slavic states in the Balkans, see the two studies of FRANCIS DVORNIK, *The Slavs: Their Early History and Civilization* (1956) and *The Slavs in European History and Civilization* (1962). All these books have substantial bibliographies for further reading. (*The Ottoman era*): The best general discussion of the Balkans in the Ottoman era is to be found in L.S. STAVRIANOS, *The Balkans Since 1453*, ch. 3–12 (1958). The Ottoman conquest of the Balkans and its effects on European diplomacy are ably described in PAUL COLES, *The Ottoman Impact on Europe* (1968). Two excellent studies of dramatic events in this period are STEVEN RUNCIMAN, *The Fall of Constantinople, 1453* (1965); and THOMAS M. BARKER, *Double Eagle and Crescent: Vienna's Second Turkish Siege and Its Historical Setting* (1967). STEVEN RUNCIMAN, *The Great Church in Captivity* (1968), covers the patriarchate in Constantinople from the Turkish conquest until the 1820s.

The Balkans from 1815 to 1914: The period from the early-19th century till the 1950s is treated in RENE RISTELHUEBER, *Histoire des peuples balkaniques* (1950; Eng. trans., *A History of the Balkan Peoples*, 1971). A short review of the Balkans in the 19th century may be found in L.S. STAVRIANOS, *The Balkans, 1815–1914* (1963); and in CHARLES and BARBARA JELAVICH, *The Balkans* (1965). See also the relevant sections in Stavrianos' excellent general history, *The Balkans Since 1453*. The comparative method is applied in L.S. STAVRIANOS, *Balkan Federation* (1944); TRAIAN STOIANOVICH, *A Study in Balkan Civilization* (1967); CHARLES and BARBARA JELAVICH (eds.), *The Balkans in Transition* (1963); DIMITRIJE DJORDJEVIC, *Révolutions nationales des peuples balkaniques, 1804–1914* (French trans. 1965); DOREEN WARRINER (ed.), *Contrasts in Emerging Societies: Readings in the Social and Economic History of South-eastern Europe in the Nineteenth Century* (1965).

National histories: (*Yugoslavia*): Different aspects of Yugoslav development may be found in ROBERT J. KERNER (ed.), *Yugoslavia* (1949); and ROBERT F. BYRNES (ed.), *Yugoslavia* (1957). Short, condensed histories include: PHYLLIS AUTY, *Yugoslavia* (1965); A.W. PALMER, *Yugoslavia* (1964); MURIEL HEPPELL and FRANK B. SINGLETON, *Yugoslavia* (1961); Z. KOSTELSKI, *The Yugoslavs* (1952); STEPHEN CLISSOLD (ed.), *A Short History of Yugoslavia: From Early Times to 1966* (1966); and WERNER MARKERT (ed.), *Jugoslawien* (1954). The best social and economic history is JOZO TOMASEVICH, *Peasants, Politics, and Economic Change in Yugoslavia* (1955). The Yugoslavs in the Habsburg monarchy are studied in articles published in the *Austrian History Yearbook*, vol. 3, pt. 2 (1967). (*Greece*): The best, although very short history of 19th-century Greece is NICOLAS G. SVORONOS, *Histoire de la Grèce moderne* (1953). Surveys of Greek development in the same period may be found in JOHN CAMPBELL and PHILIP SHERRARD, *Modern Greece* (1968); W.A. HEURTLEY *et al.*, *A Short History of Greece, from Early Times to 1964* (1965); EDWARD S. FORSTER, *A Short History of Modern Greece, 1821–1956*, 3rd ed. (1958); JOHN N. MAVROGORDATO, *Modern Greece: A Chronicle and a Survey, 1800–1931* (1931); and C.M. WOODHOUSE, *The Story of Modern Greece* (1968). The standard work for diplomatic history is J. EDOUARD DRIAULT and MICHEL LHERITIER, *Histoire diplomatique de la Grèce de 1821 à nos jours*, 5 vol. (1925–26). See also GEORGE FINLAY, *A History of Greece, from Its Conquest by the Romans to the Present Time, B.C. 146 to A.D. 1864*, new ed., 7 vol. (1877, reprinted 1970); and WILLIAM MILLER, *A History of the Greek People, 1821–1921* (1922). (*Bulgaria*): Bulgarian 19th-century development is covered in D. KOSSEV, H. KHRISTOV, and D. ANGELOV, *A Short History of Bulgaria* (Eng. trans. 1963); and MERCIA MACDERMOTT, *A History of Bulgaria, 1393–1885* (1962). For internal Bulgarian politics, see CYRIL E. BLACK, *The Establishment of Constitutional Government in Bulgaria* (1943, reprinted 1970). IRWIN T. SANDERS, *Balkan Village* (1949), gives a description of post-World War I changes in Bulgarian rural life. (*Romania*): NICOLAE IORGA, *Histoire des Roumains et de la romanité orientale*, 5 vol. (1937), is still a basic general history, as is ROBERT W. SETON-WATSON, *A History of the Roumanians, from Roman Times to the Completion of Unity* (1934, reprinted 1963). The history of the unification of the country is presented by THAD W. RIKER in *The Making of Roumania* (1931); for recent developments, see STEPHEN A. FISCHER-GALATI, *Twentieth Century Rumania* (1970). (*Albania*): The best modern study of Albania is STAVRO SKENDI, *The Albanian National Awakening, 1878–1912* (1967). An older book is JOSEPH SWIRE, *Albania: The Rise of a Kingdom* (1929, reprinted 1971). The Soviet historian, ИРИНА ГРИГОРЬЕВНА СЕНКЕВИЧ, published (in Russian): *Албания в период восточного кризиса, 1875–1881* (1965); *Освободительное движение албанского народа в 1905–1912* (1959); and with a group of authors, *Краткая история Албании* (1965). A study of the 19th-century Albanian movements may be found in THEODOR IPPEN, "Beiträge zur inneren Geschichte Albaniens im XIX. Jahrhundert," in LUDWIG VON THALLOCZY (ed.), *Illyrisch-albanische Forschungen . . .*, vol. 1, pp. 342–385 (1916). (*Ottoman Empire*): Besides the standard works of WILLIAM MILLER, *The Ottoman Empire and Its Successors, 1801–1927*, 4th ed. (1936); and NICOLAE IORGA, *Geschichte des Osmanischen Reiches*, 5 vol. (1908–13), many modern studies deal with the reform activities in the Ottoman Empire: ERNEST E. RAMSAUR, *The Young Turks: Prelude to the Revolution of 1908* (1957); RODERIC H. DAVISON, *Reform in the Ottoman Empire, 1856–1876* (1963); ROBERT DEVEREUX, *The First Ottoman Constitutional Period: A Study of the Midhat Constitution and Parliament* (1963); FRANK E. BAILEY, *British Policy and the Turkish Reform Movement: A Study in Anglo-Turkish Relations, 1826–1853* (1942). Financial and economic aspects are discussed in DONALD C. BLAISDELL, *European Financial Control in the Ottoman Empire* (1929). EDWARD M. EARLE deals with the Ottoman Empire in the era of imperialism in *Turkey, the Great Powers, and the Bagdad Railway* (1923, reprinted 1966).

The Balkans after 1914: (*General works*): HUGH SETON-WATSON, *Eastern Europe Between the Wars, 1918–1941*, 3rd rev. ed. (1967); and *The East European Revolution*, 3rd ed. (1956); ROBERT L. WOLFF, *The Balkans in Our Time* (1956); R.V. BURKS, *The Dynamics of Communism in Eastern Europe* (1961); GHITA IONESCU, *The Politics of the European Communist States* (1967); YORICK BLUMENFELD, *Seesaw: Cultural Life in Eastern Europe* (1968); TRAIAN STOIANOVICH, *A Study in Balkan Civilization* (1967); C.A. MACARTNEY and A.W. PALMER, *Independent Eastern Europe* (1962); A.W. PALMER, *The Lands Between: A History of East-Central Europe Since the Congress of Vienna* (1970); CHARLES JELAVICH (ed.), *The Balkans in Transition* (1963); STEPHEN A. FISCHER-GALATI (ed.), *Eastern Europe in the Sixties* (1963); NORMAN J.G. POUNDS, *Eastern Europe* (1969); J.F. BROWN, *The New Eastern Europe: The Khrushchev Era and After* (1966). (*On more specialized pan-Balkan topics*): W.E. MOORE, *Economic Demography of Eastern and Southern Europe* (1945); IRWIN T. SANDERS (ed.), *Collectivization of Agriculture in Eastern Europe* (1958); ALFRED BOHMANN, *Menschen und Grenzen*, vol. 2, *Bevölkerung und Nationalitäten in Südosteuropa* (1969).

By country: (*Albania*): JOSEPH SWIRE, *Albania: The Rise of a Kingdom* (1929, reprinted 1971) and *King Zog's Albania* (1937); JULIAN AMERY, *Sons of the Eagle: A Study of Guerilla War* (1948); STAVRO SKENDI (ed.), *Albania* (1956); HARRY HAMM, *Rebellen gegen Moskau: Albanien–Pekings Brückenkopf in Europa* (1962; Eng. trans., *Albania: China's Beachhead in Europe*, 1963); NICHOLAS C. PANO, *The People's Republic of Albania* (1968). (*Bulgaria*): S.G. EVANS, *A Short History of Bulgaria* (1960); JOSEPH SWIRE, *Bulgarian Conspiracy* (1939); JOSEPH ROTHSCHILD, *The Communist Party of Bulgaria: Origins and Development, 1883–1936* (1959); KOSTA TODOROV, *Balkan Firebrand: The Autobiography of a Rebel, Soldier and Statesman* (1943); JEAN KANAPA, *Bulgarie d'hier et d'aujourd'hui: le pays de Dimitrov* (1953); IRWIN T. SANDERS, *Balkan Village* (1949); L.A.D. DELLIN (ed.), *Bulgaria* (1957); J.F. BROWN, *Bulgaria Under Communist Rule* (1970). (*Greece*): JOHN CAMPBELL and PHILIP SHERRARD, *Modern Greece* (1968); NICOLAS G. SVORONOS, *Histoire de la Grèce moderne* (1953); C.M. WOODHOUSE, *The Story of Modern Greece* (1968); ARNOLD J. TOYNBEE, *The Western Question in Greece and Turkey*, 2nd ed. (1923, reprinted 1970); ALAN PALMER, *The Gardeners of Salonika: The Macedonian Campaign, 1915–18* (1965); E. SCHRAMM VON THADDEN, *Griechenland und die grossen Mächte, 1913–1923* (1933); ELISABETH BARKER, *Macedonia: Its Place in Balkan Power Politics* (1950); W.H. MCNEILL, *Greek Dilemma: War and Aftermath* (1947) and *Greece: American Aid in Action, 1947–1956* (1957); EDGAR O'BALLANCE, *The Greek Civil War, 1944–1949* (1966); J.P.C. and A.G. CAREY, *The Web of Modern Greek Politics* (1968). (*Romania*): H.L. ROBERTS, *Rumania: Political Problems of an Agrarian State* (1951); DAVID MITRANY, *The Land and the Peasant in Rumania: The War and Agrarian Reform, 1917–1921* (1930, reprinted 1968); S.D. SPECTOR, *Rumania at the Paris Peace Conference* (1962); A. HILLGRUBER, *Hitler, König Carol und Marshall Antonescu* (1954); L.D. PATRASCANU, *Sous trois dictatures* (1946); GHICA IONESCU, *Communism in Rumania, 1944–1962* (1964); STEPHEN A. FISCHER-GALATI, *The Socialist Republic of Rumania* (1969); JOHN M. MONTIAS, *Economic Development in Communist Rumania* (1967). (*Yugoslavia*): STEPHEN CLISSOLD (ed.), *A Short History of Yugoslavia: From Early Times to 1966* (1966); E. HAUMANT, *La Formation de la Yougoslàvie (XV^e–XX^e siècles)* (1930); VLADIMIR DEDIJER, *The Road to Sarajevo* (1966) and *Tito* (1953); JOACHIM REMAK, *Sarajevo* (1959); J.C. ADAMS, *Flight in Winter* (1942); IVO LEDERER, *Yugoslavia and the Paris Peace Conference: A Study in Frontiermaking* (1963); REBECCA WEST, *Black Lamb and Grey Falcon: A Journey Through Yugoslavia* (1941, reprinted 1967); J.B. HOPTNER, *Yugoslavia in Crisis, 1934–1941* (1962); L. HORY and MARTIN BROSZAT, *Der Kroatische Ustacha-Staat, 1941–1945* (1964); E. HALPERIN, *Der siegreiche Ketzer: Titos Kampf gegen Stalin* (1957; Eng. trans., *The Triumphant Heretic: Tito's Struggle Against Stalin*, 1958); ROYAL INSTITUTE OF INTERNATIONAL AFFAIRS, *The Soviet-Yugoslav Dispute: Text of the Published Correspondence* (1948); IVAN AVAKUMOVIC, *History of the Communist Party of Yugoslavia* (1964); MILOVAN DJILAS, *The New Class: An Analysis of the Communist System* (1957); PAUL SHOUP, *Communism and the Yugoslav National Question* (1968); JOZO TOMASEVICH, *Peasants, Politics, and Economic Change in Yugoslavia* (1955); JOEL HALPERN, *A Serbian Village*, rev. ed. (1967).

(M.G./G.P.M./B.Je./C.J./D.V.D./W.H.McN.)

Balkhash, Lake

Situated in the eastern part of the Kazakh Soviet Socialist Republic, U.S.S.R., Lake Balkhash (Ozero Balkhash in Russian; Ozero Balchaš in the transliteration system of the Akademiya Nauk) is contained in the vast Balkhash–Alakol Basin, 1,115 feet (340 metres) above sea level and 600 miles east of the Aral Sea. It is 376 miles long from west to east. Its area varies within significant limits, depending on the water balance. In years in which there is an abundance of water, as at the beginning of the 20th century and in the decade 1958–69, the area reaches 6,900 to 7,300 square miles (18,000 to 19,000 square kilometres). In drought-afflicted periods, however (as at the end of the 19th century and in the 1930s and 1940s), the area of the lake decreases to 6,000 to 6,300 square miles. Such changes in area are accompanied by changes in the water level of about ten feet. Such variability is caused by the structure of the basin and the lake's location. Jutting far out into the lake is the Sarymsek Peninsula (Poluostrov), which divides Balkhash into two separate hydrological parts: a western part, wide and shallow, and an eastern part, narrow and relatively deep. Accordingly, the width of the lake changes from 46 to 17 miles in the western part and six to 12 miles in the eastern part. The depth of the western part does not exceed 36 feet, whereas the eastern part reaches 85 feet. The two parts of the lake are united by a narrow strait, the Uzynaral, with a depth of about 21 feet.

The variability of its area

Hydrology. The large Ili River, flowing in from the south, spills into the western part of the lake and contributes 75–80 percent of the total influx into the lake. Only such small rivers as the Karatal, Aksu, Ayaguz, and Lepsa feed the eastern part of the lake. With almost equal areas in both parts of the lake, this situation creates a continuous flow of water from the western section to the eastern section. The water of the western part is almost fresh (total mineralization 0.74 gram per litre) and suitable for industrial use and consumption; the water of the eastern part, however, is salty (about five grams per litre).

The east bank of the lake bears traces of a historically recent union (perhaps occurring only two or three hundred years ago) of the Balkhash Basin with the basin of Ozero (Lake) Alakol in the Dzungarian Gate to the east. The similarity between the fauna of Lake Balkhash and that of the Tarim River Basin in Central Asia and the dissimilarity of the faunas of the Aral Sea and Lake Balkhash suggest that formerly Lake Balkhash, through Lakes Alakol and Ai-pi, went into the system of lakes that formerly filled the Turfan Depression in the T'ien Shan (Celestial Mountains) and had no connection with the Aral Sea to the west.

The north banks of the lake are high and rocky, with clear-cut traces of ancient terraces. Farther north a dry steppe passes into the undulating Kazakh Plain. The south banks are low and sandy, and wide belts of them are covered with thickets of reeds and numerous small lakes. The low-lying banks, periodically flooded by the waters of the lake, are being continually transformed into the desert sands of the Sary-Ishikotrau and, further away, into the foothills of the Dzhungarsky Alatau.

Climate. Extremely harsh continental conditions prevail and significantly affect the whole regime of the lake. The average (1930–67) air temperature in the western part is 44° F (7° C), with an annual range of from 80° F (June) to −1° F (January; 27° C to −18° C) and, in the eastern part, 39° F (4° C) with a range of from 72° F (June) to −13° F (January; 22° C to −25° C). The water temperature in the western part of the lake is 50° F (10° C); in the eastern part it is 48° F (9° C). Average precipitation is approximately 17 inches. The predominant winds are from the northwest (in the west) and from the northeast (in the east). They are usually fairly strong, giving rise to constantly choppy water. The lake remains frozen from the end of November to the beginning of April.

Animal life. Carbonates predominate in the ground deposits of the lake. The fauna of the lake is rich, especially in regions dense with reeds. Different types of gulls and ducks and a large number of cormorants are most frequently found here. Now and then one can spot swans and pink pelicans. Among the brushwood on the banks, pheasants and partridges can be seen. Wild boars still forage among the reeds; and wolves, foxes, and hares inhabit the thickets. In the past tigers were not infrequent here; the last one was killed in the 1940s.

Twenty species of fish inhabit the lake, of which six are peculiar to the lake itself. The remainder were introduced to the lake by man and include the sazan, sturgeon, eastern bream, pike, and the Aral barbel. The main food fish are the sazan, pike, and Balkhash perch. Bottom life is poor, and the most important sources of food for the fish are benthos (Chironomidae), zooplankton, and the larvae of Tendipedidae. Pollution is only a problem insofar as it results from some of the natural biological processes, or enters the lake via the rivers.

Fish species

Economic activity. The economic importance of the lake has greatly increased during the Soviet period. Most significant is the fishing and fish breeding begun in the 1930s and now rapidly developing. A regular shipping service with a large freight turnover has developed. Of great importance to the economic development of the region was the construction of the Balkhash copper-refining plant, around which the large city of Balkhash grew on the north shore of the lake. A railway line connects Balkhash with all the major centres of Kazakhstan and Central Asia. Cattle breeding and rice growing in the lower reaches of the Ili River are of major economic importance to the region. In 1970 the Kapchagay hydroelectric power station began operations on the Ili River. As its reservoir began to fill, the regime of Lake Balkhash began to change radically. (A.V.S.)

Ballad

The ballad is a short narrative folk song whose distinctive style crystallized in Europe in the late Middle Ages and persists to the present day in communities where literacy, urban contacts, and mass media have not yet affected the habit of folk singing. The genre is not restricted to the English-speaking world: France, Denmark, Germany, Russia, Greece, and Spain possess impressive ballad collections, and at least one-third of the 300 extant English and Scottish ballads have counterparts in one or several of these continental balladries, particularly those of Scandinavia. In no two language areas, however, are the formal characteristics of the ballad identical. For example, British and American ballads are invariably rhymed and strophic (*i.e.*, divided into stanzas); the Russian ballads known as *byliny* and almost all Balkan ballads are unrhymed and unstrophic; and, though the Spanish *romances* and the Danish *viser* are alike in using assonance instead of rhyme, the Spanish ballads are generally unstrophic while the Danish are strophic, parcelled into either quatrains or couplets.

ELEMENTS

Narrative basis. Typically, the folk ballad tells a compact little story that begins eruptively at the moment when the narrative has turned decisively toward its catastrophe or resolution. Focussing on a single, climactic situation, the ballad leaves the inception of the conflict and the setting to be inferred or sketches them in hurriedly. Characterization is minimal, the characters revealing themselves in their actions or speeches; overt moral comment on the characters' behaviour is suppressed and their motivation seldom explicitly detailed. Whatever description occurs in ballads is brief and conventional; transitions between scenes are abrupt and time shifts are only vaguely indicated; crucial events and emotions are conveyed in crisp, poignant dialogue. In short, the ballad method of narration is directed toward achieving a bold, sensational, dramatic effect with purposeful starkness and abruptness. But despite the rigid economy of ballad narratives, a repertory of rhetorical devices is employed for prolonging highly charged mo-

ments in the story and thus thickening the emotional atmosphere. In the most famous of such devices, incremental repetition, a phrase or stanza is repeated several times with a slight but significant substitution at the same critical point. Suspense accumulates with each substitution, until at last the final and revelatory substitution bursts the pattern, achieving a climax and with it a release of powerful tensions.

Rhetorical devices

Then out and came the thick, thick, blood,
  Then out and came the thin,
Then out and came the bonny heart's blood,
  Where all the life lay in.

Oral transmission. Since ballads thrive among unlettered people and are freshly created from memory at each performance, they are subject to constant variation in both text and tune. Where tradition is healthy and not highly influenced by literary or other outside cultural influences, these variations keep the ballad alive by gradually bringing them into line with the style of life, beliefs, and emotional needs of the immediate folk audience. Ballad tradition, however, like all folk arts, is basically conservative, a trait that explains the references in several ballads to obsolete implements and customs, as well as the appearance of words and phrases so badly garbled as to show that the singer does not understand their meaning though he takes pleasure in their sound and respects their traditional right to a place in his song. The new versions of ballads that arise as the result of cumulative variations are no less authentic than their antecedents. A poem is fixed in its final form when published, but the printed or taped record of a ballad is representative only of its appearance in one place, in one line of tradition, and at one moment in its protean history. The first record of a ballad is not its original form, merely its earliest recorded form, and the recording of a ballad does not inhibit tradition from varying it subsequently into other shapes, because tradition preserves by re-creating rather than by exact reproduction.

COMPOSITION

Theories. How ballads are composed and set afloat in tradition has been the subject of bitter quarrels among scholars. The so-called communal school, led by two American scholars F.B. Gummere (1855–1919) and G.L. Kittredge (1860–1941), argued at first that ballads were composed collectively during the excitement of dance and song festivals. Under attack the communalists retreated to the position that although none of the extant ballads had been communally composed, the prototypical ballads that determined the style of the ballads had originated in this fashion. Their opponents the individualists, who included the British men of letters W.J. Courthope (1842–1917) and Andrew Lang (1844–1912) and the American linguist Louise Pound (1872–1958), held that each ballad was the work of an individual composer, not necessarily a folk singer, tradition serving simply as the vehicle for the oral perpetuation of his creation. According to the widely accepted communal re-creation theory, put forward by the American collector Phillips Barry (1880–1937) and the scholar G.H. Gerould (1877–1953), the ballad is conceded to be an individual composition originally. This fact is considered of little importance because the singer is not expressing himself individually, but serving as the deputy of the public voice, and because a ballad does not become a ballad until it has been accepted by the folk community and been remolded by the inevitable variations of tradition into a communal product. Ballads have also been thought to derive from art songs, intended for sophisticated audiences, which happened to filter down to a folk level and became folk song. This view, though plausible in the case of certain folk lyrics, is inapplicable to the ballads, for if the ballads were simply miscellaneous castoffs, it would not be possible to discern so clearly in them a style unlike anything in sophisticated verse.

Technique and form. Ballads are normally composed in two kinds of stanzas, a couplet of lines each with four stressed syllables, and with an interwoven refrain:

But it would have made your heart right sair,
  With a hey ho and a lillie gay
To see the bridegroom rive his haire.
  As the primrose spreads so sweetly

or a stanza of alternating lines of four stresses and three stresses, the second and fourth lines rhyming:

There lived a wife at Usher's Well,
  And a wealthy wife was she;
She had three stout and stalwart sons,
  And sent them o'er the sea.

Reference to the tunes show that the three-stress lines actually end in an implied fourth stress to match the pause in the musical phrase at these points. The interwoven refrain is a concession to the musical dimension of the ballad; it may be a set of nonsense syllables (Dillum down dillum, Fa la la la) or irrelevant rigmaroles of flowers or herbs. A few ballads have stanza-length burdens interspersed between the narrative stanzas, a technique borrowed from the medieval carols. The lyrical and incantatory effect of refrains during the ballad performance is very appealing, but in cold print they often look ridiculous, which is perhaps why early collectors failed to note them. In the first example above, it will be noted that the gaiety of the refrain is at odds with the mood of the meaningful lines. Not infrequently the ballad stanza satisfies the music's insistence on lyrical flourishes by repeating textual phrases and lines:

So he ordered the grave to be opened wide,
  And the shroud to be turned down;
And there he kissed her clay cold lips
  Till the tears came trickling down, down, down,
  Till the tears came trickling down

Kinds of repetition used in ballads

The refrain is just one of the many kinds of repetition employed in ballads. Incremental repetition, already discussed, is the structural principle on which whole ballads ("The Maid Freed From the Gallows," "Lord Randal") are organized, and many other ballads contain long exchanges of similarly patterned phrases building cumulatively toward the denouement:

"Oh what will you leave to your father dear?"
"The silver-shod steed that brought me here."
"What will you leave to your mother dear?"
"My velvet pall and my silken gear."
"What will you leave to your brother John?"
"The gallows-tree to hang him on."

Any compressed narrative of sensational happenings told at a high pitch of feeling is bound to repeat words and phrases in order to accommodate the emotion that cannot be exhausted in one saying, a tendency that accounts for such stanzas as:

Then He says to His mother, "Oh: the withy [willow], oh:
    the withy,
  The bitter withy that causes me to smart, to
    smart,
Oh: the withy, it shall be the very first tree
  That perishes at the heart."

Much repetition in ballads is mnemonic as well as dramatic. Since ballads are performed orally, the hearer cannot turn back a page to recover a vital detail that slipped by in a moment of inattention. Crucial facts in narrative, therefore, are incised in the memory by skillful repetition; instructions given in a speech are exactly repeated when the singer reports the complying action; answers follow the form of the questions that elicited them.

Conventional imagery

The exigencies of oral performance also account for the conventional stereotyped imagery of the ballads. For unlike the poet, who reaches for the individualistic, arresting figure of speech, the ballad singer seldom ventures beyond a limited stock of images and descriptive adjectives. Knights are always gallant, swords royal, water wan, and ladies gay. Whatever is red is as red as blood, roses, coral, rubies, or cherries; white is stereotyped as snow white, lily white, or milk white. Such conventions fall into place almost by reflex action, easing the strain on the singer's memory and allowing him to give his full attention to the manipulation of the story. The resulting bareness of verbal texture, however, is

more than compensated for by the dramatic rhetoric through which the narrative is projected. In any case, complex syntax and richness of language are forbidden to texts meant to be sung, for music engages too much of the hearer's attention for him to untangle an ambitious construction or relish an original image. Originality indeed, like anything else that exalts the singer, violates ballad decorum, which insists that the singer remain impersonal.

Music. A ballad is not technically a ballad unless it is sung; but though tunes and texts are dynamically interdependent, it is not unusual to find the same version of a ballad being sung to a variety of tunes of suitable rhythm and metre or to find the same tune being used for several different ballads. And just as there are clusters of versions for most ballads, so a given ballad may have associated with it a family of tunes whose members appear to be versions of a single prototypical form.

Ballad tunes are based on the modes (that is, the archaic tonal systems of medieval church music) rather than on the chromatic scales that are used in modern music. Where chromaticism is detected in American folk music, the inflected tones are derived from Negro folk practice or from learned music. Of the six modes, the preponderance of folk tunes are Ionian, Dorian, or Mixolydian; Lydian and Phrygian tunes are rare. The folk music least affected by sophisticated conditioning does not avail itself of the full seven tones that compose each of the modal scales. Instead, it exhibits gapped scales, omitting either one of the tones (hexatonic) or two of them (pentatonic). Modulation sometimes occurs in a ballad from one mode to an adjacent mode.

Most tunes consist of 16 bars with duple rhythm, or two beats per measure, prevailing slightly over triple rhythm. The tune, commensurate with the ballad stanza, is repeated as many times as there are stanzas. Unlike the "through-composed" art song, where the music is given nuances to correspond to the varying emotional colour of the content, the folk song affords little opportunity to inflect the contours of the melody. This limitation partly explains the impassive style of folk singing. Musical variation, however, is hardly less frequent than textual variation; indeed, it is almost impossible for a singer to perform a ballad exactly the same way twice. The stablest part of the tune occurs at the mid-cadence (the end of the second text line) and the final cadence (the end of the fourth line). The third phrase of the tune, corresponding to the third line of the stanza, proves statistically the most variable. Significantly, these notes happen to coincide with the rhyming words. The last note of the tune, the point of resolution and final repose, usually falls on the fundamental tone (*i.e.*, keynote) of the scale; the mid-cadence falling normally a perfect fifth above the tonic or a perfect fourth below it. To make for singability, the intervals in the melodic progression seldom involve more than three degrees. And since the singer performs solo or plays the accompanying instrument himself, he need not keep rigidly to set duration or stress but may introduce grace notes to accommodate hypermetric syllables and lengthen notes for emphasis.

Impassive style of folk singing

TYPES OF BALLADRY

The traditional folk ballad, sometimes called the Child ballad in deference to Francis Child, the scholar who compiled the definitive English collection, is the standard kind of folk ballad in English and is the type of balladry that this article is mainly concerned with. But there are peripheral kinds of ballads that must also be noticed in order to give a survey of balladry.

Minstrel ballad. Minstrels, the professional entertainers of nobles, squires, rich burghers, and clerics until the 17th century, should properly have had nothing to do with folk ballads, the self-created entertainment of the peasantry. Minstrels sometimes, however, affected the manner of folk song or remodelled established folk ballads. Child included many minstrel ballads in his collection on the ground that fragments of traditional balladry were embedded in them. The blatant style of minstrelsy marks these ballads off sharply from folk creations. In violation of the strict impersonality of the folk ballads, minstrels constantly intrude into their narratives with moralizing comments and fervent assurances that they are not lying at the very moment when they are most fabulous. The minstrels manipulate the story with coarse explicitness, begging for attention in a servile way, predicting future events in the story and promising that it will be interesting and instructive, shifting scenes obtrusively, reflecting on the characters' motives with partisan prejudice. Often their elaborate performances are parcelled out in clear-cut divisions, usually called fits or cantos, in order to forestall tedium and build up suspense by delays and piecemeal revelations. Several of the surviving minstrel pieces are poems in praise of such noble houses as the Armstrongs ("Johnie Armstrong"), the Stanleys ("The Rose of England"), and the Percys ("The Battle of Otterburn," "The Hunting of the Cheviot," "The Earl of Westmoreland"), doubtless the work of propagandists in the employ of these families. The older Robin Hood ballads are also minstrel propaganda, glorifying the virtues of the yeomanry, the small independent landowners of preindustrial England. The longer, more elaborate minstrel ballads were patently meant to be recited rather than sung.

Difference between minstrel and folk ballads

Broadside ballad. Among the earliest products of the printing press were broadsheets about the size of handbills on which were printed the text of ballads. A crude woodcut often headed the sheet, and under the title it was specified that the ballad was to be sung to the tune of some popular air. Musical notation seldom appeared on the broadsides; those who sold the ballads in the streets and at country fairs sang their wares so that anyone unfamiliar with the tune could learn it by listening a few times to the balladmonger's rendition. From the 16th century until the end of the 19th century, broadsides, known also as street ballads, stall ballads, or slip songs, were a lively commodity, providing employment for a troop of hack poets. Before the advent of newspapers, the rhymed accounts of current events provided by the broadside ballads were the chief source of spectacular news. Every sensational public happening was immediately clapped into rhyme and sold on broadsheets. Few of the topical pieces long survived the events that gave them birth, but a good number of pathetic tragedies, such as "The Children in the Wood" and broadsides about Robin Hood, Guy of Warwick, and other national heroes, remained perennial favourites. Although the broadside ballad represents the adaptation of the folk ballad to the urban scene and middle class sensibilities, the general style more closely resembles minstrelsy, only with a generous admixture of vulgarized traits borrowed from book poetry. A few folk ballads appeared on broadsheets; many ballads, however, were originally broadside ballads the folk adapted.

Broadside ballad as a source of news

Literary ballads. The earliest literary imitations of ballads were modelled on broadsides, rather than on folk ballads. In the early part of the 18th century, Jonathan Swift, who had written political broadsides in earnest, adapted the style for several jocular bagatelles. Poets such as Swift, Matthew Prior, and William Cowper in the 18th century and Thomas Hood, W.M. Thackeray, and Lewis Carroll in the 19th century made effective use of the jingling metres, forced rhymes, and unbuttoned style for humorous purposes. Lady Wardlaw's "Hardyknute" (1719), perhaps the earliest literary attempt at a folk ballad, was dishonestly passed off as a genuine product of tradition. After the publication of Thomas Percy's ballad compilation *Reliques of Ancient English Poetry* in 1765, ballad imitation enjoyed a considerable vogue, which properly belongs in the history of poetry rather than balladry.

SUBJECT MATTER

The supernatural. The finest of the ballads are deeply saturated in a mystical atmosphere imparted by the

presence of magical appearances and apparatus. "The Wife of Usher's Well" laments the death of her children so unconsolably that they return to her from the dead as revenants; "Willie's Lady" cannot be delivered of her child because of her wicked mother-in-law's spells, an enchantment broken by a beneficent household spirit; "The Great Silkie of Sule Skerry" begets upon an "earthly" woman a son, who, on attaining maturity, joins his seal father in the sea, there shortly to be killed by his mother's human husband; "Kemp Owyne" disenchants a bespelled maiden by kissing her despite her bad breath and savage looks. An encounter between a demon and a maiden occurs in "Lady Isabel and the Elf-Knight," the English counterpart of the ballads known to the Dutch-Flemish as "Herr Halewijn," to Germans as "Ulinger," to Scandinavians as "Kvindemorderen" and to the French as "Renaud le Tueur de Femme." In "The House Carpenter," a former lover (a demon in disguise) persuades a wife to forsake husband and children and come away with him, a fatal decision as it turns out. In American and in late British tradition the supernatural tends to get worked out of the ballads by being rationalized: instead of the ghost of his jilted sweetheart appearing to Sweet William of "Fair Margaret and Sweet William" as he lies in bed with his bride, it is rather the dead girl's image in a dream that kindles his fatal remorse. In addition to those ballads that turn on a supernatural occurrence, casual supernatural elements are found all through balladry.

Romantic tragedies. The separation of lovers through a misunderstanding or the opposition of relatives is perhaps the commonest ballad story. "Barbara Allen" is typical: Barbara cruelly spurns her lover because of an unintentional slight; he dies of lovesickness, she of remorse. The Freudian paradigm operates rigidly in ballads: fathers oppose the suitors of their daughters, mothers the sweethearts of their sons. Thus "The Douglas Tragedy"—the Danish "Ribold and Guldborg"—occurs when an eloping couple is overtaken by the girl's father and brothers or "Lady Maisry," pregnant by an English lord, is burned by her fanatically Scottish brother. Incest, frequent in ballads recorded before 1800 ("Lizie Wan," "The Bonny Hind"), is shunned by modern tradition.

The Freudian paradigm

Romantic comedies. The outcome of a ballad love affair is not always, though usually, tragic. But even when true love is eventually rewarded, such ballad heroines as "The Maid Freed from the Gallows" and "Fair Annie," among others, win through to happiness after such bitter trials that the price they pay seems too great. The course of romance runs hardly more smoothly in the many ballads, influenced by the cheap optimism of broadsides, where separated lovers meet without recognizing each other: the girl is told by the "stranger" of her lover's defection or death: her ensuing grief convinces him of her sincere love: he proves his identity and takes the joyful girl to wife. "The Bailiff's Daughter of Islington" is a classic of the type. Later tradition occasionally foists happy endings upon romantic tragedies: in the American "Douglas Tragedy" the lover is not slain but instead gets the irate father at his mercy and extorts a dowry from him. With marriage a consummation so eagerly sought in ballads, it is ironical that the bulk of humorous ballads deal with shrewish wives ("The Wife Wrapped in Wether's Skin") or gullible cuckolds ("Our Goodman").

Tabloid crime. Crime, and its punishment, is the theme of innumerable ballads: his sweetheart poisons "Lord Randal"; "Little Musgrave" is killed by Lord Barnard when he is discovered in bed with Lady Barnard, and the lady, too, is gorily dispatched. The murders of "Jim Fisk," Johnny of "Frankie and Johnny," and many other ballad victims are prompted by sexual jealousy. One particular variety of crime ballad, the "last goodnight," represents itself falsely to be the contrite speech of a criminal as he mounts the scaffold to be executed. A version of "Mary Hamilton" takes this form, which was a broadside device widely adopted by the folk. "Tom Dooley" and "Charles Guiteau," the scaffold confession of the assassin of Pres. James A. Garfield, are the best known American examples.

Medieval romance. Perhaps a dozen or so ballads derive from medieval romances. As in "Hind Horn" and "Thomas Rymer," only the climactic scene is excerpted for the ballad. In general, ballads from romances have not worn well in tradition because of their unpalatable fabulous elements, which the modern folk apparently regard as childish. Thus "Sir Lionel" becomes in America "Bangum and the Boar," a humorous piece to amuse children. Heterodox apocryphal legends that circulated widely in the Middle Ages are the source of almost all religious ballads, notable "Judas," "The Cherry-Tree Carol," and "The Bitter Withy." The distortion of biblical narrative is not peculiarly British: among others, the Russian ballads of Samson and Solomon, the Spanish "Pilgrim to Compostela" and the French and Catalonian ballads on the penance of Mary Magdalene reshape canonical stories radically.

Historical ballads. Historical ballads date mainly from the period 1550–1750, though a few, like "The Battle of Otterburn," celebrate events of an earlier date, in this case 1388. "The Hunting of the Cheviot," recorded about the same time and dealing with the same campaign, is better known in a late broadside version called "Chevy Chase." The details in historical ballads are usually incorrect as to fact because of faulty memory or partisan alterations, but they are valuable in reflecting folk attitudes toward the events they imperfectly report. For example, neither "The Death of Queen Jane," about one of the wives of Henry VIII, nor "The Bonny Earl of Murray" is correct in key details, but they accurately express the popular mourning for these figures. By far the largest number of ballads that can be traced to historical occurrences have to do with local skirmishes and matters of regional rather than national importance. The troubled border between England and Scotland in the 16th and early 17th century furnished opportunities for intrepid displays of loyalty, courage, and cruelty that are chronicled in such dramatic ballads as "Edom o Gordon," "The Fire of Frendraught," "Johnny Cock," "Johnie Armstrong," and "Hobie Noble." Closely analogous to these are Spanish *romances* such as "The Seven Princes of Lara," on wars between Moors and Christians.

Disaster. Sensational shipwrecks, plagues, train wrecks, mine explosions—all kinds of shocking acts of God and man—were regularly chronicled in ballads, a few of which remained in tradition, probably because of some special charm in the language or the music. The shipwreck that lies in the background of one of the most poetic of all ballads "Sir Patrick Spens" cannot be fixed, but "The Titanic," "Casey Jones," "The Wreck on the C & O," and "The Johnstown Flood" are all circumstantially based on actual events.

Outlaws and badmen. Epic and saga heroes figure prominently in Continental balladries, notable examples being the Russian Vladimir, the Spanish Cid Campeador, the Greek Digenes Akritas, and the Danish Tord of Havsgaard and Diderik. This kind of hero never appears in English and Scottish ballads. But the outlaw hero of the type of the Serbian Makro Kraljević or the Danish Marsk Stig is exactly matched by the English Robin Hood, who is the hero of some 40 ballads, most of them of minstrel or broadside provenance. His chivalrous style and generosity to the poor was imitated by later ballad highwaymen in "Dick Turpin," "Brennan on the Moor," and "Jesse James." "Henry Martyn" and "Captain Kidd" were popular pirate ballads, but the most widely sung was "The Flying Cloud," a contrite "goodnight" warning young men to avoid the curse of piracy. The fact that so many folk heroes are sadistic bullies ("Stagolee"), robbers ("Dupree"), or pathological killers ("Sam Bass," "Billy the Kid") comments on the folk's hostile attitude toward the church, constabulary, banks, and railroads. The kindly, law-abiding, devout, enduring steel driver "John Henry" is a rarity among ballad heroes.

Epic and saga heroes

Occupational ballads. A large section of balladry, especially American, deals with the hazards of such oc-

cupations as seafaring ("The Greenland Whale Fishery"), lumbering ("The Jam on Gerry's Rock"), mining ("The Avondale Mine Disaster"), herding cattle ("Little Joe the Wrangler"), and the hardships of frontier life ("The Arkansaw Traveler"). But men in these occupations sang ballads also that had nothing to do with their proper work: "The Streets of Laredo," for example, is known in lumberjack and soldier versions as well as the usual cowboy lament version, and the pirate ballad "The Flying Cloud" was much more popular in lumbermen's shanties than in forecastles.

BALLAD CHRONOLOGY

Singing stories in song, either stories composed for the occasion out of a repertory of traditional motifs or phrases or stories preserved by memory and handed down orally, is found in most primitive cultures. The ballad habit thus is unquestionably very ancient. But the ballad genre itself could not have existed in anything like its present form before about 1100. "Judas," the oldest example found in Francis James Child's exhaustive collection, *The English and Scottish Popular Ballads* (1882–98), dates from 1300, but until the 17th century ballad records are sparse indeed. As an oral art, the ballad does not need to be written down to be performed or preserved; in any case, many of the carriers of the ballad tradition are illiterate and could not make use of a written and notated ballad. The few early ballads' records survived accidentally, due to some monk's, minstrel's, or antiquary's fascination with rustic pastimes.

Precise dating almost impossible

The precise date of a ballad, therefore, or even any particular version of a ballad, is almost impossible to determine. In fact, to ask for the date of a folk ballad is to show that one misunderstands the peculiar nature of balladry. As remarked earlier, the first record of a ballad must not be assumed to be the ballad's original form; behind each recorded ballad one detects the working of tradition upon some earlier form, since a ballad does not become a ballad until it has run a course in tradition. Historical ballads would seem on the surface to be easily datable, but their origins are usually uncertain. The ballad could have arisen long after the events it describes, basing itself, as do the Russian ballads of the Kievan cycle and the Spanish ballads about the Cid, on chronicles or popular legends. It is also likely that many historical ballads developed from the revamping of earlier ballads on similar themes through the alteration of names, places, and local details.

BIBLIOGRAPHY. F.J. CHILD (ed.), *The English and Scottish Popular Ballads*, 5 vol. (1882–98), is the canon of traditional balladry; the tunes for which are supplied in B.H. BRONSON (ed.), *Traditional Tunes of the Child Ballads*, 4 vol. (1959–71). Of the many broadside collections, the most important are *The Roxburghe Ballads*, ed. by W. CHAPPELL and J.W. EBSWORTH, 9 vol. (1871–99); and *The Pepys Ballads*, ed. by H.E. ROLLINS, 8 vol. (1929–32). For a comprehensive sampling of broadsides, see *The Common Muse: An Anthology of Popular British Ballad Poetry, XVth–XXth Century*, ed. by V. DE SOLA PINTO and A.E. RODWAY (1957). Broadside tunes are gathered in C.M. SIMPSON, *The British Broadside Ballad and Its Music* (1966). T.P. COFFIN, *The British Traditional Ballad in North America*, rev. ed. (1963); and G.M. LAWS, *Native American Balladry*, rev. ed. (1964), are valuable bibliographies. Ballad criticism and scholarship are analyzed in S.B. HUSTVEDT, *Ballad Books and Ballad Men* (1930); D.K. WILGUS, *Anglo-American Folksong Scholarship Since 1898* (1959); and A.B. FRIEDMAN, *The Ballad Revival: Studies in the Influence of Popular on Sophisticated Poetry* (1961). Key modern critical essays are C.J. SHARP, *English Folk-Song: Some Conclusions* (1907); G.H. GEROULD, *Ballad of Tradition* (1932); and M.J.C. HODGART, *Ballads* (1950). A.T. QUILLER-COUCH (ed.), *The Oxford Book of Ballads* (1910, reissued 1951); M. LEACH (ed.), *The Ballad Book* (1955); and A.B. FRIEDMAN (ed.), *Folk Ballads of the English Speaking World* (1956), are the standard anthologies.

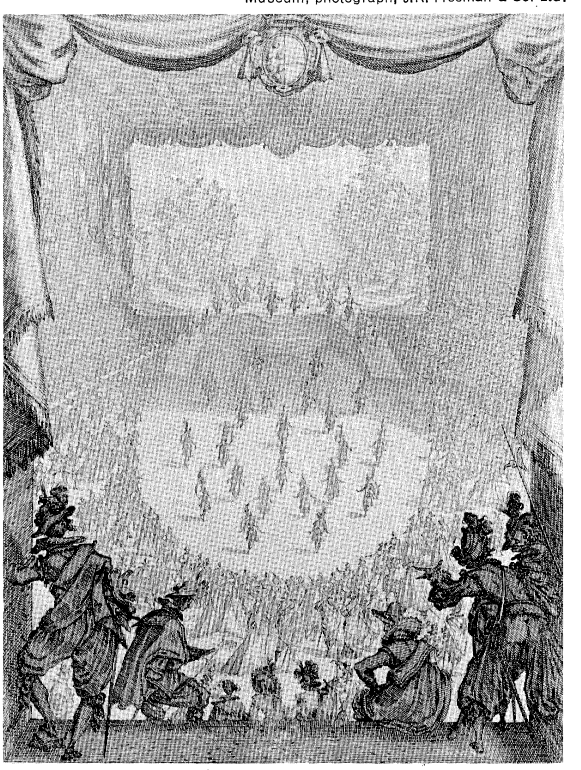
(A.B.F.)

Ballet

A theatrical art, ballet consists of extremely stylized dancing, usually with musical accompaniment and with stage settings and costumes. A ballet may, through its dance, music, and design, unfold a dramatic plot, or it may be conceived without narrative content, as a visual interpretation of the music by means of dance.

The word ballet derives from the Italian *ballare*, "to dance," but ballet involves a very special kind of dancing and of spectacle. The art of ballet had its beginnings in the 15th and 16th centuries in the spectacular dramatic entertainments of the Italian nobility, who performed these dramas themselves, usually on a theme from antiquity, combining dance, mime, song, and recitation to a musical accompaniment and with lavish settings and costumes. These extravaganzas evolved toward an emphasis on dancing, which became ballet. They sought a "perfect" form of human vocalization and movement. Over the centuries both ballet and opera refined the techniques they inherited from Renaissance court spectacles, becoming widely esteemed as the most elite forms of dancing and singing in the West. Since the 18th century, ballet has been popular outside the circles of its birth. In the 20th century, particularly, it became more widespread through extensive tours by its greatest dancers and companies, through broadened technical and thematic repertoires, and through its presentation in film and on television. It attained unprecedented popularity even where the notion of an aristocracy had become an anathema.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Ballet interlude from the spectacular Florentine opera-ballet *La liberazione di Tirreno e d'Arnea*, engraving by J. Callot after a drawing by G. Parigi, *c.* 1616. After the formal performance, the cast, composed of nobles, and the court spectators combined for dances on the main floor.

THE AESTHETICS OF BALLET

The history of ballet reveals several shifts of taste between ballet as "pure dance," or the dazzling physical execution of a repertoire of techniques, and ballet as "dramatic dance," or the presentation of a story through the stylized movements of that repertoire. Classic ballet was fully developed during the period of the romantic ballet in the 19th century, even though the terms classic and romantic are generally considered to be opposing categories in the arts. In ballet, the first applies to dance technique, the second to the subject matter of the ballets.

Like the art dance of such nations as India, Japan, and

Indonesia, ballet is based not on random exercises but on a limited collection of positions and movements that must be learned absolutely by the performer before he can add the nuances of his particular skill and personality. This technique developed much like a language. Dancers and choreographers who added to its vocabulary or enriched its subtlety of expression first mastered its existing language. It is not coincidental, therefore, that ballet was transmitted through master-to-pupil sequences.

Periods in which sheer technical virtuosity predominated were generally looked upon as decadent by ballet historians, who tended to favour dramatic expression within a narrative framework. Ballet was in such a "decadent" state when Jean-Georges Noverre, the renowned French choreographer and teacher, published in 1760 his *Lettres sur la danse, et sur la ballet*. These theories, as well as his work, were pivotal in the development of the art. Noverre called for the elimination from ballet of the meaningless gestures that had become conventions and of the leather masks worn by the dancer to indicate love, hatred, and other emotions. About the composition of ballet he wrote

Noverre's dictums

> A well-composed ballet . . . must be expressive in all its detail and speak to the soul through the eyes; . . . Steps, the ease and brilliance of their combination, equilibrium, stability, speed, lightness, precision, the opposition of the arms and legs—these form what I term the mechanism of the dance. When all these movements are not directed by genius, and when feeling and expression do not contribute their powers sufficiently to affect and interest me, I admire the skill of the human machine, I render justice to the strength and ease of movement, but it leaves me unmoved . . . Dancing is possessed of all the advantages of a beautiful language, yet it is not sufficient to know the alphabet alone. When a man of genius arranges the letters to form words, and the words to form sentences, it will cease to be dumb; it will speak with both strength and energy; and then ballets will share with the best plays the merit of affecting and moving, and of making the tears flow, and, in their less serious styles, of being able to amuse, captivate and please. (J.-G. Noverre, *Letters on Dancing and Ballets*, trans. by C.W. Beaumont, C.W. Beaumont, London, 1930.)

The basic principles that Noverre laid down were still considered valid by practitioners and critics of ballet in the late 20th century. The public for ballet, however, was by then so broad as to embrace extremes of both pure and narrative ballet.

Noverre's distinction between the means and the ends of ballet—that is, between technical virtuosity and dramatic experience—is relevant to the concepts of the classic and the romantic in ballet. The term classic ballet refers essentially to the means of all ballet since the 19th century, to the movements and positions that are its formal language—just as the metre, rhyme, number of lines, and closing couplet are an essential language of the Shakespearean sonnet. The dancer and choreographer, like the poet, provide the specific words and their sequence to spell out their thought.

Concepts of classic and romantic in ballet

The term romantic ballet, on the other hand, refers to the ballet, in favour during much of the 19th century, that dramatized tales of nymphs and other forest spirits and was an ideal expression of the ideals of the time. The romantic ballet was deeply influenced by Noverre and his direct descendants, in both technique and subject matter, and during its reign in the 19th century the dancer-choreographer-teachers Carlo Blasis and Enrico Cecchetti were among those who codified the technique of classic ballet as it came to be practiced and taught.

Apart from its theory, ballet is a practical art produced to a rigid schedule of performance by many highly and variously skilled professionals. A fuller picture of the art will emerge from a description of what each of these artists—dancer, composer, librettist, designer, and choreographer—contributes to the finished ballet.

The dancer. The dancer is both the instrument and the instrumentalist of ballet. His or her body must possess a special form of beauty, a physical beauty that an audience will find pleasing and expressive to watch and that can respond to the rigorous physical demands of the art. The height of a female dancer rarely exceeds five feet six inches (168 centimetres), and she must have an attractive and finely proportioned head well set on the shoulders and well-shaped arms and legs. Other essential features include a foot with a fine instep and toes that will serve as a platform for toe (*pointe*) dancing and the ability to turn out a 90° angle from the hips and to achieve a high extension of the legs. The male dancer, taller but rarely over six feet (183 centimetres), must have a finely built body and great strength that is not too much in evidence. He must have the grace of the efficient male athlete, a grace that is not effeminate.

There are different types of dancers. Best known are the ballerina, the pure classic dancer and the most delicate in build, and her male equivalent, the *danseur noble*. The *demi-caractère* dancer is capable of dancing soubrette or supporting roles of great variety. Finally, the "character dancer" undertakes comic or elderly roles or roles that often involve folk dance.

In addition to physical dexterity, the dancer requires a high degree of musicality, and he must be expressive from head to foot. Balletic acting is a thing apart from dramatic acting; it emerges from the dancer's translation of music into visual form as well as from his interpretation of character, period, and situation. The dancer's personality, too, is important; its quality, nearly indefinable in words, becomes obvious when the viewer senses the dancer's movements as controlled by a will rather than by a puppeteer's wires.

The art of balletic acting

The 18th-century French author Voltaire gave as one of his reasons for loving ballet that it was both a science and an art. Only after a dancer has mastered technique, the science of ballet, can he develop as an artist. Toward this mastery serious training usually begins around the age of ten or 11—not, as was once thought, at four or five. The dancer's training is at the same time both exhausting and exhaustive.

The composer. In addition to providing accompaniment to the dancers' movements, music for narrative ballet must also set the scene and provide counterpoint for the action, in effect writing the libretto in musical form. In all the most successful ballets, the score—whether especially commissioned or an existing work—corresponds inevitably to the movement on stage. Ballets have been attempted without music, but, despite their occasional success, such works are regarded largely as tours de force. Many successful ballet scores now exist in their own right as concert works.

In many long-lasting ballets, the composer and choreographer have worked closely together. Outstanding composer–choreographer relations included those of Peter Tchaikovsky with Marius Petipa in the 19th century and in the 20th century of Igor Stravinsky with Michel Fokine and George Balanchine. Working with the impresario Sergey Diaghilev in the 1920s, the French composers Francis Poulenc, Henri Sauguet, and Georges Auric wrote music that was direct and unpadded, that did not attempt to overwhelm the other elements of ballet. Music composed for an existing ballet, such as J.E. Szyfer's percussive accompaniment to Sergey Lifar's *Icare* (1935), was rare.

Among works that have used scores originally written for the concert hall are Michel Fokine's *Sylphides* (1909), which used an orchestrated version of Chopin's piano music; Léonide Massine's *Présages* (1933), to Tchaikovsky's *Fifth Symphony*, his *Choreartium* (1933), to Brahm's *Fourth Symphony*, and his *Symphonie fantastique* (1936), to Berlioz' work of that name. All were controversial but received strong support, as did *Les Sylphides*. Other creators of the so-called abstract ballet (using choreography without narrative lines) include Frederick Ashton, with *Symphonic Variations* (1946), to César Franck's music, and *Enigma Variations* (1968), based on Sir Edward Elgar's work. George Balanchine, long a specialist in this direction, produced many ballets based on existing orchestral or chamber works. The difficulty in all these cases has been to meet the choreographer's requirements without destroying the shape of the original score.

The librettist. Writers on ballet often neglect its dramatic or poetic content. The serious limitations of its narrative capacities can become sources of great strength if properly used.

Limitations of balletic communication

Ballet can speak only in the present tense. Its story must be comprehensible from the stage portrayal alone, not from program notes. Ballet can say, for example, "X did this, Y did that," but so simple a statement as "X is Y's brother-in-law" is beyond its power. In characterization, a man can only display such extreme states as hatred or love, or be jealous or generous, kind or cruel, haughty or subservient, sad or gay: the list is severely limited. Attempts to inject psychological interpretation into ballet have rarely been successful. Since the disappearance of the conventional mask and miming of early ballet, the dancer has been forced to be totally expressive within the limits of the silent medium; and the librettist, communicating only through the movement of the dancer and the music, is in constant danger of being either obvious or obscure.

Ballet can be, however, an art of great subtlety, whether telling a story or suggesting a mood or an atmosphere. The typical romantic ballet of the 19th century told a dramatic narrative, as in *Giselle*, whereas the neoromantic ballets of the early 20th century suggested an aura of romance through, for example, the suite of dances making up *Les Sylphides*. A narrative ballet such as *Petrushka* (1911) can be regarded as a simple child's story, the awakening of a soul, or even a hidden Marxist message. Its creators, Stravinsky and Fokine, repeatedly refused to offer definitive explanations of its "meaning."

The term abstract ballet is much used to describe ballet that is based solely on music rather than on a narrative line. The term is inherently ambiguous, however, for the dancers being human, every *pas de deux* (dance for two performers) may be interpreted as a courtship dance. It is best applied, perhaps, only to the kind of purely rhythmic exercise exploited in certain types of classic Indian dance, like the *bhārata-nāṭyam*.

In the 20th century, the role of the librettist has become less and less significant as choreographers have become more and more inclined to devise their own scenarios. His is now the total conception, and something may have been lost through the exclusion of the poet from the balletic domain.

The designer. The designer must understand the dance intimately. His costumes must be of the correct weight and balance for the steps to be performed, and he must plan the effect of light on materials and the mixture of colours in the choreographer's groupings. Finally, he must be in sympathy with the music, for colour and sound can easily counteract one another. The designer–composer alliance between Marie Laurencin and Poulenc in Bronisława Nijinska's *Biches* (1924), between Pablo Picasso and Manuel de Falla in Massine's *Three-Cornered Hat* (1919), and between Georges Rouault and Prokofiev in Balanchine's *Prodigal Son* (1929) was, in each case, a total artistic collaboration. Many ballets of the recent past and today have no scenery. Many of Balanchine's works, for instance, depend solely on lighting to set the mood.

Contribution of dancers to costuming

Dancers themselves contributed much to the evolution of ballet costuming. The 18th-century ballerina Marie Camargo altered the earlier long, cumbersome skirts to make leg movement easier as well as more visible. In 1729, Marie Sallé and her partner danced without the leather masks that had been customary, but only in the 1770s, under Noverre's influence, was the mask finally abandoned. Around 1838 dancing tights were invented. Although they allowed maximum freedom of movement, they were considered immoral and their acceptance was not immediate. Marie Taglioni was among the earliest dancers to use the shoes with blocked toes, invented around 1820, that permitted *pointe* dancing. In *Giselle* she introduced the midcalf-length skirt that was the prototype for the very short tutu devised by Italian ballerinas in the 1880s.

The choreographer. The choreographer is the direct descendant of the dancing master of the Renaissance, who refined the peasant dances for the courts and taught the nobles grace and style in their movements. The choreographer also recalls early ballet masters like Pierre Beauchamp, who worked with Molière and the composer Jean-Baptiste Lully to produce ballets for the court of Louis XIV of France. In the 20th century his role has become a little like that of the stage or film director, although the choreographer must first fashion what he is going to direct.

Little has been written about the science or art of choreography since 1760, when Noverre wrote his *Lettres* that outraged the leadership of the Paris Opéra, then the centre of European ballet. A seminar on choreography held in Moscow in 1969 confirmed the lack of universally agreed upon standards, but certain attributes have found general acceptance.

Nearly without exception, the choreographer is or has been a dancer, for he must be thoroughly familiar with the vocabulary of classic ballet and with the potentialities and limitations of the human body. Some choreographers collaborate with their dancers to develop a work, others present the fully conceived work almost from the first rehearsal. In either case they must work with dancers each of whom has a distinctive will, personality, and set of performing characteristics. Any important change of cast may significantly alter the work. Even if dance notations were universally accepted, therefore, it would neither alter this situation significantly nor eliminate the need for constant polishing of performances by the choreographer.

Collaboration of choreographer and dancers

The choreographer must also be a well-grounded musician. He must be familiar with the arts and styles of many cultures, primitive and advanced, contemporary and historical. In working with music he must be aware that the eye and the ear respond at different speeds. Petipa's notebooks reveal how close a working relationship he established with Tchaikovsky, who during the choreographing of *The Sleeping Beauty* sat at the piano indicating the exact length of passages.

The rules of composition pertaining in the visual arts govern its choreography: there is constant interaction of foreground and background, of colour and line, and of moving dancers and immobile scenic design. Even static moments in the dance must be expressive, not simply breaks in movement. Ballet choreography requires realism within the limits of its conventions, but stylization is essential. If a river has been established at one part of the stage, the dancers must not walk but make swimming movements at that place. On the other hand, however, if a ballet is set in China, the dancers may use *pointe* dancing even though it is not Chinese.

Finally, the subject selected by the choreographer must be effective theatre and capable of better expression through ballet than any other theatrical medium. If ballet uses subjects from other theatrical forms, it must find original perspectives in them and translate them radically to fit the conventions and means available to the dancer.

HISTORICAL DEVELOPMENT

Origins in the Renaissance. The visual arts of the Renaissance in Italy strove for an idealization of the human body. The courts of the nobility, such as those of the Medici in Florence, vied to outdo one another in their magnificence.

This desire for opulence and the quest for idealization were combined to produce spectacular entertainments for distinguished guests, often several hours in length and consisting of a series of entrées, or pageants of music, poetry (sung or declaimed), pantomime, and dancing hung loosely upon a classical theme. Costumes were lavish, and the settings, whether in a garden or in the centre of the banquet hall, were splendidly designed. One such spectacle, produced in Tortona in 1489, has been called the first actual ballet. Between courses of the feast, members of the nobility presented the story of Jason and the Golden Fleece; the dances were based on the formal styles of the court.

Renaissance ballets

Another candidate for designation as the first ballet was the *Ballet comique de la reine* (1581), presented by Cath-

erine de Médicis, who had brought her musicians from Italy when she became queen of France. The *Ballet comique* told of Ulysses' escape from Circe's wiles, a mythological source typical of ballet librettos for centuries. The word *comique* referred to the gaiety of the occasion (a betrothal) and to the dance's happy ending. The ballet was devised by Catherine's director of court festivals, Baltazarini di Belgioioso, an Italian known in France as Balthazar de Beaujoyeulx. He left a long account of the work, which he referred to as "a geometrical arrangement of many persons dancing together under a diverse harmony of instruments."

By courtesy of (top) the Bibliotheque Nationale, Paris; (bottom) the National Portrait Gallery, London





Renaissance court ballet.
(Top) The prologue from the *Ballet comique de la reine*, choreographer, Balthazar de Beaujoyeulx, engraving by J. Patin, 1581. Three-dimensional sets representing Circe's garden and palace (arranged around the ballroom) moved on rollers in order to be seen from at least three sides. (Bottom) English masque as seen in the portrait of Sir Henry Unton, oil on a panel by an unknown artist, *c.* 1596. In the National Portrait Gallery, London.

This statement might serve as a partial definition of choreography at the time. It was a period in which the idealizations of the earlier Renaissance assumed in the visual arts an extreme stylization that became known as mannerism. Ballet, certainly, was mannered. Belgioioso's stress on geometry was of interest. The long, cumbersome costumes limited the dancers, and without the elevations, toe steps, and leaps developed by future generations, attention was focussed on the ground patterns.

The *Ballet comique* cost 3,600,000 francs and represented a great change from the banquet intermezzi. For the next century such court entertainments were danced by aristocratic amateurs under the direction of ballet masters. Louis XIII injected bawdy humour into the productions. In England, the Italian masquerades introduced during the reign of Henry VIII evolved into the masque. This form was closely akin to the French ballet; its master librettist, the poet and dramatist Ben Jonson, owned a copy of the *Ballet comique*. In 1632 performances were opened for the first time to the general public.

17th century. In France, the building of the Palais Cardinal (1636) changed ballet history. The dance moved from the ballroom to the theatre, and both the audience's angle of vision and the placement of stage scenery were permanently altered.

Louis XIV's influence on ballet

A number of circumstances turned an aristocratic pastime into a profession. The young Louis XIV, the future "Sun King," was bored with opera, for he loved dancing and at 13 had performed a leading role. He gathered about him outstanding artists from many mediums: the dramatist Molière, the poet Isaac de Benserade, the composer Jean-Baptiste Lully, and his own dancing master and intimate, Pierre Beauchamp. Molière demanded complete devotion to dancing, but he was often forced to introduce awkward intermissions for costume changes so he could use his few trained dancers in several roles. In 1661, Louis created the Académie Royale de Danse, comprising the dancing masters of the nobility with Beauchamp as chairman.

The preamble to the letters patent of the Académie's foundation, written by Louis, was the first full recognition of the art of ballet.

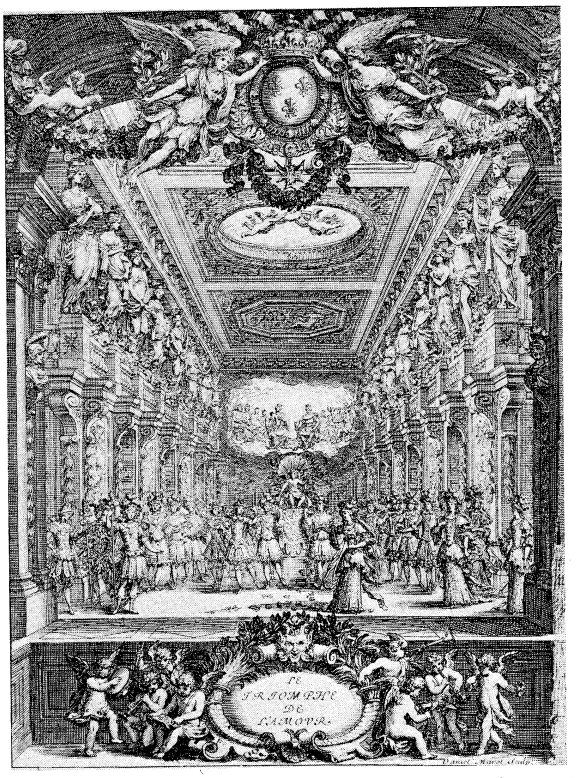
> Although the art of dancing has always been recognized as one of the most honourable and most necessary for the training of the body . . . many ignorant persons have tried to disfigure and spoil it . . . so that we see few among those of our court and suite who would be able to take part in our ballets . . . Wishing to establish the said art in its perfection and to increase it as much as possible, we deemed it opportune to establish in our good town of Paris a Royal Academy of Dancing.

The members of the new Académie, most of them illiterate, left no written records, but they systematized the techniques recognized at that time. The success of the undertaking prompted Louis to set up the Académie Royale de Musique in 1669 and a dancing school in 1671. These institutions, combined as the Académie de Musique and known as the Paris Opéra, were the major cradle of ballet and remained a centre of ballet production through succeeding centuries.

Innovations of the Académie

The school rapidly brought about many changes. Originally the ladies of the court took part only as a decorative background; as in the Elizabethan theatre, female roles were performed by men. Women first appeared professionally in Lully's *Triumph of Love* (1681); its ballerina, Mlle Lafontaine, was quickly acclaimed "queen of the dance." There was also an extension of balletic techniques. Pierre Beauchamp, who had danced in *The Triumph*, clarified styles of turns and elevation and codified the five positions of the feet that remained basic to ballet.

The form of ballet also received attention. The French ballet was still a mixture of dance and music with singing and declamation, whereas the English masque had taken the path of poetic drama, and in Italy singing predominated. In 1708, at a fete given by the Duchess du Maine, Françoise Prévost and Jean Balon (Ballon) danced and mimed without words the last act of *Les Horaces*, based on Pierre Corneille's tragedy *Horace;* the audience is said



Pierre Beauchamp's *The Triumph of Love*, 1681, with set and costumes by Jean Berain, and music by Jean-Baptiste Lully, engraving by Daniel Marot (1663–1752). Originally Beauchamp danced the feminine role in a *pas de deux* with Louis XIV, but later gave his part to the first ballerina, Mlle Lafontaine.
By courtesy of the Bibliotheque Nationale, Paris

to have been moved to tears. At about the same time, John Weaver was making similar experiments in dramatic dancing in England. Probably influenced by the pantomimes of John Rich, a famous Harlequin and rival producer, Weaver staged such narrative ballets as *The Cheats of Scapin, or, The Tavern Bilkers* (1702) and *The Loves of Mars and Venus* (1717).

18th century. The rivalry of Marie Camargo and Marie Sallé in the first half of the 18th century increased public interest in the dance. Both ballerinas altered ballet technically and aesthetically. The shortened skirts and heelless slippers of "la Camargo" permitted leg extensions not previously possible, as well as the perfection of the *entrechat*, a leap during which the feet are crossed rapidly. Her favourite vehicle was the plotless divertissement, a series of solo minuets, courantes, sarabands, gigues, and other dances in contrasting styles.

In contrast to Camargo's balletic agility, Marie Sallé emphasized the dramatic potentialities of the dance. She studied with Rich in London and, in 1734, danced in the dramatic ballet *Pygmalion* with a loosely draped costume and flowing hair. Sallé's attempts to unify music, costume, and dance, as well as her interest in the dramatic, anticipated Noverre's reforms by a quarter of a century.

Camargo's dances of purely decorative beauty and virtuosity and Sallé's dances seeking the utmost dramatic expression were typical of the polarities in 18th-century ballet. The dancers at the Paris Opéra, including such brilliant performers as the three Dumoulin brothers, Jean-Barthélemy Lany, and Louis Dupré, continued to perfect their techniques. Appearances by Italian acrobats and comic dancers in Paris further stimulated an emphasis on physical display. Gaetano Vestris perfected the broad, high leap, or *grand jeté*, and elaborated on the multiple turn known as the *pirouette*. By 1800 male dancers were executing spectacular leaps and series of *pirouettes*.

Technical virtuosity of 18th-century dancers

This dazzling virtuosity presented a danger noted as early as 1741 by St. Mard in his *Reflections sur l'Opéra:* the dancers, he complained, lacked both variety and intelligence and were like "cardboard that is made to move like machines." Despite such strictures and the work of such ballet masters as Jean Baptiste de Hesse, who staged pantomimic ballets at the Théâtre Italien, Paris, and trained his students as dance actors, the Opéra became by the middle of the 18th century a bastion of technical purity. Its dancers offered entertainment without involvement.

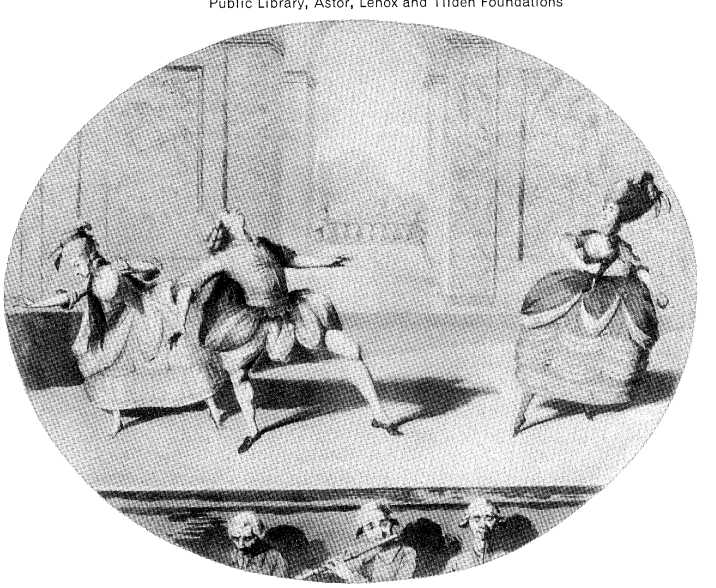
Noverre, the teacher of a generation of great dancers, published his *Lettres* in 1760. His ideas were rejected at the Opéra, and he spent 16 years choreographing in Stuttgart, London, Vienna, and Milan. A curious Gaetano Vestris observed him in Stuttgart, danced in his ballets, and brought a version of his *Medée et Jason* to the Opéra in 1770. When Vestris discarded the leather mask as Jason, other dancers soon followed. Despite opposition, Noverre's ideas prevailed and he was called to the Opéra in 1776. There he found an ideal interpreter of his *ballets d'action* in the ballerina Madeleine Guimard. His pupils Jean Dauberval and Charles Le Picq carried his ideas throughout Europe. In 1789 Dauberval produced *La Fille mal gardée*, which was still in the ballet repertoire of the late 20th century and, next to Vincenzo Galeoti's *Whims of Cupid and the Ballet Master* (1786), was the oldest work in performance.

Although Noverre was the most influential innovator, the Italians Gasparo Angiolini and Salvatore Viganò also furthered dramatic, narrative ballet. Angiolini, who concurred with Noverre's theories but thought his ballets obscure, collaborated with the German composer Christoph Willibald Gluck in the ballets *Don Juan* (1761) and *Semiramide* (1765) and choreographed the dances for the opera *Orfeo ed Euridice* (1762). Gluck's experiments in opera were further aided by work with Noverre himself on *Iphighénie en Tauride* (1779). Viganò, a nephew of the composer Luigi Boccherini and a student of Dauberval, worked mainly at La Scala theatre in Milan, where he built the roles of his dancers in the manner of a stage director, assigning them individual characteristics. Typical of his heroic works, which were designed on a grand scale, was *Otello* (1818). For Viganò Beethoven composed his only ballet score, *The Creatures of Prometheus* (1801).

Romantic and classical traditions in the 19th century. The Romantic spirit found in the arts of the early 19th century was expressed in the dance with the production of *La Sylphide* at the Paris Opéra in 1832. This ballet, danced by Marie Taglioni and choreographed by her father, Filippo Taglioni, permanently changed the art and

The "golden age of ballet"

By courtesy of the Dance Collection, the New York Public Library, Astor, Lenox and Tilden Foundations



Medée et Jason, choreography by Gaetano Vestris in 1770, adapted from the original ballet by Jean-Georges Noverre, engraving by John Boydell, 1782. Vestris as Jason (centre), Adelaide Simonet as Medea (right), and Giovanna Baccelli as Creusa (left) appear without the traditional leather masks in Boydell's caricature of the tragic ballet.

ushered in what has been called the "golden age of ballet."

"After *La Sylphide*," wrote the French critic Théophile Gautier,

> . . . the Opéra was given over to gnomes, undines, salamanders, nixes, wilis, peris—to all that strange and mysterious folk who lend themselves so marvellously to the fantasies of the ballet master.

Greek mythology and sunlight gave way to the spirit-haunted, moonlit, and mist-covered forests of Europe. Costuming underwent rapid change to accommodate the new subjects and the seemingly disembodied dancing of Taglioni and her successors, and the music came into closer rapport with the action. The most drastic innovation of the age was dancing *sur les pointes*, on the tips of the toes, essential for the portrayal of sylphs, naiads, and bird-women.

Gautier, who became high priest of the movement, condemned the male dancer as essentially coarse and unromantic. With some notable exceptions the male was virtually swept from the ballet stage outside Russia and Denmark. Consequently, audiences of the period watched ballerinas, many of whom became legends. Taglioni and Fanny Elssler reproduced the Camargo–Sallé rivalry of a century earlier. Carlotta Grisi created the title role in *Giselle* (1841), usually considered, with *La Sylphide* and Saint-Léon's *Coppélia* (1870), the finest achievement of the romantic ballet outside Russia. In 1845, together with the equally renowned dancers Fanny Cerrito and Lucile Grahn, Taglioni and Grisi appeared in London in Perrot's *Pas de quatre*, a gathering of brilliance rare in any period.

By courtesy of the Theatrical Museum of La Scala, Milan



Pas de quatre, choreography by Jules Perrot, 1845, with music by Cesare Pugni. The ballerinas are (left to right) Carlotta Grisi, Marie Taglioni, Lucile Grahn, and Fanny Cerrito; Fanny Elssler, Taglioni's great rival, did not appear.

Contributions of Carlo Blasis

In Italy, Carlo Blasis, a prolific writer as well as a choreographer, codified the techniques of classic ballet in his *Elementary Treatise Upon the Theory and Practice of the Art of Dancing*, published in 1820. Ten years later, his *Code of Terpsichore* advised young dancers of the rigours demanded by the dance and of the vitality their work must exhibit. His school at La Scala, which he conducted from 1837, attracted both students and such famous dancers as Cerrito and Grisi.

Ballet historians regard the last third of the 19th century as a period of decadence in the art, much like the one that Noverre had first denounced and then transformed. The romantic subject matter, new in *La Sylphide* and *Giselle*, was endlessly repeated, as Greek mythology had been before. The replacement of square-blocked shoes by the satin slippers made ballet and toe-dancing synonymous, and the neglect of ensemble dancing for the brilliant virtuosity of the ballerina deprived ballet of artistic unity.

Only Denmark and Russia were immune from these excesses. Antonio Sacco and Vincenzo Galeotti had established dramatic ballet on the Danish stage before 1800, but the prime architect of the Royal Danish Ballet was Auguste Bournonville. A native Dane, Bournonville in 1829 brought to Copenhagen the preromantic style of the Paris Opéra, which emphasized masculine virtuosity and expressive mime. The tradition was meticulously preserved by Bournonville and his successors.

Early history of Russian ballet

On the continent of Europe, however, empty formulas might have destroyed ballet as a serious art had it not been for the Russian ballet. The westernization of Russia begun by Peter the Great in the early 18th century brought the European court dance to the newly founded capital city of St. Petersburg. The old-guard nobles, who held one masquerade lasting a week, were compelled by the Emperor to modify their costumes and dance the minuet with the court ladies. The early history of French ballet was repeated, and again it was only a short step from court to theatre. In 1734 the empress Anna brought Jean-Baptiste Landé from France to organize a school, "to give instruction to the young people under his charge, to teach them with honesty, sincerity, seriousness and all the qualities of a good man." His pupils were the children of the poor.

By 1740 the Imperial School of Ballet was established at the Winter Palace, and Russian soloists appeared in a fully established company. Such teachers and choreographers as Franz Hilverding—Noverre's successor at Stuttgart—Angiolini, and Le Picq were brought to Russia by Catherine the Great and Alexander I. The half million rubles expended on the scenery for one ballet suggested the exuberance of the Russian balletomane. Ivan Valberg became the first Russian choreographer, but it was the Frenchman Charles Didelot who, after 1801, created the grand spectacles that remained hallmarks of Russian ballet. The Taglionis as well as Elssler, Cerrito, Grahn, and Grisi appeared in St. Petersburg. There is some evidence that the ballerina Mariya Danilova, who died at 17, was the first Russian to use *pointes*.

The vigour of the Russian ballet, even throughout the period of decline elsewhere, was attributable in some part to its assimilation of native Russian materials. Apart from the imperial companies, nobles sponsored companies composed of their serfs. To them, the Russian folk dance, not the foreign, was the living reality. The Imperial Russian Ballet constantly renewed itself, by purchase and legacy, from these companies. In 1829 the Moscow ballet, now the Bolshoi Ballet, founded in 1825, acquired 21 dancers from a Mr. Rzhevsky, and after 1824 serfs of the imperial theatres were emancipated, nearly 50 years before emancipation became general.

The allegation that Russian ballet was the product of foreigners thus requires modification, for it was very Russian in character from the start of the 19th century. The indigenous dance influenced the foreign teachers themselves, most notably the Russianized Auguste Poirot, who spent 30 years studying Russian dance and incorporating it into ballet. The ballerina Ylena Andreyanova, who rivalled Taglioni when she danced in Paris and Milan, lent a special quality to *Giselle* in 1842. Gautier praised the serious outlook of the Russian dancers. Critics pointed to the softening influence of the Russian temperament on the brilliant but sometimes brittle Italian dancers.

Petipa's influence on the Russian ballet

The major foreign influence on Russian ballet came from the Frenchman Marius Petipa, the Swede Christian Johansson, and the Italian Enrico Cecchetti. Petipa, whose brother Lucien was one of the few outstanding male dancers of the romantic period, came to Russia in 1847. He became virtual dictator of Russian ballet from his appointment as choreographer in 1862 until his return to France in 1910. His output included 57 full-length ballets, 34 opera ballets, and 17 revivals. His greatest triumphs were the Tchaikovsky ballets, although some of

their most effective passages were the work of a Russian, Lev Ivanov. Petipa's original choreography of *Swan Lake* (1895)—Act II was Ivanov's work—and *The Sleeping Beauty* (1890), and his versions of *Giselle* and *Coppélia* remained in the repertoire of most ballet companies in the U.S.S.R. and elsewhere.

Christian Johansson, trained by Bournonville, came to St. Petersburg in 1841 as a dancer. His major contribution was as principal teacher at the Imperial School of Ballet, where he formed an entire generation of ballerinas. The international character of ballet is nowhere so clear as in the case of Cecchetti. He first visited Russia in 1874 and returned 13 years later with an Italian company for a summer of dancing. The technical brilliance of his dancers dazzled the Russians, who were accustomed to the gentler grace of their Franco-Russian ballet. Pierina Legnani created the role of Odette in *Swan Lake*, and Cecchetti himself danced the bluebird in *The Sleeping Beauty*. He remained for 15 years as dancer and teacher before going on to Warsaw and London.

Early 20th century. When the initial impetus for a great movement slows down, innovation may cease and inspiration may become a formula. This occurred toward the end of Petipa's reign, certainly by 1903, when his *Magic Mirror* was a failure. The audiences of St. Petersburg were connoisseurs of the dance, critical of each step but no longer interested in choreography.

The Diaghilev period. A new direction was provided by Sergey Diaghilev, a man who was neither dancer, composer, poet, designer, nor choreographer. As a law student in St. Petersburg, Diaghilev became part of a group of avant-garde writers, painters, and musicians who called themselves "the Pickwickians of the Neva." They aimed to take the best in the Russian arts to the West and to overturn the academicism that insisted that every picture tell a story. Not at first concerned with ballet, they were redirected by two events. The first was the dancing of the Italian Virginia Zucchi, which demonstrated to "Pickwickian" Alexander Benois the potentiality of a great dancer given the right work. The second was the appearance in Russia of the American Isadora Duncan, dancing in flowing costumes to the music of the great composers. She aroused fierce controversy between the connoisseurs and the reformers. Diaghilev led ballet on a middle path, and Duncan became a major force behind what came to be known as modern dance.

Role of Isadora Duncan

In 1909, Diaghilev organized a troupe of Russian dancers to present a season in Paris. As choreographer he chose Michel Fokine, whose student work had been acclaimed but who found himself denied the limelight that Petipa still commanded. As early as 1904 Fokine had restated Noverre's principles and elaborated them. Ballet would no longer be broken into separated dances but would flow without interruption, like Wagnerian opera. Dancing would vary in style with the demands of the theme and be the sole means of expression, disdaining conventionalized mime. Most important, the several arts of ballet were to be equal partners and the style was to be consistent in all elements. As a result, the full-length ballet, with its padding and its dances interpolated to show off a particular ballerina, was abandoned.

The French public was accustomed to isolated stars and to buxom women dancing male roles, but not to the homogeneous ensemble dancing of the Polovtsian warriors in the fierce *Prince Igor* nor to the reintroduction of the leading male dancer by Vaslav Nijinsky in *Petrushka*. This first phase of Diaghilev's life as impresario consisted of the neoromantic choreography of Fokine in those and such other works as *The Firebird* (1910), *Scheherazade* (1910), and *Le Spectre de la rose* (1911). All of them were intensely Russian and had the scenic design of Alexandre Benois, Léon Bakst, and others.

Dancing and choreography of Nijinsky

A bridge in Diaghilev's career occurred when Nijinsky became a choreographer as well as a dancer. The scandal of his *Afternoon of a Faun* (1912) was dwarfed by that of *The Rite of Spring* (1913), when Stravinsky's music, considered incomprehensible, caused a fashionable Parisian audience to riot.

Following the disruptions of World War I and tours through America and southern Europe, the choreography of Léonide Massine provided a new aesthetic direction. Varying national styles assimilated in the travels of the Diaghilev troupe appeared in such Massine works as *The Three-Cornered Hat* (1919) and *Pulcinella* (1920). Cut off from Russia by the Revolution, Diaghilev made Paris his artistic centre, commissioning designs from such painters as Picasso, Rouault, Matisse, and Derain.



Original design by Léon Bakst for *Scheherazade*, 1910. In the Musée des Arts Décoratifs, Paris. The other artists contributing to this Diaghilev production were: Michel Fokine, choreographer; Nikolay Rimsky-Korsakov, music; Alexandre Benois, book; Ida Rubinstein, Vaslav Nijinsky, and Enrico Cecchetti, principal dancers.
Marc Garanger

Diaghilev animated a group and reformed an art; he had a counterpart who popularized ballet and inspired a generation of dancers. The solitary genius Anna Pavlova left Russia in 1907 and toured her company over the globe. Tirelessly dedicated to the dance, she performed on stages large and small, good and bad, in villages and capitals, bringing ballet to people who had never suspected its existence. It has been said that she, rather than Diaghilev, was the guardian of the traditions of ballet.

(A.L.H.)

Ballet since Diaghilev. The immediate heirs of Diaghilev were René Blum, who was in charge of the ballet in Monte-Carlo from Diaghilev's death in 1929 until the Nazi invasion of France in 1940, and Colonel W. de Basil, who formed a rival Russian ballet company in Paris, went into partnership with Blum in 1932, and later separated from him again. From the early 1930s until after World War II, these and various other allegedly Ballets Russes companies toured the world. The Blum–de Basil company first visited the United States in 1933, and a rival company run by Sergey Denham made its home there during and after the war. Such Diaghilev choreographers as Fokine and Massine worked with these companies and new Russian dancers emerged. In particular the de Basil company became famous for its teenage "baby" ballerinas—Irina Baronova, Tatyana Ryabushinska, and Tamara Toumanova—who were all trained by the former St. Petersburg ballerina Olga Preobrajenska in Paris. The Ballets Russes were instrumental in bringing ballet to many new audiences and in keeping the love of ballet alive. One of de Basil's dancers, Édouard Borovansky, eventually settled in Australia and started the first ballet company there. After his death in 1959 the company was run by Dame Peggy Van Praagh, who was later joined by Sir Robert Murray Helpmann.

The days of privately run touring ballet companies were, however, numbered, and the more influential heirs of Diaghilev were those who created or directed permanent national companies in London, Paris, and New York. It took all the genius of a Diaghilev to maintain creativity and artistic integrity without a permanent home or a public subsidy; recent imitators, including the late Marquis de Cuevas in France and Rebekah Hark-

ness in the United States, though they succeeded in running popular companies at their own expense, did not contribute anything lasting to the art of ballet.

The French national ballet

The French national ballet, which had fallen into decline, received a new lease of life from Sergey Lifar, one of Diaghilev's stars, who was principal dancer, choreographer, and artistic director at the Paris Opéra from 1932 until 1958. His choreography did not on the whole find much favour outside France, but during his reign ballet gained an important status in Paris. A large number of dancers trained at the Opéra became international stars, in some cases going on to form companies of their own. Roland Petit, for example, left the Opéra in 1944 to form Les Ballets des Champs-Elysées. For some time he was regarded as the most important young choreographer in western Europe, producing *Les Forains* (1945), *Le Jeune Homme et la mort* (1946), and *Carmen* (1949), among other works. Unfortunately, French ballet went back into decline toward the end of Lifar's regime.

Great Britain and the United States. In Britain and the United States, on the other hand, ballet can be said virtually to have been born after the death of Diaghilev. British ballet was largely the creation of two remarkable women, Dame Marie Rambert and Dame Ninette de Valois. Neither was English—Rambert was born in Poland and de Valois in Ireland. Both had served their apprenticeships with Diaghilev and benefitted from the appetite for ballet he had created, though they suffered at first from the public's belief that ballet must be Russian. Rambert did not have the ability, or maybe the will, to create a large organization, but her Ballet Rambert nurtured the two most important choreographers in the first decade of British ballet, Frederick Ashton and Antony Tudor, and many of the dancers she trained went on to become stars elsewhere. De Valois had the apparently hopeless vision of creating a British national ballet. Later, she had the satisfaction of watching the tiny Vic-Wells Ballet she started in 1931 grow into the Sadler's Wells Ballet and then into the present Royal Ballet, based at the Royal Opera House, Covent Garden, London, and recognized as one of the world's major companies. The Vic-Wells started with British dancers inherited from Diaghilev, notably Dame Alicia Markova and Anton Dolin; de Valois soon acquired the services as choreographer of Frederick Ashton, and she also did choreography for the company herself, her *Rake's Progress* (1935) being an outstanding work in a truly English idiom. Tudor spent most of his later career in the United States, though his early British ballet *Lilac Garden* (produced for Marie Rambert's Ballet Club, 1936) is still his most successful and best known work.

Balanchine's role in American ballet

In the United States, another Diaghilev protégé, George Balanchine, founded the School of American Ballet in 1934. Since then, in close association with Lincoln Kirstein, a businessman, he has directed various ballet companies, culminating in the New York City Ballet, formed in 1948. Just as the Sadler's Wells Ballet achieved international recognition when it appeared at the Metropolitan Opera House, New York, in 1949 on the first of what were to become almost annual visits, so New York City Ballet took its place among the world's leading companies when it appeared in London at Covent Garden in 1950. It acquired a permanent home in 1964 at the new New York State Theater in New York City's Lincoln Center for the Performing Arts.

The distinctive styles of the Royal Ballet and of New York City Ballet are the result of years of work with their respective resident choreographers, Ashton and Balanchine. The other major American company, Ballet Theatre (now the American Ballet Theatre), founded in 1939 by Lucia Chase and Richard Pleasant, has always had a more eclectic repertoire and a wider range of styles, making it closer to the Ballets Russes tradition. During and immediately after World War II, Ballet Theatre was the most important American company, and it was the first one to dance at Covent Garden and in the U.S.S.R. But it never had a permanent home and always depended on private finance. In 1971 it became one of the official companies in residence at the new John F. Kennedy Center for the Performing Arts in Washington, D.C.

Not only have Ashton and Balanchine created vast and varied repertoires for their own companies; they have also influenced the whole development of ballet in the Western world. Ashton's style might be described as a romantic approach to classicism. His most popular work, *La Fille mal gardée*, is a re-creation of a French ballet first staged in 1789. He has also made three-act works to Prokofiev's scores for *Cinderella* and *Romeo and Juliet* and to Hans Werner Henze's especially commissioned *Ondine*. Among his most successful short ballets have been abstract works like *Symphonic Variations* and *Monotones*, lighthearted divertissements like *Les Rendezvous, Les Patineurs*, and *Jazz Calendar*, and narrative ballets like *Daphnis and Chloë* and *Enigma Variations*. He played a crucial part in the development of Margot Fonteyn, the first ballerina to emerge from within de Valois' organization, and he created many roles with lyrical and technically demanding *pas de deux* for her and her successive partners, Sir Robert Helpmann, Michael Somes, and Rudolf Nureyev.

Ashton's choreographic style

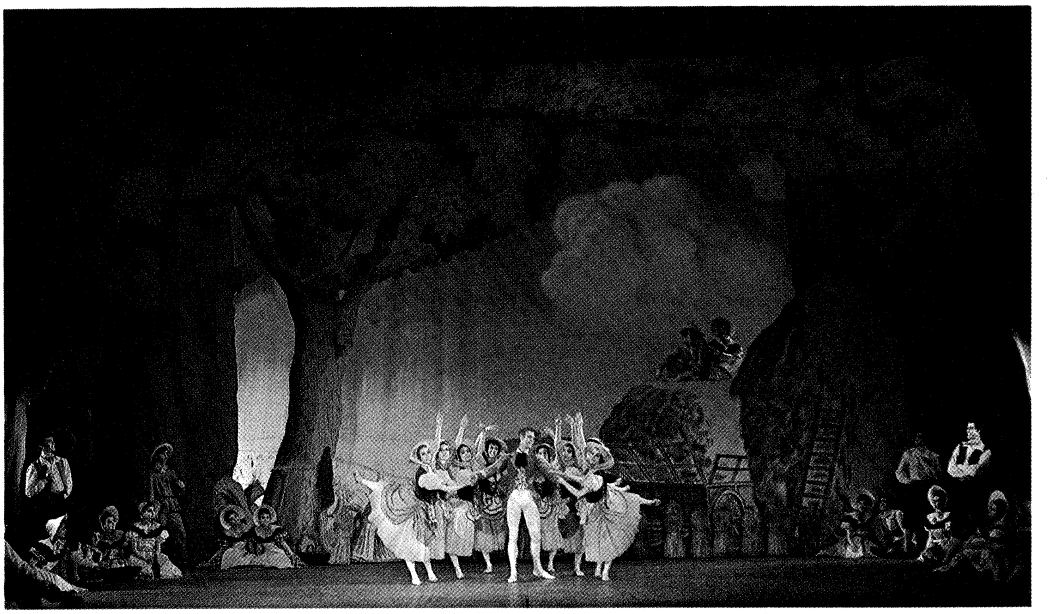
Ashton's success in making three-act ballets, which had not previously been attempted by Western choreographers, influenced the younger British choreographers John Cranko and Kenneth MacMillan to follow the same path. Cranko made *The Prince of the Pagodas* to an especially commissioned score by Benjamin Britten, and in Stuttgart, where he became director of the ballet in 1961, he created several full-length ballets. MacMillan created *Romeo and Juliet* and *Anastasia* for the Royal Ballet, of which he succeeded Ashton as director in 1970.

The creation of new full-length works, and new versions of 19th-century classics, is one of the surprising developments in ballet since the early 1950s. The other is the great popularity of abstract, pure dance works, largely inspired by Balanchine. (Nineteenth-century classics and the idea of dance for its own sake were once thought to have been killed by the Diaghilev–Fokine revolution.) Although Balanchine has created full-length story ballets, he is mainly associated with plotless works in which dancers simply move to music. His genius lies in his ability to make every sort of music—from Bach, Mozart, Tchaikovsky, and John Philip Sousa to Charles Ives, Anton von Webern, and Yannis Xenakis—seem as if it were written for ballet. His special sympathy and close friendship with Igor Stravinsky resulted in a long series of ballets and led him to devise new styles of movement to match the percussive, nonmelodic scores of the composer's later period. One of Balanchine's favourite devices was to choreograph a symphony or concerto using different dancers for each movement. He was generally more interested in choreography for women than for men, and there developed a "Balanchine-type" of ballerina—young, long-legged, fast-moving, technically brilliant, but usually rather lacking in personality. In general, Balanchine's works were more coolly classical, less sentimentally romantic, than Ashton's.

Heirs of Ashton and Balanchine

Just as Ashton's heirs are Cranko and MacMillan, so Balanchine's are Jerome Robbins and a number of younger American choreographers, especially John Clifford, Eliot Feld, and John Neumeier. Robbins first achieved prominence in 1944 with *Fancy Free*, a ballet derived from the style of Broadway musicals. He is perhaps best known for his direction and choreography of the musical *West Side Story*, but he also created abstract classical ballets, among them *Dances at a Gathering* and *Goldberg Variations*. Clifford, working with New York City Ballet, Feld with his own American Ballet Company, and Neumeier, in Frankfurt, West Germany, all created promising works in this style.

These are just some of the choreographers who contributed to the mid-20th-century ballet boom. In the West, it centred on Britain and the United States, where regional ballet began to flourish. Classical companies, often working in association with Balanchine, were especially successful in Washington, Philadelphia, Boston, and Salt Lake City, but there were also many others, and dance departments were established in many American univer-



La Fille mal gardée. Sir Frederick Ashton created a new version of the 18th-century Dauberval ballet in 1960, with the music of François-Joseph Hérold adapted by John Lanchbery, and decor by Osbert Lancaster. Performed by the Royal Ballet.
Houston Rogers

sities. In Britain, small regional companies were set up in Glasgow and Manchester. There were also touring companies in both countries. In the U.S., the City Center Joffrey Ballet had an eclectic repertoire of standard 20th-century works and new ones—including some multimedia experiments—by Robert Joffrey and Gerald Arpino, the company's directors. In Britain, London's Festival Ballet relied mainly on the classics.

Modern European ballet. The ballet boom extended to many other countries. Classical ballet companies were attached to virtually every German opera house, and, under Cranko, the Stuttgart company won international recognition. British influence was also dominant in establishing the Australian Ballet and the National Ballet of Canada. British ballet masters worked in Scandinavia, Turkey, Israel, Argentina, and many other countries. Balanchine lent the ballet in Geneva, Switzerland, ballet masters and dancers, and he also advised the company in West Berlin.

In Belgium, the French-born choreographer and producer Maurice Béjart staged large-scale dance spectacles in sports arenas and circuses as well as in the opera house. Béjart directed a large company of talented dancers with a particular appeal to young audiences. The ballet companies at La Scala, Milan, and the Vienna State Opera have inevitably taken second place to the opera, but even these companies gained status in recent years.

Copyright © 1961 Mirisch Pictures Inc. All Rights Reserved



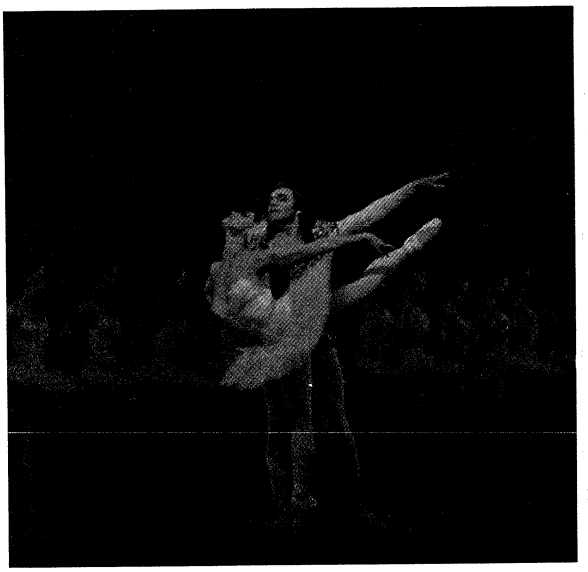
Members of the street gang the "Sharks" dance the choreography of Jerome Robbins to the music of Leonard Bernstein in *West Side Story*, film version, 1961.

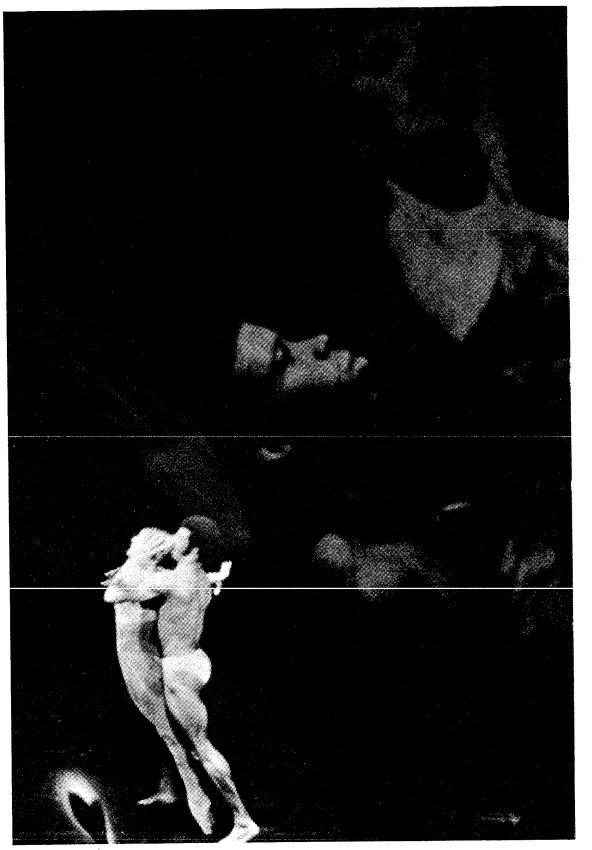
The Scandinavian companies

Great efforts were made to revive and strengthen the ballet companies in Norway, Finland, and Sweden; the Royal Swedish Ballet, one of the oldest companies in the world, benefitted from regular work with the British ballet mistress Mary Skeaping, the choreographer Antony Tudor, and the Danish *premier danseur* Erik Bruhn, who was director of the company from 1967 to 1971. The Royal Danish Ballet is even older than the Swedish and is the only Scandinavian company in the top international class. It preserves the charming and romantic 19th-century works created for it by Auguste Bournonville. Under its director, Flemming Flindt, the Royal Danish Ballet broadened its repertoire to include representative modern works in all styles, as well as preserving and sometimes restaging the Bournonville works.

Modern ballet in Soviet Russia. The Soviet government encouraged ballet companies in many provincial cities of the U.S.S.R. and helped to establish and maintain ballet in the east European Communist countries. Soviet companies made frequent visits to the West, and the Soviet style of dancing—more flamboyant and acrobatic than the traditional style inherited from Petipa—influenced Western dancers and choreographers. The former Leningrad dancer Rudolf Nureyev played a vital role in raising standards of male dancing and in staging various works by Petipa that had been forgotten outside Russia.

Immediately after the Revolution, ballet masters in the U.S.S.R. experimented with avant-garde styles and tried to give political content to their work, but under Stalin Soviet ballet reverted to a more traditional form. It concentrated on the 19th-century classics and on such neoclassical works as *Romeo and Juliet* and *Cinderella.* Soviet choreographers have attempted to modernize the classics by replacing conventional mime with naturalistic acting and by giving the principal male characters more dancing. Soviet productions of *Swan Lake* often end with the Prince killing Rothbart and "living happily ever after" with Odette, as the arts in the U.S.S.R. are intended to be uplifting and Russian audiences today are not supposed to believe in life after death. One recent full-length work, Yury Grigorovich's version of *Spartacus* (1968), succeeded in telling a revolutionary story entirely through dance and exploited the virility and athleticism of Russian male dancers in a new and exciting way.





Classical and experimental ballet in the 20th century.
(Left) The Bolshoi Ballet performs Aleksandr Gorsky's 1911 restaging of the 1895 version of *Swan Lake*, with choreography by Marius Petipa and Lev Ivanov and music by Peter Ilich Tchaikovsky. (Right) *Astarte*, Robert Joffrey's psychedelic *pas de deux*, was composed in 1967 as a multimedia work combining ballet with film, light, and rock music. Performed by the City Center Joffrey Ballet.
(Left) Novosti Press Agency, (right) Herbert Migdoll

Experiment of a more determinedly avant-garde kind became common in the West during the late 1960s and early 1970s. Especially in the United States, Britain, Belgium, and The Netherlands, producers and choreographers used electronic and concrete music, tried to achieve psychedelic effects with stroboscopic and projected lighting, and combined speech, song, and technical effects in an effort to break down the barriers between dance and other forms of theatre. In accordance with similar trends in the theatre and cinema, dancers were presented in the nude and realistic imitations of various sexual activities were devised. It is not yet clear whether these trends will lead to the emergence of a new form of ballet, or a new art form, or whether they will be largely forgotten like the similar experiments of the 1920s. It may be significant that the world's leading ballet companies and choreographers still preferred evolution to revolution and still put their faith in traditional styles and techniques and that audiences continued to flock to see their work. (O.Ke.)

BIBLIOGRAPHY. The most comprehensive general reference books on ballet are: ANATOLE CHUJOY and P.W. MANCHESTER (eds.), *The Dance Encyclopedia* (1967); and G.B.L. WILSON (ed.), *A Dictionary of Ballet*, rev. ed. (1961). Histories of ballet include books covering the entire history as well as those devoted to ballet in specific countries or historical periods. Some authoritative general histories are: ARNOLD L. HASKELL, *Ballet* (1938), *A Picture History of Ballet* (1954), and *Ballet Retrospect* (1964); LINCOLN KIRSTEIN, *Dance: A Short History of Classical Theatrical Dancing* (1935); and JOAN LAWSON, *A History of Ballet and Its Makers* (1964).

Books on ballet history that are devoted to single countries and periods include: CYRIL W. BEAUMONT, *A History of Ballet in Russia, 1616–1881* (1930); V.M. BOGDANOV-BEREZOVSKII, *Ulanovna and the Development of the Soviet Ballet* (1952; orig. pub. in Russian, 1949); EDWIN DENBY, *Looking at the Dance* (1949), essays on ballet in the United States; THEOPHILE GAUTIER, *The Romantic Ballet As Seen by Théophile Gautier* (Eng. trans. 1932); IVOR GUEST, *The Ballet of the Second Empire*, 2 vol. (1953–55), a study of ballet in 19th-century France; PETER LIEVEN, *The Birth of the Ballets Russes* (1936); JOHN MARTIN, *The World Book of Modern Ballet* (1952), mainly on the United States; DEIRDRE PRIDDIN, *The Art of the Dance in French Literature from Théophile Gautier to Paul Valéry* (1952); NATALIA ROSLAVLEVA, *Era of the Russian Ballet, 1770–1965* (1966); MARY GRACE SWIFT, *The Art of the Dance in the U.S.S.R.* (1968); and WALTER TERRY, *The Dance in America* (1956).

Interesting works that approach ballet history by focussing on its major figures, either as biography or reminiscence, include: ALEXANDRE BENOIS, *Reminiscences of the Russian Ballet* (1947); VICTOR DANDRE, *Anna Pavlova* (1932); ARNOLD L. HASKELL and W. NOUVEL, *Diaghileff: His Artistic and Private Life* (1935); TAMARA KARSAVINA, *Theatre Street*, rev. ed. (1950); DERYCK LYNHAM, *The Chevalier Noverre: Father of Modern Ballet* (1950); LILLIAN MOORE, *Artists of the Dance* (1938); ROMOLA NIJINSKY, *Nijinsky* (1933); and LYDIA SOKOLOVA, *Dancing for Diaghilev* (1960).

Valuable discussions of ballet technique are in: ERIK BRUHN and LILLIAN MOORE, *Bournonville and Ballet Technique* (1961); JEAN GEORGES NOVERRE, *Lettres sur la danse sur les ballets et les arts*, rev. ed. (1803; Eng. trans., *Letters on Dancing and Ballets*, 1930); and AGRIPPINA VAGANOVA, *Fundamentals of the Classic Dance*, trans. and ed. by ANATOLE CHUJOY (1946).

Stories of ballet libretti are contained in: GEORGE BALANCHINE, *New Complete Stories of the Great Ballets* (1968); CYRIL W. BEAUMONT, *Complete Book of Ballets* (1937), and *Ballets, Past and Present* (1955); and WALTER TERRY, *Ballet: A New Guide to the Liveliest Art* (1959).

(A.L.H.)

Ballistics

Ballistics is the subject concerned with the phenomena and laws of projectiles, their propulsion, flight, and impact. These three aspects of the field can be studied separately when applied to guns and are called individually interior (or internal), exterior (or external), and terminal ballistics, respectively, and a wide range of phenomena is subsumed under them. Rocket propulsion in free flight involves the special problem of interior and exterior ballistics at the same time because the thrust of the projectile and its aerodynamics are interdependent during flight. Ballistics beyond the atmosphere (that is, space flight) is a subcategory of exterior ballistics known as geoballistics. The ballistics of cartridge-actuated devices concerns the use of propellants to move heavy loads short distances, or to work against large resistance in short periods of time. Still another branch of study is hydroballistics, the study of projectiles moving through the

water. The general effects treated in ballistics are reproducible muzzle velocity; the lowest gun pressure that will produce the desired velocity within a given length of the gun barrel; the stability of the projectile in flight; long and reproducible trajectories; and maximum interaction with a selected target. In this article "ballistic" is used to apply to all the properties, behaviour, reactions, etc., of projectiles of any sort—under any and all conditions.

HISTORY OF BALLISTICS THEORY

Ballistics theory depends on many scientific disciplines. Interior ballistics involves the application of theories of chemical combustion, compressible gas dynamics, and mechanics; exterior ballistics theory is built on a knowledge of mechanics, aerodynamics, and, sometimes, geodesy—the shape and physical character of the Earth; terminal ballistics draws on the theories of explosive technology, metallurgy, solid-state physics, and mechanics. Because all ballistic phenomena occur at extremely high rates, and a great complexity of interactions between them develops, the fastest and most capacious computers, microsecond photography, supersonic wind tunnels, and pressure gauges capable of measuring several thousand megabars (a megabar is equivalent to about one million atmospheres of pressure) in microseconds are required in ballistic study.

Ballistics problems have always challenged the most powerful and sophisticated scientific techniques available and have stimulated new ones. Among the prominent early contributors to ballistics were Leonardo da Vinci, Galileo, and Newton. Men like Leonhard Euler, a Swiss astronomer and mathematician; Pierre-Simon Laplace, a French physicist and astronomer who became Examiner of the Royal Artillery; and Antoine-Laurent Lavoisier, the French father of chemistry and director of the French National Powder Factory, also made vital contributions. The first hurled object was a problem in ballistics. The advantages of being able to hurl farther or harder led to devices for extending the arm, slings and throwing sticks. Further increases in effectiveness were achieved with the bow, which combined increased mechanical advantage and the accuracy derived from a more stable platform than provided by the human in the act of throwing. Arrows could reach the parapets of a defensive structure, but breeching the walls or causing internal structural damage necessitated the invention of the ballista from which ballistics derives its name. Ballistics shares an etymological root with the word diabolic, or devilish, both deriving from the Greek word *ballein*, "to throw." *Diaballein* means to throw across or, alternatively, to calumniate. The intimate relationship between the human body and ballistics in propulsion by throwing or using the ordinary bow makes it doubtful that hunters and warriors felt a need for an intellectual ballistics principle. With the advent of the ballista, however, the intimate sense of feel and the integration of arms and eyes are absent. The ballista propelled missiles in one of three ways: through torque, equipoise dropping of weights, or flexure. Its capacity was determined by the tension, lever, and weight values applied to the device. By the time of Leonardo da Vinci, the principles of levers were well known, so that methods of increasing initial velocity of various weights could be understood characteristically, if not analytically. Because Leonardo and his contemporaries had mistaken views of the magnitude of gravitational effects, the effective use of the ballista required experiments to test the range of each missile and each variation in the ballista's design and setting; a completely empirical set of data, therefore, had to be obtained for accurate use of the machine. Knowledge of terminal ballistics was also empirical, although the chemical formulation of Greek fire was a refined and highly secretive art.

The work of Leonardo da Vinci

Prior to the advent of guns, then, principles of ballistics included the application of the principles of levers to ballista and of geometry to aiming.

It might also be noted that although Leonardo had an excellent knowledge of static mechanics, his version of dynamics was empirical and theoretically incorrect. He incorrectly deduced that force is directly proportional to velocity, rather than a change of velocity with time; that is, he did not seem to comprehend that force is directly proportional to acceleration. Thus, he could have no correct theory of interior or exterior ballistics. He did write on the trajectory of cannon shot, and it is not surprising that a sculptor would have an interest in the cannon foundry. Leonardo, as a late 15th-century Florentine, lived only a century after the discovery or invention of gunpowder and guns, and knew only their early use in warfare—two centuries before the English began to regulate the propulsive properties of gunpowder.

Until the first half of the 16th century it was assumed that a projectile travelled in a straight line to the zenith of its trajectory, and then fell on the other leg of an isosceles triangle. It was observed that flight was curved, but the principle applied was simply to approximate the curvature of flight with a circular arc in the upper portion of flight. No concept of drag is included in this description, and it was 1740 before Benjamin Robins, an English mathematician and engineer, first measured with any accuracy the loss of projectile velocity in flight. His results were incredible to scientists of that time. It had been thought that the influence of drag was about the same as that of gravity; but experiments have shown it to be many times greater. The measuring device that Robins invented was a ballistic pendulum that operated on the principle that the swing of a pendulum of large mass, when struck by a projectile with less mass, can be used to calculate the impact velocity.

Gunpowder. Muscle and gravity were eventually replaced by gunpowder, but there is dispute over its origin. It could have been imported from China, though Marco Polo never mentioned such a curiosity in his writings; or it could have been invented in Italy, because chemicals of the proper quality were mined there; or it could have been discovered by Sir Francis Bacon. Whatever the source, gunpowder provided a human experience with dynamical forces far, far beyond the effects of muscle, levers, or weights. There were no theoretical tools in chemistry, physics, or mathematics with which to begin a theory of gun ballistics, but the history of science since that time is replete with mathematical, physical, and chemical discoveries that were made in a search for solutions to problems in ballistics.

PRINCIPLES OF BALLISTICS

Interior ballistics. *General considerations.* It is useful to begin with the simplest ballistic case—a bullet fired from a gun. The bullet is driven along the barrel by the gas pressure generated when the gunpowder propellant burns at extreme rate in a chamber (called the breech) at the rear of the gun. The process is analogous to that of a piston in an internal-combustion engine, the projectile being the piston. The gross variables of this process are the weight and cross-sectional piston area of the projectile, the gas pressure applied to its base, and the length of travel in the gun along which the propellant gases can exert force. If a mathematical relationship can be found for these factors, then the ballistic behaviour of any bullet fired from any gun using any propellant can be approximated.

In the broadest sense, the classical laws of motion are applicable to any moving object: the force (F), causing the mass (m) to accelerate (a), is equal to the product of the mass and the acceleration; that is, $F = ma$. Since the projectile's mass is usually constant throughout the period defined by interior ballistic period, its acceleration will result from the applied force. That force is the sum of: (1) the gas pressure (p) working on an area (A) (effectively, the cross-sectional area of the gun bore that is normally taken as a constant), and (2) the resistance to motion resulting from starting forces, friction, and spin (in the case of rifled gun barrels). The latter forces are small relative to that derived from the pressure, but are required in useful formulations of the interior ballistic trajectory—especially the shot-start pressure (pressure at instant of firing) required to engrave the projectile's soft metal band, usually of copper, that then rotates the pro-

jectile along the grooves of the barrel. By using classical physics, the foregoing ballistic data is related in an equation such that the acceleration (a) equals the pressure (p) times the area (A) plus the resistant force (f), divided by the mass (m), the equation being written: $a = \frac{pA + f}{m}$.

The principal variable in this process is the pressure on the projectile base; if it can be measured, then the acceleration can be calculated. The pressure that is effective on the base of the projectile depends on the propellant's properties, on the amount of it that is burned, and on the spatial distribution of gas down the bore. The outcome of these variables is a pressure that is usually measured by instrumentation at a fixed location—the breech end of the gun.

Chemistry of propellants. Propellants are substances that ignite and burn extremely fast, producing voluminous quantities of gas in a very short time. In general, the burning process is a chemical reaction involving oxidation, an oxidizing agent being any substance that removes electrons. The shorter the period during which a given mass of propellant burns, the more explosive will be the production and expansion of the gas. Thus, propellants are mixtures of chemical compounds that contain both fuel and oxidizer, in contrast with internal-combustion engines into which a mixture of fuel vapour and atmospheric oxygen are metered. Propellants must be chemically and physically stable in ordinary environments, and must react only to specific conditions that can be reproduced acceptably. They are ignited by a primer propellant that has been in turn ignited by heat either from an electrical-resistance wire or from an explosive percussion primer. Properly manufactured, prepared, placed, and ignited, the propellant material burns at a rate that depends on the pressure of the product gases and, secondarily, on the temperature of the unburnt material. The higher the temperature, the faster will be the reaction. This burning rate, usually expressed in terms of length burned per unit time, is determined by burning a strand of the material at various controlled pressures. When the strand is inhibited from burning along its sides and made to burn evenly across its face, the linear rate of burning is measured and an equation is derived in which an exponent of the pressure varies from 0.1 to 1, with conventional propellants; such a range of values makes determination of this variable a critical factor in calculating the ballistic behaviour of any particular gun or projectile.

Geometry of propellant grains

The propellant charge in a gun is composed of many small pieces, or grains, that are ideally of the same shape, and, thus, the weight of propellant consumed is a function of the accumulated linear consumption at any time in the combustion process. The actual shape of the grains is an important factor and, therefore, the problem is a geometric one. Solid cylindrical grains present a smaller and smaller area as they burn and represent a regressive geometry. Cylinders with central perforations along their length present a nearly constant surface, representing a neutral geometry. With a number of longitudinal perforations (usually seven) that expand, the amount of propellant burned increases as the cylinder burns away up to the point at which the expanding perforations meet. This is called a progressive geometry. The ballistic advantage of the latter is that it generates gas at a lower rate, initially when the total volume behind the projectile is small, and produces gas at an accelerated rate as the projectile gathers velocity and creates volume behind it at an increasing rate.

Behaviour of gases. The chemistry of propellants involves the so-called equation of state of the material; *i.e.*, the relationship between pressure, volume, and temperature of a gas. If p is pressure, V the volume available to the gas, n the number of moles of the gas (weight in grams equal to the molecular weight), R the universal gas constant, and T its temperature (in degrees absolute), then the product of pressure and volume will equal the product of the number of moles, the gas constant, and the temperature, thus: $PV = nRT$.

Since the gas itself must be accelerated down the gun bore behind the projectile, it is distributed at any time in such a way that local pressure decreases from the breech to the projectile's base as the latter travels. The average pressure along the gun behind the projectile can be used to calculate the instantaneous rate of burning. The energy for the work done in moving the propellant gas itself is obtained from the total pressure generated, and the projectile, therefore, receives that much less. Accordingly, acceleration can be estimated by dividing the force of the combustion (the mean pressure of the combustion times the effective area of the bullet) by the mass of the bullet plus one-third the weight of the propellant. The thermodynamic, or heat, efficiency of the system is also reduced by any gas leakage past the projectile, by potential work available in the gas remaining when the projectile leaves the muzzle, and by heat lost to the gun itself. When all these factors are considered, what is left for the kinetic energy in the projectile is about one-third of the potential chemical energy contained in the propellant charge.

There is clearly an exceedingly intricate relationship between the rate of propellant consumption, the pressure at which the propellant burns, and the volume of gas behind the projectile; the relationship is expressed in several simultaneous differential equations which are frequently programmed on an analogue computer to provide curves of gas pressure and projectile travel versus time. These vary for each assembly of devices, but when for any one system enough measurements have been made, the formula can be used to predict desired factors (for example, the amount of a propellant needed to move a certain projectile at a desired acceleration).

Projectiles are often given a spin in order to obtain a more stable and uniform flight pattern. This rotational acceleration of projectiles, as well as other secondary variables, such as frictional resistance to motion and leakage of gas, must be taken into account in a full development of interior ballistics. Although they use, collectively, only about 5 percent of the energy consumed in the whole process, they can affect projectile performance significantly and are, thus, important to ballistic calculations. For example, the inferred empirical value for shot-start pressure is vital in establishing the conditions for burning during projectile travel. Some ballistic devices have been designed with leakage as a critical parameter. If a salvo of several projectiles is to be fired at once with a particular pattern of impact ranges, this can be accomplished either by pre-aiming several barrels with appropriate firing angles relative to one another to achieve the pattern or by making the barrels with various effective lengths; the same effect is achieved with barrels of identical length that leak gas after the projectile has travelled different distances in the bore.

Absence of a satisfactory ignition theory

No theory has been satisfactorily developed for ignition of propellants. Ignition has remained a facet of technology in which the aims are to ignite reliably and quickly at all propellant temperatures and without high rates of pressure rise.

As noted above, the propellant gas is accelerated down the gun bore behind the projectile. Since the weight of the propellant charge is usually of the same order of magnitude as the projectile weight, a considerable amount of the propellant energy is expended on its own acceleration. The predominant products of combustion are usually oxides of nitrogen and carbon, with molecular weights on the order of 30 or higher. If the propelling gas is much lighter, as, for example, hydrogen is, then much higher velocities of gas and projectile are feasible. For research experiments two-stage guns using a light gas have been built, and projectile velocities of five to ten kilometres per second (three to six miles per second) have been achieved at their muzzles.

Gun wear. An undesired product of the interior ballistic cycle is gun wear. Chemical corrosion from the gases and erosion from heat and friction steadily enlarge the bore diameter of a gun mostly near its ends. The effects of this wear include different initial conditions for projectile motion, leakage, and even projectile instability

in the bore. The gun tube, as a pressure vessel, must be designed at each position along its length to withstand the maximum pressure it will receive, increased by a suitable safety factor. To the extent that the pressure pulse the gun receives is below the design level the gun is inefficient. A pressure gauge at any point along a gun measures a pressure that rises quickly to a maximum and then decays slowly, the gun at that point being able to withstand the design maximum pressure for the entire interior ballistic cycle. The ratio of the mean pressure on the projectile to the maximum pressure is called piezometric efficiency. Its value taken at breech pressure is normally between 0.5 and 0.75.

The interior ballistic portion of the total ballistic trajectory ends with the velocity, spin, and direction imparted to the projectile at the gun muzzle as well as to the projectile's angle of attack to its direction of motion.

Exterior ballistics. Exterior ballistics is concerned with the flight of projectiles through the atmosphere, and of primary interest are: the manner in which projectiles are slowed as a result of a factor called drag; how they maintain a consistently favourable attitude (a variable factor called stability), and how they are influenced by variations in interior ballistics and by their flight environment.

General considerations. First, the interaction between the projectile and the medium through which it passes must be examined and measurements made in terms of coefficients that indicate the amount of changes in the projectile's behaviour as the conditions change. The aerodynamics of a nonspinning projectile are represented in terms of three dimensionless parameters which appear as coefficients of drag, of lift, and of moment (a moment measures the forces tending to turn an object) in equations describing how a projectile is slowed by the air, how it is moved laterally (lifted or depressed) by the aerodynamic pressure distribution over the projectile in flight, and how the projectile tends to tumble because of an overturning aerodynamic moment. These coefficients are all functions of the speed of the projectile and properties of the air such as density and elasticity. They are expressible as functions of projectile velocity divided by the local speed of sound (that is, Mach number; Mach 1 being a speed equal to the velocity of sound in the same medium) and of the so-called Reynolds number (based on the velocity at which streamlined flow of a fluid becomes turbulent) of the air or fluid through which the projectile is passing. The theory of the Reynolds number is complex, and a good approximation to conventional ballistics can be obtained by ignoring it and regarding the three coefficients as functions of Mach number only. (This is not the case for high speed—hypersonic—or hydroballistics. For those the change in Reynolds number is sufficient to require its use in conversion of experimental data.) These coefficients are also different for different projectile shapes. Specifically, for typical shapes the drag coefficient is near 0.08 at low velocities, increases very sharply to about 0.2 at the speed of sound, and decreases slightly at higher velocity. The coefficients are normally determined experimentally either from direct force measurements in a wind tunnel on a fixed projectile shape with air blown over it, or by inference from fitting flight trajectory data to a mathematical model. If ρ is the air density, d the projectile diameter, v the projectile velocity, and δ the angle of yaw—or the angle between the projectile's long axis and its direction of motion—and K_D, K_L, K_M are coefficients of drag, lift, and moment, and these forces are combined into the following linearized equations, then drag equals density times diameter squared times velocity squared times the drag coefficient; lift equals the same factors but substituting the coefficient of lift for the coefficient of drag, times the sine of the angle of yaw; and the moment equals the product of the density, diameter cubed, velocity squared, the coefficient of moment, and the sine and cosine of the angle of yaw. Written out symbolically, the equations are

Mach and Reynolds numbers

$$\text{Drag } D = \rho d^2 v^2 K_D$$
$$\text{Lift } L = \rho d^2 v^2 K_L \times \sin \delta$$
$$\text{Moment } M = \rho d^3 v^2 K_M \times \sin \delta \times \cos \delta.$$

Drag and spin. The drag effect at zero yaw can be broken up into three influences: that due to pressure of the atmosphere on the bow of the projectile resulting in a backward force component; that due to skin friction of the atmosphere along the body of the projectile and varying with its surface roughness and minor convolutions; and that due to base drag resulting from the turbulence induced in the atmosphere at the rear of the projectile. At usual projectile velocities, say 300 to 900 metres per second (1,000 to 3,000 feet per second), pressure drag is reduced by sharpening and shallowing the curve of projectile noses, frictional drag by body smoothness, and base drag by drawing the base more toward a conical shape.

Factors in stable flight

In stable projectile flight the orientation of lowest drag is presented to the airstream. Unstable flight leads to high drag, short range, and reduced accuracy. Static stability is achieved when the centre of pressure (resulting from combination of forces including drag) acting on the body is behind its centre of gravity. This is sometimes achieved by the use of fins as on bombs, mortar shells, and nonspinning rockets; the fins acquire the centre of pressure, whereas the centre of gravity remains somewhere around the midsection of the projectile. Dynamic stability is achievable even when the centre of pressure is forward of the centre of gravity by spinning the projectile and thereby inducing a gyroscopic action that counterbalances the wobbling or tumbling that the pressure would induce. Thus, unless suitable fins are provided, spinning the projectile is necessary for the stability of its longitudinal axis in flight. Too much spin will make a projectile so stable at high angles of fire that it will resist following its velocity vector near the zenith of its trajectory and assume an increasingly sideways orientation, and eventually it will tumble. The mathematical treatment of spinning projectiles introduces complex variable notation to the force system described above; a number of separately indentifiable torques (or moments) and forces result that are expressible as functions of the same basic parameters that are identified without spin, with the addition of the spin parameter and new dimensionless coefficients. These coefficients are also empirical. The cross forces and moments, which result from spinning in the airstream, are called Magnus forces or Magnus moments. To make the mathematical problems tractable, projectile yaw is assumed to be small, and a linearized theory is used. There has been some analytical development of nonlinear yawing motion.

Experiments with projectiles in flight are of two kinds. In one, flight is simulated in wind and water tunnels by making air or fluid flow past a stationary projectile that is hooked up to instruments. Alternatively, projectiles can be fired down dark tunnels called spark ranges where flash photographic images are taken. The measurements obtained in either case are complex, but the equations developed from such data can then be used for approximate solutions of problems in exterior ballistics.

Geoballistic factors. Over long ranges of flight, including an "infinite" range postulated by Newton (*i.e.*, that of orbital satellites), effects that can be neglected in conventional ballistic systems become prominent and, at extreme altitudes, the significance of aerodynamic effects are reduced. For very long ranges and times of flight, the spinning of the Earth imposes a small force, called the coriolis force, which causes projectiles in the Northern Hemisphere to have a rightward curving flight, and in the Southern Hemisphere a leftward curving flight. The explanation of this is that since the surface of the Earth, spinning on its axis, moves at different speeds at different Earth latitudes (a point on the Equator moves faster than any point north or south because the circumference at the Equator is greatest; while a point at either pole is stationary), when trajectories cover a range of latitudes this must be accounted for. Moreover the oblate nonspherical configuration of the Earth affects orbital trajectories as does, to lesser degree, local variation in the magnitude and direction of gravity forces. The forces resulting from the attractions of Sun and Moon are negligible. All these effects are treated in a modern branch of ballistics called geoballistics.

Aiming bombs

Nonspinning projectiles. Bombs are fin stabilized and are dealt with as nonspinning projectiles. The problem of aiming a bomb introduces the problem of locating a release point from a moving platform in order to make the bomb's trajectory intersect a target. Since a bomb-carrying aircraft counteracts air drag with its engine thrust while the falling bomb does not, after its release, the bomb immediately trails the aircraft while it falls. To select the point at which to release the bomb requires mathematical manipulations of data, and the aim is to find the most efficient way of mechanizing the selection of release conditions. While being carried and during its release, the bomb is subjected to complicated airflow conditions around the aircraft. The effects of this are not well understood.

Meteorological effects. Corrections for variation from standard, of either gun muzzle velocity or of projectile weight, are supplied by exterior ballistics. A more difficult correction problem is posed by the meteorological conditions. Wind speed, direction, and air density vary with altitude and horizontal distance from the gun. Measurements are not generally available for the specific flight path intended, but, rather, one must use data gathered at a more or less nearby location and at an earlier time. Internationally agreed-upon standard atmospheres are used as a first approximation in standard ballistics tables giving range and time of flight for various gun elevations and selections of ammunition. Meteorological differentials are then applied, either for a single "average" value of wind and density variation from standard, or even for variations with altitude.

Ideal shape for projectiles. The ideal projectile for interior ballistic performance differs from the exterior ballistic ideal. The latter has a long finely pointed nose and tail for low drag and is heavy for a high ballistic coefficient. Best interior ballistic performance is attained with projectiles that are lighter, have higher muzzle velocities, and have a nearly cylindrical shape which provides better bearing surfaces through the gun bore for low wear. The terminal ballistic ideal may be different from both of these, since usually what is wanted is a large payload weight but low accelerations in the gun. The practical design is a compromise.

Rocket ballistics. *General considerations.* The general principles underlying rocketry involve propellant combustion in a chamber that has some suitable aerodynamic shape and that expels the gaseous products of the combustion through one or more nozzles at the rear of the chamber. If the rocket is well designed, the internal combustion pressure remains nearly constant during burning. Otherwise, the pressure vessel must be designed for the maximum pressure generated and at lesser pressures the extra protection is wasted; the design is then inefficient. The rocket nozzle is designed to effect a smooth flow of gas from the combustion chamber, accelerating the escaping gas velocity to the local speed of sound at its most constricted point (throat), and expanding it to nearly atmospheric pressure and supersonic velocity. According to Newton's second law of motion, the speed of change of the momentum of the gas leaving the rocket represents a force, and, according to his third law, this force is reacted against by a thrust in the opposite direction. It is this thrust that drives the rocket forward. Clearly, the thrust is increased with higher velocity and increasing propellant gas density. Thrust is the change in the product of mass (m) and velocity (V) of the gas divided by the change in time. This is written as $\frac{dm\,V}{dt}$, the d indicating an infinitesimal change.

For well designed rockets the exhaust velocity is substantially constant. Since the rocket mass includes the steadily reducing mass of the propellant, which is used up during burning, and since the rocket is accelerating, both mass and velocity are changing. The thrust of a rocket is approximately proportional to the product of the smallest cross section area in the nozzle and the chamber pressure, with a factor of proportionality between one and two depending on the expansion ratio.

Propellants. The propellant may be solid or liquid. When liquids are used the fuel is usually stored separately from the oxidizer (that is, the substance that "burns" it), and the two are mixed in proper proportion at rates designed to cause them to ignite spontaneously and burn evenly. Examples of fuels are liquid hydrogen and ammonia. Liquid oxygen, nitric acid, and fluorine have been used as oxidizers. Solid rockets require that the grain of the propellant have a geometry designed to present a nearly constant surface area as it burns away, and that the grains have a suitable rate of burning. Rocket thrust is a function of time, varying from the instant of ignition through flight, and it can be measured on a gauge that indicates force. The ideal curve derived in this fashion rises quickly to a pressure plateau and, at the end of burning, drops quickly. Another factor affecting flight is the fact that rocket propellants burn at a rate that depends on the propellant temperature. Fortunately, the conduction of heat through a mass of rocket propellant is low in relation to the actual linear rate of burning so that the temperature tends to remain constant during burning. That is, the propellant burns before the heat of combustion can spread ahead of the flame. Higher initial temperature of the propellant, however, increases the rate of burning and, thus, the thrust is raised to a value higher than the designed value. Finally, although curves for rocket thrust versus time may vary from the roughly trapezoidal form referred to above, it is the integral which corresponds to final rocket velocity at burn out, and the latter tends to be less variable than individual values of thrust during burning.

When a rocket is ignited inside a gun barrel, with or without an independent cartridge charge, another variable must be taken into account. Gas pressure difference across the exhaust nozzle of the rocket is considerably lower within the gun barrel than in free air. This application of rocketry is efficient only if short, high-velocity trajectories are desired.

Special problems. Problems peculiar to rocketry arise during flight. Thrust malalignment may occur either because the thrust is at an angle relative to the longitudinal axis of the rocket or is along a line displaced from, but parallel to, the rocket axis. The latter is called linear malalignment. Angular malalignment gives the thrust a sideways component that must be reacted against by the rocket body and fins. Linear malalignment produces a torque moment (a tendency to turn due to uneven forces acting on different parts of the body) relative to the centre of mass. The peculiarities of rocket flight also include the fact that a yawing rocket has a thrust component along the direction of its attitude as compared with its direction of motion. Subjected to a crosswind, finned rockets tend to weathercock or turn into the wind and to fly upwind. After burnout the weathercocking is still present, but drag due to the atmosphere results in a downwind motion. The reaction to a changing angle of rocket attack as reflected in the angular change in gas momentum is a yaw damping in the rocket called jet damping torque.

Weathercock effect

Rockets may be given spin stability by canting two or more nozzles relative to the centreline of the rocket. This angular and linear malalignment, if properly designed and constructed, results in a rolling moment, or spin, being applied to the rocket without also adding pitch and yaw.

Rocketry is a most efficient technique when low reactions by the launcher are required, as is the case with a light platform, and also when the weight of the gun pressure vessel exceeds that which is ordinarily allowable. Although these advantages are not always realizable in practice, another advantage of rockets has been important in achieving space flight: beyond the Earth's atmosphere drag is negligible and the propulsive energy of the rocket can be used efficiently. Rockets also provide the capability of delivering heavy loads to targets thousands of miles away and, with onboard guidance systems, achieving a high degree of accuracy in that application.

A combination of effects is achieved in recoilless guns.

By venting propellant gas through a suitable nozzle at the rear, or breech, of the gun, the gun is effectively a rocket receiving a forward thrust. This tends to counteract the projectile's momentum (that is, to work against the recoil). It is an obviously inefficient use of propellant in order to reduce the weight of the gun platform.

A disadvantage of rockets as compared with guns is that they are not normally re-usable, while the gun is. Nozzle erosion in most rocket designs significantly affects efficiency so that reloading, even if the rocket were recovered, would require replacement of the nozzle as well as the propellant and usually would also require repairs to the motor case.

Terminal ballistics. *General considerations.* Of the various events that occur when a projectile hits a target, the simplest to discuss is penetration of the target by the projectile. Attempts to state analytically the three-dimensional process of penetration accounting for the material properties of target and missile have been only partially successful. Even the simplest two-dimensional attempts have encountered considerable computational difficulty. The phenomenology of penetration has evolved out of experimental observations and known characteristics of material failure under shock loading. It is most easily approached through projectile impact on a single plate of some thickness. Penetration completely through the plate is called perforation. Partial penetration that pushes whatever is left of the surface at the rear face clear of the plate around the penetration path is called scabbing. If the projectile's impact causes a pressure shock of sufficient amplitude in the plate, and if the plate thickness and its material permit it, a piece of the plate, thinner than the plate thickness, may be separated and driven by shock from the rear face of the plate without the projectile penetrating an appreciable distance. This piece, called a spall, is split off the plate by the extreme tensile force within the plate, caused by the rarefaction pulse which results when the shock front is reflected from the rear surface of the plate that is not penetrated. A theoretical description of spallation can be given from the impact geometry and velocity and from the equation of state of the materials. The depth of penetration for a projectile that survives impact without fracture is roughly proportional to the impact velocity and the projectile length, and it increases with the density of the penetrator, which may not be identical with the density of the projectile before impact. As the angle of impact on the plate varies from the right angle to the plate, the possibility of ricochet increases. The impact angle at which half of the projectiles in a series being tested ricochet is called the ricochet angle.

Spallation

Explosive projectiles. Projectiles may contain high explosives which cause the shell casing of the projectile to fragment before or after impact with a target. Fragmentation increases the region affected, called the lethal volume, if the fragments resulting from impact cause the kind of damage sought. The process of fragmentation is understood primarily from empirical data but the reproducibility of material failure under shock loading suggests that fragmentation is a more uniform process than any that could occur under very slow loading rates; *i.e.*, slow increase in pressure. The projected fragments are characterized by their masses, velocities, and directions. The terminal ballistic process for fragmenting projectiles is then completed by impact of fragments with the target.

Effects of shaping charges

Different kinds of action can be achieved with high-explosive projectiles by preforming and shaping the interfaces between explosive and metal, as is done, for example, in the case of the so-called rod warhead. If the ends of metal rods are properly attached in a cylindrical bundle around the explosive, the rods can be caused to expand into a hoop shape, with a resulting cutting action. Another kind of action is achieved with a cylinder filled with explosive and terminated at one end in a metal cone the apex of which is in the explosive. This is called a conical-shaped charge. Detonation at the apex end of the charge drives the cone material forward and toward its axis, and the collision can be explained in terms of fluid mechanics. The resulting jet of metal achieves velocities above 3,000 metres per second (10,000 feet per second). The fluid properties of this system result in jet elongation that, up to the point of separation and dispersal, increase penetration. The laws of penetration regarding the velocity and length of the penetrator indicate greatly increased penetration for such devices and experiments verify it. Similar techniques are used for fragment accelerators to study the effects of meteoroid impacts in space.

Theory of terminal ballistic impact. A theory of terminal ballistic impact is now evolving as a result of simulation using digital computers, based on a number of hydrodynamic representations (considering the materials under high pressure as fluids). By the use of computer graphics, a picture of the fluid can be presented as it is being compressed or rarefied and as its geometry changes. This is done by calculating and plotting a lattice of connecting points of the fluid as a function of time. Initially, the lattice may look like a section of graph paper, but then with time the lattice is seen to distort and change configurations just as would the fluid when disturbed. Indeed, it is a motion picture of its history. Inputs in such computations include impact velocity and physical variables of projectile and target, such as density, speed of sound in the materials, material strengths, and equation of state, all relating pressure and density in the material. Output variables include dimensional deformation and internal pressure distributions. Varying impact conditions in the formulations beyond the range of available experimental evidence can suggest new directions, but the validity of computed extrapolation beyond experience is questionable until verified by experiment. The great speed and analytical ability of the digital computer nevertheless allows some success in attempts to smooth projectiles for flight, to increase muzzle velocity, and to penetrate the target more deeply. Thus, in all areas of ballistics, a process of computer-simulated experimentation supported by selected physical experiments has become useful.

Theory of terminal ballistics. A theoretical structure for terminal ballistics is a relatively current development, having begun about 1960. Although all fields of ballistics are centuries old, in an experimental sense, terminal ballistics is in its theoretical youth and has considerable promise in explaining physical relationships in materials at very high pressures.

RECENT DEVELOPMENTS

Cartridge-actuated devices

Since World War II, propellants and explosives have found increasing use in applications other than conventional guns. Devices that take advantage of the large amounts of potential chemical energy stored in small-volume propellant cartridges to do work in an emergency or in a place remote from large power supplies are called cartridge-actuated devices. They are widely used in aircraft personnel escape systems and in aircraft bomb and missile ejectors, as well as for attaching and detaching structure, inflation, emergency positioning, cutting, and electrical switching.

The ballistic properties of cartridge-actuated devices differ from those of guns in both degree and kind. Many cartridge-actuated devices move weight or move against a resisting force in a substantially closed system so that only interior ballistic theory is applicable. The action of explosive separators is best described by terminal ballistic principles as in shock-separated explosive bolts and linear-shaped charge for cutting aircraft escape capsules away from the rest of the aircraft.

Differences in degree and kind can be seen in comparing pilot ejection catapults with guns. In the late 1940s, when aircraft performances exceeded the pilot's ability to bail out safely, the pilot ejection seat was developed. The system employed a catapult gun powered by gun propellant. Early catapult guns shot man and seat away from the aircraft at a velocity of about 18 metres per second (60 feet per second), achieved in about one metre of travel. The prime ballistic constraints in such a system are the ejection velocity necessary for the ejection seat to clear the aircraft's vertical stabilizer and the maxi-

mum acceleration below the level that the human spine can sustain (about 25 times the force of gravity, or 25 g). Although the system can be regarded as a simple gun, its exterior ballistic properties are special with regard to the high initial yaw of the "projectile," a large overturning moment at the muzzle, and its unusual ballistic shape. The terminal ballistic problem is to deploy a parachute at the right time and at the right attitude of the "projectile." The interior ballistic factors are also unusual for a gun: operating at perhaps 3 percent of the usual maximum pressure over about 20 percent of the usual barrel length against 10 times more projectile weight than is conventional to achieve muzzle velocities one-fiftieth of the usual.

In more extreme interior ballistic situations, aircraft ejectors employ propellant charges to thrust against as much as 50,000 times their weight over distances on the order of 15 centimetres (six inches).

Many cartridge-actuated devices burn gun propellants at pressures much lower than desirable for uniform combustion. Also, as described above, pressure is dependent upon the projectile mass, which can vary by a factor of 10 in the same cartridge-actuated device. This can lead to inefficient and non-reproducible ballistics. However, if the propellant is burned in a chamber separated from the working piston and cylinder by a wall perforated with holes of suitable size, the propellant can be burned in a reproducible way, the flow of gases through the perforations will be sonic, and the pressure exerted in the working cylinder will be roughly independent of initial temperature and projectile weight.

As a family, cartridge-actuated devices experience heat loss during the interior ballistic cycle of 15 to 20 percent of the potential propellant heat energy, as compared with 5 percent in guns. This factor must be accounted for in the interior ballistic theory of cartridge-actuated devices.

New uses for ballistic theory and devices

Ballistic theory is applicable to guns, mortars, rockets, and any thrown, hurled, or freely propelled missile, as well as to closed systems that use stored or generated gas as a source of energy. Consequently, the principles are widely applied in space exploration as well as in many of man's Earth-bound activities. Examples include the inflation of flexible containers by propellant gas, which activates life rafts and crash-protection bags in motor vehicles, and explosive devices for welding, the forming of complex metal shapes, and emergency cutting. The principles that have evolved from studies of terminal ballistics have civilian as well as military application; police control measures, for example, include the use of rubber bullets to reduce the incidence of human injury.

BIBLIOGRAPHY

History: W.Y. CARMAN, *A History of Firearms, from Earliest Times to 1914* (1956), a comprehensive treatise on the origins and evolution of guns and cannons with inferential information on ballistics.

Interior ballistics: J. CORNER, *Theory of the Interior Ballistics of Guns* (1950), a standard work; F.R.W. HUNT *et al.* (eds.), *Internal Ballistics* (HMSO, 1951), a practical work accessible to the inexperienced ballistician with practical formulations capable of hand calculation.

Exterior ballistics: E.J. MCSHANE, J.L. KELLEY, and F.V. RENO, *Exterior Ballistics* (1953), contains a historical appendix with a thorough account of the history of exterior ballistics with reference to other branches of ballistics.

Terminal ballistics: RAY KINSLOW (ed.), *High-Velocity Impact Phenomena* (1970), the first book on terminal ballistics, with emphasis on experimental techniques and computer solutions.

Rocket ballistics: G.P. SUTTON, *Rocket Propulsion Elements*, 3rd ed. (1963), a standard work, with bibliography; L. DAVIS, JR., J.W. FOLLIN, JR., and L. BLITZER, *Exterior Ballistics of Rockets* (1958), a practical and theoretical treatment of rocket ballistics without interior ballistic coverage; a companion to Sutton's book.

(R.I.Ro.)

Baltic Languages

The Baltic languages form a branch of the Indo-European language family and are more closely related to Slavic, Germanic, and Indo-Iranian (in that order) than to the other branches of the family. They comprise modern Lithuanian and Latvian (Lettish), the languages of the Balts inhabiting the eastern coast of the Baltic Sea, as well as the now extinct Old Prussian language, Yotvingian (also spelled Yatvingian, Jotvingian, Jatvingian), Curonian (Kurish), Semigallian, and Selonian (Selian); the speakers of this group are here referred to as the B-Balts. There also existed languages and dialects of the Balts (D-Balts) who lived east of the above-mentioned groups in the areas of the upper reaches of the Dnepr River.

Languages of the group. Because its dialects are more archaic in their forms than those of the other living Indo-European languages, Lithuanian is of particular importance in the study of comparative Indo-European linguistics. The language has 2,507,000 speakers in Lithuania (1970) and several thousand speakers in Belorussia and Poland, and until 1945 there were several thousand Lithuanians in East Prussia as well. More than 500,000 Lithuanians live abroad, mostly in the United States. Lithuanian is sharply divided into dialects whose differences are quite marked. The two major ones are Low (or Western) Lithuanian, with three subdialects, and High (or Eastern) Lithuanian, with four subdialects. The Low dialect is spoken by the Lowlanders, who live in the west and along the Baltic Sea; High Lithuanian is spoken by the Highlanders, who live in the eastern (and greater) part of Lithuania. Standard Lithuanian, formed at the end of the 19th and the beginning of the 20th century, is based on the southern subdialect of West High Lithuanian.

Lithuanian, most archaic living Indo-European language

The language most closely related to Lithuanian is Latvian, spoken by 1,342,000 speakers in Latvia (1970) and about 370,000 abroad, mostly in the United States. Latvian is divided into dialects, the major ones being the Central dialect, Livonian (also called Tahmian, or West Latvian), and High (or East) Latvian. Standard Latvian, established at the end of the 19th and the beginning of the 20th century, is based on the Central dialect.

By the 16th century the Selonians, Semigallians, and Curonians, who lived in areas of Latvia and Lithuania, had completely lost their national identities and were assimilated by the Latvians and the Lithuanians. They left no written records. Nor did the Yotvingians (or Suduvians), who lived in southwest Lithuania and farther to the south (in the territory of the present-day Poland). They became extinct around the 16th–17th century, being assimilated by the Lithuanians in the north and the Slavs in the south. Information on the extinct Baltic languages is extremely scarce (mostly place-names). Only Old Prussian, of all the extinct Baltic languages, left any written records, and they are quite poor. The Prussians lived in East Prussia (*i.e.*, between the lower reaches of the Vistula and Neman [Lithuanian Nemunas] rivers on the Baltic coast). They became extinct (*i.e.*, were assimilated by the Germans) at the beginning of the 18th century.

Extinction of several Baltic languages

Linguistically, the Yotvingians were very closely related to the Prussians. They made up one ethnic Baltic group, commonly called the Western Balts, as opposed to the so-called Eastern Balts—the Lithuanians, Latvians, Selonians, Semigallians, and Curonians. The traditional terms Western Balts and Eastern Balts are inaccurate when used for all of the Balts—*i.e.*, including the Balts for whose languages there are no records (the D-Balts). These Balts, who were assimilated by Slavs in the 7th–14th century, lived in the upper reaches of the Dnepr.

Historical survey. Proto-Baltic, the ancestral Baltic language from which the various known languages evolved, developed from the dialects of the northern area of Proto-Indo-European. These dialects also included the Slavic and Germanic protolanguages (and possibly also Tocharian). The quite close historic relationship of the Baltic, Slavic, and Germanic languages is shown by the fact that they alone of all the Indo-European languages have the sound *m* in the dative plural ending (*e.g.*, Lithuanian *vilkà-m-s* "wolf," Common Slavic **vilko-m-ŭ*, Gothic *wulf-am*). (An asterisk [*] indicates that the following sound or word is unattested and has been reconstructed as a hypothetical linguistic form.) This relation-

Features common to Baltic, Slavic, and Germanic

ship is suggested not only by the morphology and word-formation but also by the vocabulary—*e.g.*, Lithuanian *draũgas* (Latvian *dràugs*) "friend," Old Church Slavonic *drugŭ*, Gothic *driugan* "to fulfill military service"; Lithuanian *vãškas* (Latvian *vasks*) "wax," Russian *vosk*, Old High German *wahs*. Probably the earlier close contact of the Balts and the Slavs with the Germanic tribes broke off around the 2nd millennium BC, when the Balts moved from the south (not, however, losing contact with the Slavs) and settled a large area of the eastern coast of the Baltic Sea and the upper reaches of the Dnepr.

Relationship between Baltic and Slavic. Because contact between the Balts and Slavs from the time of Proto-Indo-European was never broken off, it is understandable that Baltic and Slavic should share more linguistic features than any of the other Indo-European languages. Thus, Indo-European **eu* passed to Baltic *jau* and Common Slavic **jau* (which became *ju*)—*e.g.*, Lithuanian *liáudis* "people," Latvian *làudis*, Old Church Slavonic *ljudije*. Tonal correspondences are found between Lithuanian and Serbo-Croatian (a Slavic language of Yugoslavia), and there are also similarities in stress; *e.g.*, Lithuanian *dűmai* "smoke" and Russian *dym* have the stress on the root, as do Lithuanian *rañką* "hand" (accusative singular) and Russian *rúku*, while both Lithuanian *rankà* "hand" (nominative singular) and Russian *ruká* are stressed on the second syllable.

Baltic and Slavic have specific morphological features in common. Among them, for example, is the genitive plural form. In Lithuanian, *műsų* "of us" (= Latvian *mūsu*), evolved from the older form **nūsōn*, which comes from Baltic **nōsōn* and corresponds to the genitive plural form in Common Slavic, **nōsōn*, from which developed Old Church Slavonic *nasŭ* "of us." Baltic also shares some syntactic features with Slavic; *e.g.*, the genitive case is used in place of the accusative with verbs expressing negation (Lithuanian *jis* nieko *nežino* "he does not know anything," Latvian *viņš* nekā *nezin*, Russian *on* ničego *ne znajet*). There are also many lexical items common to Baltic and Slavic. More than 100 words are common in their form and meaning to Baltic and Slavic alone, among them Lithuanian *bėgu* "I run," Latvian *bēgu*, Old Church Slavonic *běgǫ*; Lithuanian *liepa* "linden tree," Latvian *liẽpa*, Old Prussian *lipe*, Old Church Slavonic *lipa*; Lithuanian *rãgas* "horn," Latvian *rags*, Old Prussian *ragis*, Old Church Slavonic *rogŭ*.

In addition to these features common to all the Baltic and Slavic languages, there are certain quite archaic features that Slavic has in common with Lithuanian and Latvian but not with Old Prussian. The most striking example is the genitive singular ending in Lithuanian *vìlk-o* = Latvian *vìlk-a* "of a wolf," which comes from Baltic **-ō*, historically paralleled by the genitive singular ending in Common Slavic **vĭlk-ā*. Old Prussian, however, has a different ending for the same inflection (*deiw-as* "of God"). In certain instances the Slavic languages, differing from Lithuanian and Latvian, come closer to Old Prussian; *e.g.*, the Prussian possessive pronouns *mais* "my, mine," *twais* "your, yours," *swais* "one's own" are different from Lithuanian *mãnas, tãvas, sãvas* and from Latvian *mans, tavs, savs* but similar to Old Church Slavonic *mojĭ, tvojĭ, svojĭ*.

It is possible to conclude that there was close contact between the Baltic and Slavic protolanguages at the time when they began to develop as independent groups (*i.e.*, from about the 2nd millennium BC) and that the Proto-Slavic area might have been a part of peripheral Proto-Baltic, although a specific part. That is, Proto-Slavic at that time was in direct contact with both the corresponding dialects of the peripheral Proto-Baltic area (*e.g.*, with Proto-Prussian) and the corresponding dialects of the central Proto-Baltic area. All this shows that the Proto-Slavic area of that time (south of the Pripyat River) was much smaller than the Proto-Baltic area. Proto-Slavic began to develop as a separate linguistic entity in the 2nd millennium BC and was to remain quite unified for a long time to come. Proto-Baltic, however, besides developing into an independent linguistic unit in the 2nd millennium BC, also began gradually to split. Among other things, the size of the Proto-Baltic area had an influence on the development of Proto-Baltic in that it considerably reduced contact between its dialects (see also SLAVIC LANGUAGES).

Development of the individual Baltic languages. By the middle of the 1st millennium BC, the Proto-Baltic area was already quite sharply split into dialects. From the middle of the 1st millennium AD, the Baltic language area began to become considerably smaller; at that time the greater part of Baltic territory, the eastern part, began to be inhabited by Slavs migrating from the south. The Balts there became gradually assimilated by the Slavs; complete assimilation probably occurred around the 14th century. One of these Baltic tribes, the Galindians (Goljadĭ), is mentioned in a chronicle as late as the 12th century. The protolanguage of the so-called Eastern Balts split into Lithuanian and Latvian (Latgalian) around the 7th century. The other languages of the so-called Eastern Balts became separated probably at the same time. Selonian and Semigallian could have been transitional languages between Lithuanian and Latvian. Only Curonian, which some consider to be a transitional language between East and West Baltic, might have developed somewhat earlier. Moreover, the name of the Curonians occurs in historical sources earlier (AD 853: Latin Cori) than the names of the other tribes of the so-called Eastern Balts.

Split between Lithuanian and Latvian

Old Prussian. In historical sources the Prussians are called Aistians from the 1st century AD (by Tacitus) until the 9th century AD (by the Anglo-Saxon seafarer Wulfstan). They are referred to by their own name (by a Bavarian geographer using the form Bruzi, "Prussians") for the first time in the 9th century AD. About 1230 the Teutonic Order began to plunder the lands of the Prussians and finally conquered the Prussians and the Yotvingians (Suduvians) in 1283. From that time the slow extinction of the two Baltic groups began, with the Germanization of the Prussians being completed at the beginning of the 18th century.

The earliest Old Prussian (and, for that matter, Baltic) written record is a German–Prussian vocabulary—the so-called Elbing vocabulary, compiled about 1300 and extant in a copy dated around 1400. This vocabulary, consisting of 802 Old Prussian words (and the same number of German words), was written in a South Prussian dialect (in Pomesania). Somewhat poorer than the Elbing vocabulary is the vocabulary compiled by Simon Grunau, consisting of 100 Old Prussian (and German) words, written between 1517 and 1526. The most important Old Prussian written records are the three catechisms of the 16th century based on the dialects of Sambia and translated from the German; the first two catechisms, which are very short and anonymous, date from 1545, and the third catechism, or *Enchiridion*, dates from 1561 and was translated by Abelis Vilis (Abel Will), a pastor of the church at Pobeten (Pabečiai; modern Romanovo). The language of all the Old Prussian catechisms is rather poor: the translations are excessively literal, and there are many errors in language and orthography. In spite of this, it is from these Old Prussian catechisms that scholars can learn most about the Old Prussian language.

Lithuanian. Lithuanians are first mentioned in historical sources in AD 1009. Old Russian (more precisely, an East Slavic language based primarily on Belorussian), Latin, and Polish were used in official matters in the Grand Duchy of Lithuania, which was established in the mid-13th century and lasted until the 18th century. Lithuanian writings begin to appear in the 16th century, first in East Prussia (where many Lithuanians lived) and, somewhat later, in the Grand Duchy of Lithuania. In East Prussia, a quite uniform written Lithuanian language, based on the West High Lithuanian dialect, had already been established by the second half of the 17th century. In Lithuania, however, a uniform written Lithuanian came into use only at the end of the 19th and the beginning of the 20th century—*i.e.*, when a standard Lithuanian language, based on the (Southern) West High Lithuanian dialect (spoken in both East Prussia and Lithuania), was established. Martynas Mažvydas (died 1563),

Establishment of uniform written Lithuanian

who published the first Lithuanian book (a catechism) in Königsberg (Lithuanian Karaliaučius; modern Kaliningrad) in the year 1547, is purported to be the first person to use Lithuanian as a written language. Others, in particular Baltramiejus Vilentas, Jonas Bretkūnas, and the pastor-poet Kristijonas Donelaitis, also took part in the formation and standardization of a written Lithuanian language in the 16th–18th century in East Prussia. Great influence was exerted by the first grammars of Lithuanian, by Danielius Kleinas (1653 and 1654), and the works of Donelaitis (1714–80), the first Lithuanian writer to become well known. In the Grand Duchy of Lithuania the first to use Lithuanian as a written language is held to be Mikalojus Daukša (died 1613), who published a catechism in 1595 and a prayer book (*Postilė*) in 1599. Among later writers who helped to standardize written Lithuanian were Konstantinas Sirvydas, who prepared the first dictionary of Lithuanian (1629), Jonas Jaknavičius (1598–1668), and Saliamonas Slavočinskis (17th century). The works of Daukša and Sirvydas in particular, based on the Middle and East High Lithuanian dialects, did much toward establishing the practice of drawing on the various dialects in the creation of a written Lithuanian. This tradition became weakened in the 18th century but was again revived at the beginning of the 19th, when the formation of a standard Lithuanian was undertaken. The practice became most apparent at the end of the 19th and the beginning of the 20th century, during the establishment of Standard Lithuanian. The process of the mixing and levelling of the Lithuanian dialects started at the beginning of the 20th century because of the influence of a standard language, and it was especially intensified after the creation of the Lithuanian Soviet Socialist Republic in 1940. Standard Lithuanian is the official language of the Lithuanian S.S.R., as it was of the Republic of Lithuania (from 1918).

Latvian. The Latvian (Latgalian) people achieved a separate identity around the 16th century AD, when they completely assimilated the other Balts, as well as a greater part of the Livs (also called Livonians, Livians), who are of Finnic descent and live on Latvian territory. As a result of the conquering of Latvian territory by the German Knights of the Sword by 1290, close contact between all of the so-called Eastern Balts (the Latvians with the Lithuanians as well) was considerably weakened for a long period of time. The first Latvian book was the *Catechismus Catholicorum* of 1585. In 1638 the first Latvian (–German) dictionary, by Georgius Mancelius, appeared; the first grammar of the Latvian language, by Johann Georg Rehehausen, was published in 1644; and a Latvian translation of the Bible was published in 1685. The Latvian writings of the 16th–18th century are translations of religious works, as are the Lithuanian. The language of these Latvian works, however, is somewhat poorer than that of the Lithuanian writings of the same period. The works of the Latvians Juris Alunāns (1832–64) and Atis Kronvalds (1837–75) exerted a great influence on the development of a standard Latvian language, based on the Central dialect, at the beginning of the 19th century. Standard Latvian was finally established at the end of the 19th and the beginning of the 20th century, and the levelling influence of this standard language on the Latvian dialects began at this time. Standard Latvian is the official language of the Latvian S.S.R.

Early Latvian books

Characteristics of the Baltic languages. All of the Baltic languages are inflected. Old Prussian is the most archaic of the recorded Baltic languages (although it also has innovations of its own), and it differs considerably from Lithuanian and Latvian.

Old Prussian. In contrast to Lithuanian and Latvian, Old Prussian retained the Baltic diphthong *ei*—Old Prussian *deiws* "God," Lithuanian *diẽvas*, Latvian *dìevs;* Old Prussian *deinan* "day" (accusative singular), Lithuanian *dienà*, Latvian *dìena*. In place of Lithuanian *š* and *ž* (from Indo-European **ḱ*, **ǵ*, and **ǵh*), however, Old Prussian, like Latvian (as well as Curonian, Semigallian, and Selonian), has *s* and *z*—thus, Old Prussian *assis* "axle," Latvian *ass*, Lithuanian *ašìs;* Old Prussian (*po*)*sinnat* "to confess," Latvian *zināt*, Lithuanian *žinóti* "to know." The cluster *s* + *j* (and *z* + *j*) in Old Prussian, as in Latvian, passed to *š* (and *ž*): Old Prussian *schan* (from **sjan*) "this" (accusative singular feminine), Latvian *šùo* "this," Lithuanian *šią*. In contrast to Lithuanian and Latvian, Old Prussian did not replace the clusters *t* + *j* and *d* + *j* with affricate sounds (begun with complete stoppage of the breath stream from the lungs and released with incomplete closure and friction): Old Prussian *median* "forest," Lithuanian *medžias*, Latvian *mežs*.

Word stress was free in Old Prussian, as it is in Lithuanian (in contrast to Latvian, in which the stress is predictable and falls on the first syllable). Old Prussian also made use of intonations (tones), the character of which is similar to that of the Latvian (*i.e.*, more archaic than that of Lithuanian intonations). The Proto-Baltic circumflex intonation corresponds to the falling tone in Old Prussian, while the acute intonation corresponds to the rising tone.

Old Prussian morphology

Old Prussian, moreover, had a substantive neuter gender, lost by Lithuanian and Latvian: Old Prussian *assaran* "lake," Lithuanian *ežeras*, Latvian *ezers;* Old Prussian *lunkan* "bast," Lithuanian *lùnkas*, Latvian *lūks*. It differs in morphology from Lithuanian and Latvian in more than one instance—*e.g.*, in the genitive singular ending, Old Prussian *deiw-as* "of God" (Lithuanian *dievo* = Latvian *diev-a*) and, in the dative singular, Old Prussian *tebbei* "to you" (Lithuanian *tavi* = Latvian *tev*), among others. Old Prussian did not have the dual number, only the singular and plural. Nouns were declined according to seven types. There were five cases: nominative, genitive, dative, accusative, and vocative. All verbs had three separate forms in the plural, but not in the singular. The 3rd person was the same in both the singular and the plural. There were three tenses: present, preterite, and future.

In vocabulary Old Prussian is quite similar to Lithuanian and Latvian (closer to Lithuanian than Latvian). It should be emphasized, however, that Old Prussian differs from Lithuanian and Latvian in that it retained a greater number of archaisms than either.

Comparison of Lithuanian and Latvian. The differences between Lithuanian and Latvian can be summarized in very broad terms by saying that Lithuanian is far more archaic than Latvian and that modern written Lithuanian could in many instances serve as a "protolanguage" for it. For example, Lithuanian has quite faithfully preserved the old sound combinations *an, en, in, un* (the same is true of Old Prussian, Curonian, Selonian, and, possibly, Semigallian), while they have passed in every case to *uo, ie, ī, ū* in Latvian; thus, Lithuanian *rankà* (Old Prussian *rancko*) = Latvian *rùoka* "hand," Lithuanian *peñktas* (Old Prussian *penckts*) = Latvian *piekt(ai)s* "fifth," Lithuanian *pìnti* = Latvian *pīt* "to weave, to twine," and Lithuanian *jùngas* = Latvian *jūgs* "yoke." The diphthongs *ei* (as well as *ai*) and *au* in final position were monophthongized and later shortened in Latvian—*e.g.*, Lithuanian *ved-eĩ* (2nd person singular preterite) = Latvian **ved-ie*, which became *ved-i* "you led"; Lithuanian *med-aũs* = Latvian **med-uos*, which became *med-us* "of honey." Long vowels at the end of polysyllabic words have been shortened in Latvian, and short vowels have been dropped—*e.g.*, Latvian *sak-a* "says" (which derives from **-ā*) = Lithuanian *sãk-o*, Latvian *pel-e* "mouse" (from **-ē*) = Lithuanian *pel-ẽ*, Latvian *vìlk-u* "wolf" (from **-uo*) = Lithuanian *vĩlk-ą*, Latvian *daikts* "thing" (from **-ăs*) = Lithuanian *dáiktas*, and Latvian *nakts* "night" (from **-ĭs*) = Lithuanian *naktìs*. Palatalized *k* and *g*, formed with the blade of the tongue closer to the hard palate than nonpalatalized *k* and *g*, were retained in Lithuanian (as in Old Prussian and Semigallian) but changed to *c* (pronounced like *ts*) and *dz* in Latvian (as in Selonian and Curonian): Lithuanian *ãkys* "eyes" (Old Prussian *ackis*) = Latvian *acis*, and Lithuanian *gérvė* "crane" (Old Prussian *gerwe*) = Latvian *dzeŗve*. The change of the old clusters *t* + *j* and *d* + *j* progressed further in Latvian. Most Lithuanian dialects have *č* (as *ch* as in "church") and *dž* (as *j* in "jam"), whereas Latvian has *š* (as *sh* in "shore") and *ž* (as *z* in "azure")—*e.g.*, Lithuanian *trẽčias* = Latvian

trešs "third"; Lithuanian *bríedžiai* = Latvian *briẽži* "elks." Another difference between Lithuanian and Latvian is that, instead of Lithuanian *š* and *ž*, Latvian (like Selonian, Semigallian, Curonian, and Old Prussian) has *s* and *z* sounds—*e.g.*, Lithuanian *širdìs* = Latvian *sirds* "heart"; Lithuanian *žiemà* = Latvian *zìema* "winter." Proto-Latvian (and Prussian) *s* + *j* and *z* + *j* have passed to *š* and *ž*: Latvian *šūt* "to sew" = Lithuanian *siūti;* Latvian *eža* "of a hedgehog" (from Latvian **ezjā*) = Lithuanian *ežio*. Lithuanian has retained the initial clusters *pj* and *bj*, which in Latvian (and similarly in Slavic) have passed to *pļ* and *bļ*—*e.g.*, Lithuanian *piáuti* (*pi* is pronounced as *pj*) = Latvian *pļaũt* "to cut"; Lithuanian *biaurùs* = Latvian *bļaũrs* "hideous, nasty."

Stress and tone in Lithuanian and Latvian

Lithuanian has a free stress in contrast to Latvian fixed stress, which occurs on the first syllable. Latvian is more archaic than Lithuanian in the intonations inherited from Proto-Baltic: the Proto-Baltic circumflex intonation has preserved its falling character in Latvian (it became rising in Lithuanian), and the Proto-Baltic acute intonation retained its rising character (it is falling in Lithuanian), although in some cases (because of stress retraction) it has been changed to the broken intonation; *e.g.*, Latvian *pìrsts* "finger" = Lithuanian *pir̃štas* (falling in Latvian and rising in Lithuanian from the Proto-Baltic circumflex), Latvian *vārna* "crow" = Lithuanian *várna* (the rising or extended intonation in Latvian and the falling intonation in Lithuanian from the Proto-Baltic acute intonation), Latvian *zâle* "grass" (the Latvian broken intonation from the Proto-Baltic acute intonation through stress retraction).

There are really no differences in the older morphological features between Lithuanian and Latvian if one does not take into account innovations such as the Latvian debitive verb form (*man ir jāmācās* "I must study" or "it is necessary for me to study") and the Lithuanian frequentative past (*jie eidavo* "they used to go"). Lithuanian and Latvian have two grammatical genders (masculine and feminine) and two numbers (singular and plural), while some Lithuanian dialects also have the dual number. Both Lithuanian and Latvian have seven cases—nominative, genitive, dative, accusative, instrumental, locative, vocative. Standard Lithuanian has five declensions of nouns with 12 inflectional types; Latvian has six declensions with eight inflectional types. Lithuanian adjectives have three declensions, Latvian adjectives have one. The comparison of adjectives in the two languages is different. Both Lithuanian and Latvian have indefinite adjectives (Lithuanian *mãžas*, masculine, *mažà*, feminine, "a small one" = Latvian *mazs*, *maza*) and definite adjectives (Lithuanian *mažàsis*, *mažóji* "the small one" = Latvian *mazais*, *mazā*) with their own specific inflection. The verb in Lithuanian and Latvian has three conjugations (genetically different). There are three persons, the third of which is the same (apparently from the time of Proto-Indo-European) in both the singular and the plural (as well as the dual); for example:

Lithuanian		Latvian
Singular		Singular
1. *kertù*	("I cut, I strike")	1. *certu*
2. *kertì*	("you cut, you strike")	2. *certi*
3. *kir̃ta*	("he cuts, he strikes")	3. *cert*
Plural		Plural
1. *kir̃tame*	("we cut, we strike")	1. *certam*
2. *kir̃tate*	("you cut, you strike")	2. *certat*
3. *kir̃ta*	("they cut, they strike")	3. *cert*

Verb forms

The verb in Lithuanian and Latvian has three tenses (present, preterite, future)—*e.g.*, Lithuanian *kertù*, Latvian *certu* (present); Lithuanian *kirtaũ*, Latvian *cirtu* (preterite); Lithuanian *kir̃šu*, Latvian *ciršu* (future). In contrast to Latvian, Lithuanian also has a frequentative past tense—*e.g.*, *kir̃šdavau* "I used to cut, strike." Lithuanian and Latvian have many compound tense forms, compounded from the forms of the verb *būti* "to be" and participles. There are several moods in both languages, although they are different. The system of participles (active and passive) in Lithuanian and Latvian is quite similar, although complicated—*e.g.*, Lithuanian *kertą̃s*, Latvian *certuošs* (present active); Lithuanian *kir̃tamas*, Latvian *certams* (present passive). Lithuanian and Latvian sentence word order is quite free, and, in general, the syntax of both languages is quite similar.

Words are formed in Lithuanian and Latvian basically by means of suffixes, prefixes, and compounding. The languages are very similar in their early vocabulary, and the differences that do occur tend to be more of a semantic nature—*e.g.*, Lithuanian *móša* "husband's sister" = Latvian *māsa* "sister"; Lithuanian *žam̃bas* "corner, angle (acute)" = Latvian *zùobs* "tooth." Some older lexical differences do occur, however (*e.g.*, Lithuanian *kraũjas* = Latvian *asins* "blood"; Lithuanian *sūnùs* = Latvian *dēls* "son"). In the newer vocabulary, there are now many differences between Lithuanian and Latvian.

Loanwords in Baltic. The Baltic languages have loanwords from the Slavic languages (*e.g.*, Old Prussian *curtis* "hunting dog," Lithuanian *kùrtas*, Latvian *kuŕts* come from Slavic [*cf.* Polish *chart*]; Lithuanian *muĩlas* "soap" [*cf.* Russian *mylo*]; Latvian *suods* "punishment, penalty" [*cf.* Russian *sud*]). There are also a few loanwords from Gothic (*e.g.*, Old Prussian *ylo* "awl," Lithuanian *ýla*, Latvian *īlens*) and possibly from Scandinavian, and many from German, especially in Old Prussian and Latvian, as a consequence of the German colonization of the Prussians, Latvians, and, in part, of the Lithuanians in the 13th century.

Relation between the Baltic and Finnic languages

The Balts first came in close contact with their northern neighbours, the Baltic Finns, about 2000 BC. This contact left traces in both the Baltic and the Finnic languages, perhaps most clearly in the vocabulary. Baltic has very few early loanwords from Finnic, but Finnic has many early loans from Baltic. Latvian, with many loanwords from Livian (Livonian) and Estonian (both Finnic languages), has been more influenced by Finnic than has any other recorded Baltic language.

Orthography. The Lithuanian alphabet is based on the roman (Latin) alphabet. It has 33 letters, several employing diacritical marks (*ą, č, ę, ė, į, š, ų, ū, ž*), and is phonetic (*i.e.*, written as it is pronounced). In linguistic literature ′ is used for falling tones, and ~ for rising tones; the grave accent (`) is used for short, stressed vowels. The Latvian alphabet has 33 letters, 11 with diacritical marks: *ā, č, ē, ģ, ī, ķ, ļ, ņ, š, ū, ž*. A macron (¯) over a vowel indicates that it is long. In linguistic literature the following accents are used for the Latvian intonations: ` (falling), ~ (extended, or rising), ˆ (broken).

The Old Prussian orthography is almost wholly based on the German orthography of that time and is quite inconsistent. Furthermore, every Old Prussian written record has its own specific orthography.

BIBLIOGRAPHY. There are very few works on the Baltic languages in English aside from LEONARDAS DAMBRIUNAS, ANTANAS KLIMAS, and WILLIAM R. SCHMALSTIEG, *Introduction to Modern Lithuanian* (1966); TEREZA BUDINA LAZDINA, *Teach Yourself Latvian* (1966); and JANIS ENDZELINS, *Baltų kalbų garsai ir formos* (1957; Eng. trans., *Comparative Phonology and Morphology of the Baltic Languages*, 1971). *Baltic Linguistics*, ed. by THOMAS F. MAGNER and WILLIAM R. SCHMALSTIEG (1970), is a collection of papers on various aspects of Baltic linguistics.

The following publications are written in German, Lithuanian, Latvian, and Polish. REINHOLD TRAUTMANN, *Die altpreussischen Sprachdenkmäler* (1910); JOHANN ENDZELIN, *Lettische Grammatik* (1922; trans. into Latvian, 1951); KAZIMIERAS BUGA, *Lietuvių kalbos žodynas* (1924–25), the introduction to this dictionary contains much valuable information on the history of the Baltic languages; *Rinktiniai raštai*, 3 vol. (1958–61); JANIS ENDZELINS, *Senprūšu valoda* (1943; trans. into German, 1944, without a glossary); *Ievads baltu filoloģijā* (1945), an introduction to Baltic linguistics, in Latvian; ERNST FRAENKEL, *Die baltischen Sprachen* (1950), a general introduction; ALFRED SENN, "Die Beziehungen des Baltischen zum Slavischen und Germanischen," *Zeitschrift für vergleichende Sprachforschung*, vol. 71 (1954), discusses the relationship of Baltic to Slavic and Germanic; *Mūsdienu latviešu literārās valodas gramatika*, 2 vol. (1959–62), a grammar of the modern Latvian literary language; ARTURS OZOLS, *Veclatviešu rakstu valoda* (1965), treats Old Latvian; JAN OTREBSKI, *Gramatyka języka litewskiego*, 3 vol. (1956–65), a Lithuanian grammar, in Polish; MARTA RUDZITE, *Lat-*

viešu dialektologija (1964), treats Latvian dialectology; *Lietuvių kalbos gramatika*, 2 vol. (1956, 1971), the most authoritative grammar of the Lithuanian language; ZIGMAS ZINKEVICIUS, *Lietuvių dialektologija* (1966), a valuable treatment of Lithuanian dialectology; CHRISTIAN S. STANG, *Vergleichende Grammatik der Baltischen Sprachen* (1966), the only scholarly comparative grammar of the Baltic languages; VYTAUTAS J. MAZIULIS (comp.), *Prūsų kalbos paminklai* (1966), contains and discusses all the photographed Old Prussian texts; ALGIRDAS SABALIAUSKAS, "Lietuvių kalbos leksikos raida," *Lietuvių kalbotyros klausimai*, 8:5–140 (1966), treats the development of the vocabulary of Lithuanian; JONAS PALIONIS, *Lietuvių literatūrinė kalba XVI–XVII a.* (1967), a treatment of the Lithuanian literary language in the 16th and 17th centuries; JONAS KAZLAUSKAS, *Lietuvių kalbos istorinė gramatika* (1968), the only historical grammar of Lithuanian; VYTAUTAS MAZIULIS, *Baltų ir kitų indoeuropiečių kalbų santykiai* (1970), treats the relationship of Baltic and the other Indo-European languages.

The largest dictionaries of Baltic languages are: K. MULENBACHS, *Latviešu valodas vārdnīca*, 4 vol. (1923–32); *Lietuvių kalbos žodynas*, 8 vol. (1941–70); and ERNST FRAENKEL, *Litauisches etymologisches Wörterbuch*, 2 vol. (1955–65).

(V.J.M.)

Baltic Religion

The term Baltic religion covers the religious beliefs and practices of the Balts, ancient inhabitants of the Baltic region of eastern Europe, who spoke languages belonging to the Baltic family of languages.

SOURCES AND PROBLEMS IN THE STUDY OF BALTIC RELIGION

Problems. The study of Baltic religion has developed as an offshoot of the study of Baltic languages—Old Prussian, Latvian, and Lithuanian (see BALTIC LANGUAGES). These form a separate group—the oldest one—of the Indo-European languages, which are closely related to the ancient Indian language Sanskrit.

Although the study of Baltic languages is important in the study of Indo-European linguistics, the study of Baltic religion has not assumed a similar level of importance in the study of comparative religion. In 1875 it was shown that the religious concepts of the Balts, when compared with those of other European peoples, are found to be marked by many older features that agree with Vedic (ancient Indian) and Iranian ideas. At least one scholarly reconstruction of ancient Indo-European religion depended mainly on Baltic religious traditions. International research in Baltic religion has, however, been greatly hindered by the fact that the languages of these small Baltic countries (Latvia and Lithuania) are but little known and because Baltic scholars have been able to work in this field only relatively recently. Thus, a comprehensive review of Baltic religion is possible only on the express understanding that many findings are only hypothetical and require further research. But, as will be seen below, even under these circumstances Baltic religious concepts help greatly in understanding the formation and structure of the oldest phases of Indo-European religion.

Sources of data. There are four main sources of data, each with its own relevance and each requiring its own specific methodology: archaeological material, historical documents, linguistics, including toponymy (the study of the place-names of a region or language), and folklore. Since the last half of the 19th century, archaeological material has furnished much information about burial and sacrificial rites. The remains of sacral buildings have also been found. This material is of special interest in that it corroborates old religious traditions preserved by folklore, which gives added reliability to both of these sources. But archaeological material can at best furnish only a partial and incomplete picture, even though it is meaningful in some respects. Historical documents, already partially compiled and published, could be expected to yield much more information. Their value, however, is made problematic by the fact that all such documents were written by foreigners, mainly Germans who, in the course of their centuries-long eastward expansion, subjugated the Baltic peoples and exterminated some of them. Since the conquerors did not understand the Baltic languages, many documents contain the names of gods and other divinities that are without basis in fact. Baltic religion was viewed dogmatically and negatively in the light of Christian interpretations. Linguistic source material, also compiled by foreigners, shows fewer signs of interpretation, especially in regard to toponymy. Baltic folklore—one of the most extensive folklores of all European peoples—contains the greatest amount of material, especially in the form of *dainas* (short folk songs of four lines each) and folktales. Folklore is especially valuable because it contains many concepts that elsewhere have been lost under the influence of Christianity. Old religious beliefs have persisted because the Germans, after conquering the Baltic lands in the 13th and 14th centuries, made practically no attempt at Christianization and contented themselves with only economic gains. The positive result of this policy is the preservation of old traditions and religious beliefs; some researchers have also noted the similarity between the metrical structure of the *dainas* and that of the Old Indian short verses in the Ṛgveda (a Hindu sacred scripture).

The student of Baltic religion still encounters two difficulties. First, as has been noted, since written documents were established in Christian times, Christian influences in them are inescapable. Such influences cause difficulties and make a critical approach mandatory. Second, after the establishment of political independence of the Baltic countries following World War I, there arose a certain national romanticism that has attempted to identify Baltic culture with that of the ancient Indo-Europeans. Thus, an uncritical approach has led even to the introduction of "gods" that are actually only etymological derivations from the names of Christian saints. On the other hand, those western European scholars who are unfamiliar with the special historical and social circumstances of the Balts have assumed Baltic folklore to be on a level with the thoroughly Christianized western European folklore and thus have underestimated its importance.

MYTHS AND GODS

Cosmology. In the traditions of the Baltic peoples, there are no epic myths about the creation of the world and its structure. This fact is explained by the historical and social circumstances mentioned above, which either have hindered the formation of these types of myths or, more likely, have simply made their preservation possible. Furthermore, there has been no significant research concerning Baltic myths and their intrarelationships. Fragmentary evidence found exclusively in folklore indicates only two complexes of ideas with any certainty: the first concerns the structure of the world, the second the enmity between Saule (the Sun) and Mēness (Latvian; Lithuanian Mėnulis; the Moon).

There is disagreement as to whether the Balts pictured the world as consisting of two regions or of three. The two-region hypothesis seems to be more plausible and is supported by a dualism found frequently in the *dainas*: *šī saule* (literally "this sun") and *viņa saule* (literally "the other sun"). The metaphor *šī saule* symbolizes ordinary everyday human life, while *viņa saule* indicates the invisible world where the sun goes at night, which is also the abode of the dead.

The two-region hypothesis

The evidence does not show conclusively whether it is located in the direction of the setting sun or under the earth, beneath which the sun travels back to the east. The sky is considered to be a mountain, sometimes of stone, and is the residence of the sky gods. Saule rides over the sky in a chariot drawn by a varying number of horses, Mēness rides to be married, and Pērkons (Latvian; Lithuanian Perkūnas; the Thunderer) makes weapons and jewelry in the sky.

The concept that Saule, unseen during the night, makes her way from west to east under the earth so that she can start her course anew over the sky mountain is also familiar. It is also possible to see here the ancient idea of a world ocean on which the earth, as a round plate, swims, an idea that has disappeared under the influence of Christianity.

The notion of a sun tree, or world tree, is one of the most important concepts regarding the cosmos. This tree grows at the edge of the path of Saule, and the setting sun (Saule) hangs her belt on the tree in preparation for rest. It is usually considered to be an oak but is also described as a linden or some other kind of tree. The tree is said to be located in the middle of the world ocean or generally to the west.

The gods. *Dievs.* The Baltic words Latvian *dièvs*, Lithuanian *dievas*, and Old Prussian *deivas* are etymologically related to the Indo-European *deiu̯os;* among others, the Greek Zeus is derived from the same root. It originally meant the physical sky, but already in Old Indian and other religions the sky became personified as an anthropomorphic deity. Dievs, the pre-Christian Baltic name for god, was used by Christian missionaries (and still is) to denote the Christian God. The etymology of the word indicates that the Balts preserved its oldest forms, which is also true of the functions and attributes of the personified Baltic sky god Dievs, who lives on his farmstead on the sky mountain but does not participate in the work of the farm. Importantly, Dievs is a bridegroom who rides together with the other gods to a sky wedding in which his bride is Saule. Dievs' family is a later development; in the family, Dieva dēli (God's Sons) play the primary role. Thus Dievs is pictured as the father of a family of sky gods. Besides such anthropomorphic characteristics, another characteristic that gives Dievs a universal significance may be observed: he appears as the creator of order in the world on the one hand, and as the judge and guardian of moral law on the other. From time to time he leaves the sky mountain and actively takes part in the everyday life of the farmers below. His participation in various yearly festivals is vividly described. In spite of this, the Baltic Dievs is similar to the Old Indian Dyaus, the Greek Zeus, and other personifications of the sky. Such divinities have a tendency, in comparison with other gods of their religions, to recede into a secondary role.

The sky god

Pērkons. In Baltic, as in other Indo-European religions, there is, in addition to Dievs, the Thunderer (Latvian Pērkons, Lithuanian Perkūnas) with quite specific functions. Pērkons is described in the oldest chronicles and in poetic and epic folklore, but, though he is a primary divinity there is no reason to believe that he is the main god. His abode is in the sky, and, like Dievs, he sometimes descends from the sky mountain. He has two main characteristics. First, he is a mighty warrior, metaphorically described as the sky smith, and the scourge of evil. His role as adversary of the devil and other evil spirits is of secondary importance and has been formed to a great extent under the influence of Christian syncretism. Secondly, he is a fertility god, and he controls the rain, an important event in the life of farmers. Various sacrifices were made to him in periods of drought as well as in times of sickness and plague. No other god occupied a place of such importance at the farmer's table during festivals, especially in the fall at harvest time. Like the other sky gods, he also has a family. Even though his daughters are mentioned occasionally, originally he had only sons, and myths depicting sky weddings portray his role vividly, as a bridegroom and as the father in his sons' weddings.

Saule. The sun, Saule, occupies the central place in the pantheon of Baltic gods. The divinity of the sun has been recognized all over the world, and the Balts were no exception. The Baltic description of the sun as divinity is so complete and specific that it was one of the first to be studied by scholars. Of greatest importance is the similarity in both functions and attributes of Saule and the ancient Indian god Sūrya. Similarities between the two gods are so great that, were not the two peoples separated by several thousand miles and several millennia, direct contact between them would be indicated instead of only a common origin.

The sun goddess

The representation of Saule is dualistic in that she is depicted as a mother on one hand, and a daughter on the other. Her attributes are described according to the role she plays. As a daughter she is mentioned only when she is a bride to the other sky gods. But as her daughters frequently are in the same role, it is difficult to differentiate between them. As a mother, however, she is depicted much more extensively and completely. Her farmstead on the sky mountain borders that of Dievs, and both Dieva dēli and Saules meitas (Daughters of the Sun) play and work together. Sometimes Dievs and Saule become enraged at each other because of their respective children, as, for example, when Dieva dēli break the rings of Saules meitas or when Saules meitas shatter the swords of Dieva dēli. Their enmity lasts three days, which some scholars explain through natural phenomena; *i.e.*, the three days before the new moon when Dievs, a substitute for the moon, is not visible.

That Saule, richly described in mythology, also had a cult devoted to her is suggested by the many hymns in her honour. They contain either expressions of thanks for her bounty or prayers seeking her aid, not only in relation to agriculture but to life in general. In agriculture Saule is a sanctifier of the fertility of the fields; in the life of the individual she is a typical sky goddess, interfering in her omniscience. She has human moral characteristics and punishes the immoral and aids the suffering. Though the question of where Saule's places of worship were located is not solved, the occasions for rituals pertaining to Saule have been definitely established, the most important of which was the summer solstice. Besides song, recitative, and dance, a central place in the ceremonies was occupied by a ritual meal, at which cheese and a drink brewed with honey (later beer) were consumed.

Mēness. Mēness, the moon, also belongs to the sky pantheon. Detailed analysis only recently has shown that he has a role as a war god in Baltic religion. Such a role is indicated not only by his dress and accoutrements but especially by his weapons and expressions used in times of war. The influence of syncretism, however, has erased the outlines of his characteristics so far as to make a description of his role and any cult he may have had very difficult. The sky wedding myths furnish a somewhat more complete picture in which he is represented as a conflict-creating rival suitor of Auseklis (the Morning Star).

The moon god

Auseklis, his sons, Dieva dēli, and Saules meitas form a separate group of divinities. Although they are mentioned in the sky myths, they have remained only as personifications of natural phenomena, characterized by the most beautiful metaphors.

It is notable that a common characteristic of the sky gods, and, in fact, of all Baltic divinities, is the express tendency for each to have a family.

All the divinities mentioned above are closely associated with horses: they either ride or are drawn in chariots across the sky mountain and arrive on earth in the same fashion. The number of horses is indeterminate but usually varies from two to five or more. This trait also confirms the close ties between Baltic and Indo-Iranian religions.

Although males form the majority of the sky gods, the chthonic (underworld) divinities are mostly female. In both Latvian and Lithuanian religions the earth is personified and called Earth Mother (Latvian Zemes māte, Lithuanian Žemyna). But the Lithuanians also have Žemėpatis, Earth Master. Latvians in general refer to mothers, Lithuanians to masters. Zemes māte is the only deity in addition to Dievs who is originally responsible for human welfare. Based on the writings of the Roman historian Tacitus, it has been asserted that she is the mother of the other gods, but there is no support for this view in other sources. Under the influence of Christian–pagan syncretism, the Virgin Mary has assumed some of the functions of Zemes māte. Furthermore, some of these functions have been acquired and differentiated by various other later divinities, who, however, have not lost their original chthonic character. Thus, a deity of the dead has developed from Zemes māte, called in Latvian Smilšu māte (Mother of the Sands), Kapu māte (Mother of the Graves), and Veļu māte (Mother of the Ghosts). Libations and sacrifices were offered to Zemes māte.

Such rituals were also performed in connection with the other divinities at a later stage of development. The fertility of the fields is also guaranteed by Jumis, who is symbolized by a double head of grain, and by various mothers, such as Lauka māte (Mother of the Fields), Linu māte (Mother of the Flax), and Mieža māte (Mother of the Barley).

Forest and agricultural deities. A forest divinity, common to all Baltic peoples, is called in Latvian Meža māte (Mother of the Forest, Lithuanian Medeinė). She again has been further differentiated into other divinities, or rather she was given metaphorical appellations with no mythological significance, such as Krūmu māte (Mother of the Bushes), Lazdu māte (Mother of the Hazels), Lapu māte (Mother of the Leaves), Ziedu māte (Mother of the Blossoms), and even Sēņu māte (Mother of the Mushrooms). Forest animals are ruled by the Lithuanian Zvėrinė opposed to the Latvian Meža māte.

The safety and welfare of the farmer's house is cared for by the Latvian Mājas gars (Spirit of the House; Lithuanian Kaukas), which lives in the hearth. Similarly, other farm buildings have their own patrons—Latvian Pirts māte (Mother of the Bathhouse), Rijas māte (Mother of the Threshing House); Lithuanian Gabjauja.

Gods of natural phenomena

Because natural phenomena and processes have often been raised to the level of divinities, there is a large number of beautifully described lesser mythological beings whose functions are either very limited or completely denoted by their names. Water deities are Latvian Jūras māte (Mother of the Sea), Ūdens māte (Mother of the Waters), Upes māte (Mother of the Rivers), and Bangu māte (Mother of the Waves; Lithuanian Bangpūtys), while atmospheric deities are Latvian Vēja māte (Mother of the Wind), Lithuanian Vėjopatis (Master of the Wind), Latvian Lietus māte (Mother of the Rain), Miglas māte (Mother of the Fog), and Sniega māte (Mother of the Snow). Even greater is the number of those beings related to human activities, but only their names are still to be found, for example Miega māte (Mother of Sleep) and Tirgus māte (Mother of the Market).

Goddess of destiny. Because of peculiarities of the source materials, it is difficult to determine whether the goddess of destiny, Laima (from the root word *laime*, meaning "happiness" and "luck"), originally had the same importance in Baltic religion as later, or whether her eminence is due to the specific historical circumstances of each of the Baltic peoples. In any case, a wide collection of material concerning Laima is available. The real ruler of human fate, she is mentioned frequently together with Dievs in connection with the process of creation. Although Laima determines a man's unchangeable destiny at the moment of his birth, he can still lead his life well or badly within the limits prescribed by her. She also determines the moment of a person's death, sometimes even arguing about it with Dievs.

The devil. The devil, Velns, has a well-defined role, which is rarely documented so well in the folklore of other peoples. Besides the usual outer features, several characteristics are especially emphasized. Velns, for instance, is a stupid devil. In addition, the Balts are the only colonialized people in Europe who have preserved a large amount of folklore that in different variations and situations portrays the devil as a German landlord. Another evil being is the Latvian Vilkacis, Lithuanian Vilkatas, who corresponds to the werewolf in the traditions of other peoples. The belief that the dead do not leave this world completely is the basis for both good and evil spirits. As good spirits the dead return to the living as invisible beings (Latvian *velis*, Lithuanian *vėlės*), but as evil ones they return as persecutors and misleaders (Latvian *vadātājs*, Lithuanian *vaidilas*).

PRACTICES, CULTS, AND INSTITUTIONS

Temples and other holy places. Recent archaeological excavations have indicated the existence of temples made of wood. The only remains of these temples are postholes. Such temples were circular, approximately five metres in diameter, in the centre of which a statue of a god may have been erected. At present, however, the existence of such temples must be regarded only as conjecture within the realm of probability. On the other hand, the existence of open-air holy places or sites of worship among the Balts is confirmed by both the earliest historical documents and folklore. Such places were holy groves, called *alka* in Lithuanian. Later the word came to mean any holy place or site of worship (Lithuanian *alkvietė*). Considerable research has shown that the usual sites were little hills, where the populace gathered and sacrificed during holy festivals, all of which supports the idea that wooden buildings could have been built at these sites.

Sites of worship

Other holy places were also recognized. The most important of these appear to be bathhouses, whose function some researchers have compared to that of churches in Christianity. A large amount of evidence indicates that religious–magical rites, from birth ceremonies to funerals, were performed in such bathhouses. There are various opinions as to whether the so-called holy corner (*heilige Hinterecke*); *i.e.*, the dark corner of a peasant's house in which a deity or patron lives, belongs to pre-Christian concepts or not. On the other hand, various places in the house proper, such as the hearth and the doorstep, were considered to be abodes of spirits. In general, the more important work sites each had its own guardian spirit. Sacrifices were performed at each spot to assure successful completion of work. Because they supplied the farmstead with water, streams and rivers were also especially important.

Religious personages. There is no reliable information that the Balts had a priestly class, let alone religious hierarchy. The 11th-century German historian Adam of Bremen, in describing conflicts between Christian missionaries and Latvians, said that "every house is filled with seers, augurers, and necromancers," which indicates that the Balts had sacral persons, probably the patriarchs of large extended families or heads of clans. As even 18th-century church inspection records show, the Christian Church had great difficulty in curbing their influence, especially within their clans. Their religious functions were twofold. First, they were responsible for the welfare and means of existence of the people through the performance of appropriate rites both at work sites and during the holy festivals. Second, they assured that the proper procedure would be followed in rituals connected with the important occasions of human life, such as birth, marriage, and death. In the syncretistic amalgam of Christianity and the religion of the Balts, those persons were called sorcerers (*Zauberer*) and, according to church records, were treated by the Balts with the same reverence as bishops were by Christians.

Sacred times. Special rites evolved for the festivals of the summer solstice and the harvest, while other rites were used specifically for beginning various kinds of spring work. Such spring work included sending farm animals to pasture or horses to forage for the first time, plowing the first furrow, and starting the first spring planting. The birth of a child was especially noted; it usually took place in the bathhouse or some other quiet spot. Laima was responsible for both mother and child. One birth rite, called *pirtīžas*, was a special sacral meal in which only women took part. Marriage rites were quite extensive and corresponded closely to similar Old Indian ceremonies. Fire and bread had special importance and were taken along to the house of the newly married couple. These rites persisted until quite late and were to be seen even at the end of the 19th century, though in many cases only as games. In this connection, fire in general occupied a central place in Baltic religion. Considered holy, it was worshipped, and sacrifices were offered to it.

Death rites and customs

It seems unbelievable that even as late as 1377 and 1382, respectively, the Lithuanian king Algirdas and his brother Kęstutis could still be buried according to the old traditions in a Christian Europe; dressed in silver and gold, they were burned in funeral pyres together with their best possessions, horses, hunting dogs, birds, and weapons. In spite of a ban by the church and subsequent persecution, this rite still persisted in the 15th century. The tenacious preservation of this ancient Indo-European

ritual casts light on other features of Baltic religion. Chronicles relate that Lithuanians, after losing a battle, joyfully committed suicide; this was also true of the widows of soldiers killed in battle. Such voluntary immolation and the articles buried with the dead are evidence of a belief in life after death. It is said that at the funeral of a nobleman his companions threw lynx and bear claws into the fire to aid his climb up the mountain to God, an indication of Christian influence. Archaeological excavations have also yielded evidence of fire funeral rites: the bones of men and animals, metal jewelry, and weapons found at the sites of the funeral pyres.

In funeral rites several different phases are discernible during the period between death and burning. The deceased was laid out in his house for a longer or shorter period depending on his social position and the size of his estate. During this time a meal lasting several days was held for the deceased's relatives and friends. In the course of the festivities the participants conducted fights on horseback. Lamentations, leave-takings, and praises of the deceased, as well as wishes for a safe journey to the world of the dead, accompanied the corpse on the way to the funeral pyre. In spite of persecution by the church, the tradition of lamentation has lasted until modern times, though in a somewhat modified form. One of the peculiarities of Baltic funeral rites was their similarity to wedding ceremonies. The corpse and a partner selected from the living were dressed in elaborate wedding costumes, wedding songs were sung, and dancing took place. The basis of these ceremonies was the belief that the dead anticipate a new companion with the same joy as the living do a new in-law. The corpse's living partner was a symbolical substitute for the new comrade awaited by the dead.

The above suggests a dominant concept in Baltic religious thought, namely, that the boundary between the worlds of the dead and the living was not real. The dead continued to live invisibly and were present at all important occasions. A place was set for them at the festival table and no one else might sit there. The extensive practice of feeding the dead was a consequence of the concept that the living were responsible for their welfare. Originally, their food must have been placed at the hearth. In later development, meals for the dead were also placed in other buildings, such as the threshing house or the bathhouse. Under the influence of Christianity, these living dead (Latvian *velis*, Lithuanian *vėlės*) have been confused with the devil. A widespread view was that the souls of the dead dwell in the *zalktis* (Latvian; Lithuanian *žaltys;* "green snake"); thus special care was taken in its feeding. But the *zalktis* was also closely associated with fertility and sexual symbolism.

SUMMARY AND CONCLUSIONS

Three main characteristics

Three main characteristics are discernible in Baltic religion. First, it is a typical astral religion in which the personified sky and main heavenly bodies play a major role. Saule, Mēness, Auseklis, and other gods have their own traits, frequently based on counterparts in nature. Although they are all related as one family, their roles within the family are varied. Depending on the cult or the plot of the myth, each divinity can assume various functions; religious man, in general, does not experience such fluctuations as a contradiction. The second main characteristic is personified happiness, luck, and fate, Laima, who has assumed the role of a goddess of destiny. Because happiness is not an external, datable event, other gods besides Laima can help determine happiness in human life. The differentiation of Laima's functions has led to the establishment of some of her functions as independent entities with sometimes a poetic, sometimes a religious, meaning. The concept of destiny in Baltic religion has not, however, resulted in passive resignation or quietism but rather full exploitation of opportunities within the limits set by it. The third characteristic is the fertility cult. Here the primary force is the personified earth, called Mother, with all her functions and characteristics. It must be understood that the concept of a fertility cult entails a wider meaning, that of the assurance of human welfare in general.

These three main typological traits hardly describe Baltic religion in all of its details and nuances. The religion can also be analyzed as having two strata: one, expressed in the above three features, can be called the stable surface layer. The second, visible below the first, contains only the outlines of undifferentiated, fluid mythological and religious beings that, because of their vague character, appear in various guises and have no stable role. They are the countless house, field, and wood spirits of the nature myths.

Baltic religion, typologically, is an agricultural religion, and it is useless to speculate whether any other basis—such as nomadism, hunting, or fishing—can be found for it, because no information regarding such possibilities can be derived from any source. The amorphous agricultural clan defines the nature of Baltic religion. The farmer's gods are also farmers, though they live in great glory on their farmsteads on the sky mountain, from which they descend to help their lesser image—man. If necessary, Dievs, Saule, and Laima dress themselves in farmer's clothes and walk his fields with him. This religion does not recognize contemplation or mysticism but rather exhibits a healthy rationalism. Just as the gods are part of the cosmic order and are responsible for its maintenance, so man obeys it and becomes a part of the divine rhythm of life set by the gods. In this way, man crosses the boundary that otherwise separates him from the world of the gods. Various specific historical circumstances explain why the Balts, in their language as well as in their religion, have preserved many elements undoubtedly belonging to the oldest phase of Indo-European religion.

BIBLIOGRAPHY. H. BERTULEIT, "Das Religionswesen der alten Preussen mit litauischlettischen Parallelen," in *Prussia*, vol. 25 (1924), is the only complete review of the Old Prussian religion. A critical examination of sources and research may be found in HARALDS BIEZAIS, "Die Religionsquellen der baltischen Völker und die Ergebnisse der bisherigen Forschungen," *Arv*, 9:65–128 (1953). The most important and complete bibliography is ZENONAS IVINSKIS, *Senoves lietuviu religijos bibliografija* (1938). For a comprehensive collection of historic records of the Prussian, Lithuanian, and Lettish religion, see W. MANNHARDT, *Letto-Preussische Götterlehre* (1936). HARALDS BIEZAIS, *Die Hauptgöttinnen der alten Letten* (1955), *Die Gottesgestalt der lettischen Volksreligion* (1961), and *Die himmlische Götterfamilie der alten Letten* (1972), are devoted to central problems of Baltic religion, with exhaustive bibliographies. MARIJA GIMBUTAS, *The Balts*, pp. 179–204 (1963), gives a concise summary. For essays on all conceptions, see the *Wörterbuch der Mythologie*, fasc. 7, pp. 373–454 (1965).

(H.Bi.)

Baltic Sea

As the largest expanse of brackish water in the world, the semi-landlocked and fairly shallow Baltic Sea is of conspicuous interest to scientists, while to historians it represents the economic core of the Hanseatic League, the great medieval trading group of north European ports. The many names for the sea—Ostsee (German); Östersjön (Swedish); Morze Bałtyckie (Polish); Baltiyskoye More (Russian); Itämeri (Finnish)—attest to its strategic position as a meeting place of many nation-states.

General description and location

The Baltic Sea is roughly finger-shaped, covering 160,000 square miles (420,000 square kilometres) and extending northeast on the eastern side of the Scandinavian peninsula from latitude 54° N to very near the Arctic Circle. The catchment area drained by the rivers bringing fresh water into the Baltic is about four times as large as the sea itself. Its major axis, from eastern Denmark to south Finland, is just over 1,000 miles long, with an average width of about 120 miles. The western Baltic is connected to the North Sea by the channel known as the Skagerrak, a deep inlet that separates southern Norway from the tip of the Jutland peninsula; to the immediate east of the Skagerrak, but, at a right angle to it, the shallower Kattegat separates northeast Denmark from Sweden. The large islands of Bornholm

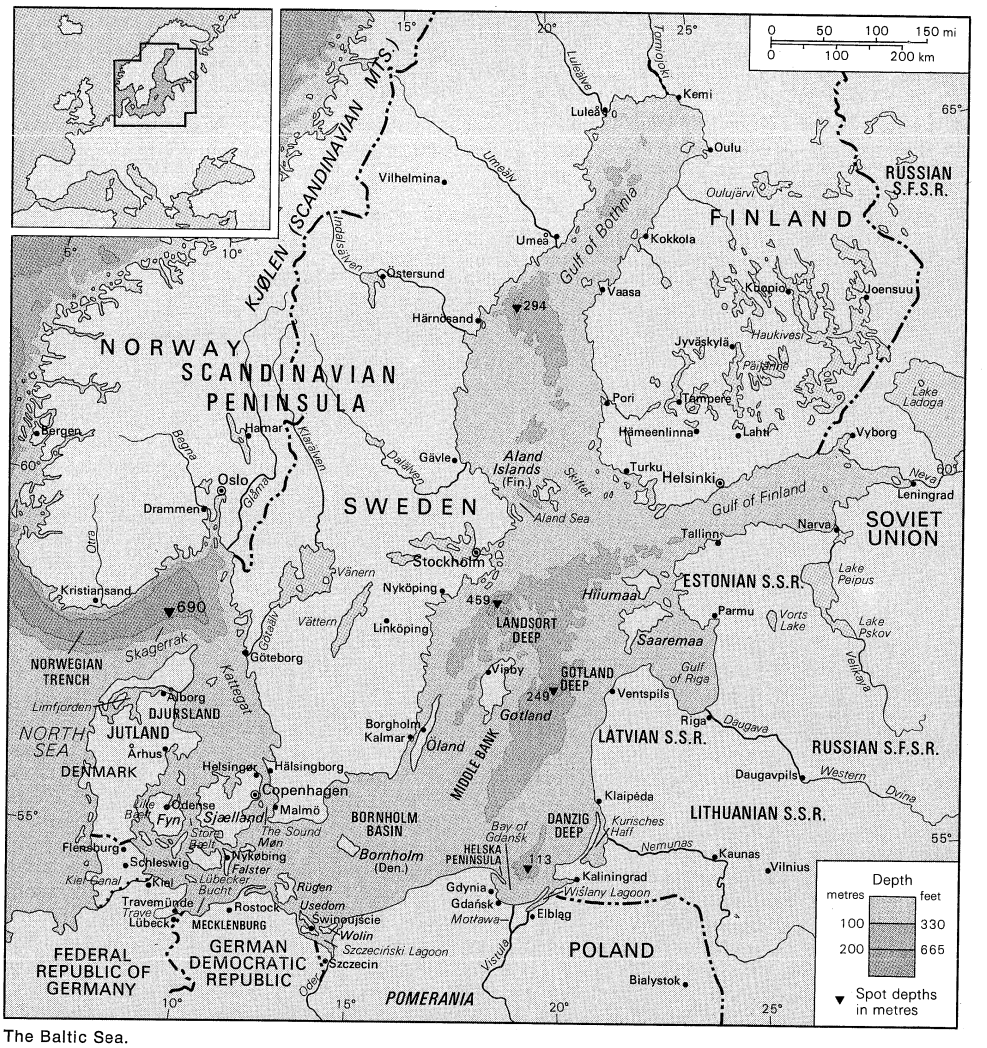
(Denmark) and then Öland and Gotland (Sweden) lie in the western Baltic while the Aland Islands (Finnish Ahvenanmaa; Swedish Åland), further north, rise from a narrows between Sweden and Finland and mark the entrance to the arm of the Baltic, known as the Gulf of Bothnia (Swedish Bottenhavet). Just to the south of the Aland Islands, the narrow Gulf of Finland stretches eastward between Finland and the Soviet Union, with Leningrad at its head. Proceeding clockwise from the west, the states bounding the Baltic are Norway; Sweden; Finland; four of the constituent republics of the Soviet Union (the Russian Soviet Federated Socialist Republic and the Estonian, Latvian, and Lithuanian S.S.Rs.); Poland; the German Democratic Republic; the Federal Republic of Germany; and Denmark. (For history, see BALTIC STATES, HISTORY OF THE, and SCANDINAVIA, HISTORY OF; for economic aspects, see HELSINKI; LENINGRAD; and STOCKHOLM; see also the articles on the states bordering the Baltic.)

The Baltic Basin. *Physical evolution.* The Baltic Sea is a shrunken remnant of the water-covered region that emerged as the melting Scandinavian Ice Sheet retreated toward the Arctic at the close of the Ice Age. In 12000 BC, ice covered all of northern Europe as far as the present German–Polish coastline; by 7700 BC glacial meltwater had formed the Yoldia Sea, which stretched east from the present Skagerrak across what is now lake-strewn southern Sweden as far as the present Lake Ladoga, beyond the bend of the Gulf of Finland. A thousand years later, only limited areas of stagnant ice remained in northern Sweden, leaving the freshwater Ancylus Lake stretching from Arctic Sweden and Finland to the present southern Baltic. Later changes around 4500 BC led to a breach of the land bridge between the present Baltic and North seas, forming The Sound (Øresund), the Store Bælt (Great Belt), and the Lille Bælt (Little Belt), which fragment the Jutland peninsula. Today there is a rise in sea level equivalent to several feet per century in the Stockholm region—a feature absent in southern Sweden and along the Danish coast—while, on the other hand, deposition is building up the head of the almost tideless Gulf of Bothnia.

Features of the sea floor. The shallowest part of the Baltic is the continental shelf, from which rise the islands of the Danish archipelago. Here the Lille Bælt divides east Jutland from the island of Fyn (Funen), which is itself separated from Sjælland (Zealand) by the deeper Store Bælt. The narrow channel of The Sound between Denmark and Sweden is too shallow for ocean-going vessels, so that Göteborg is the Swedish transatlantic shipping terminal on the coast of the Kattegat. The greatest deeps (over 1,500 feet) lie off the southeast coast of Sweden between Nyköping and the island of Gotland and also in the Gulf of Bothnia in the Aland Sea between Sweden and the Aland Islands. A deep-water channel also extends along most of the Gulf of Finland. The Baltic Sea proper contains a series of basins (*e.g.*, in the Zatoka Gdańska [Bay of Gdańsk]) divided by shallow shelves.

Hydrography. *Salinity and ice formation.* The Baltic Sea is so nearly land-locked (and its outlet so shallow) that its waters are remarkably fresh. Its longest rivers, the Vistula and the Oder, drain regions of temperate continental climate; swollen by spring snowmelt they have low evaporation rates, thus further reducing the salinity. The highest salinity is recorded in the west Baltic, where the figure is about 10‰ (parts per thousand) at the surface and about 15‰ near the bottom, while the

Factors influencing low salinity



The Baltic Sea.

lowest is at the head of the Gulf of Bothnia, where it is less than a third of this amount. The low salinity and the shallow coastal waters cause pack ice to accumulate at the head of the Gulf of Bothnia and off Finland in most winters, sometimes banked up in pressure ridges almost 50 feet high. Drift ice forms at and north of the Aland Islands area and also in the inner reaches of the Gulf of Finland, reaching a depth of about a yard. Navigation between Stockholm, Turku, and Helsinki is, however, possible, except in the most severe winters. Along the Polish and German coasts, ports are normally closed by ice for at least four weeks a year. In summer, the surface waters remain cool, while Arctic temperatures persist at lower depths.

Tides and currents. The Baltic, sheltered from the strong North Sea tides, is remarkable for the general absence of tidal currents, although the great influx of river water and surface runoff, including spring snowmelt, causes an outflow of surface water into the North Sea, while a deeper countercurrent of more dense salt water flows into the Baltic. Strong northeasterly winds may cause high waves along the southern shores, resulting in coastal flooding; conversely, prevalent southwesterly winds have aided the accumulation of sand dunes along the German and Polish coasts and also cause a piling up of water in the northern Baltic.

Coastal features. The coasts of Sweden and Finland are highly fretted and generally rocky, whereas those of the southern Baltic are flat and rather featureless. Where the crystalline rocks of the ancient rock mass known as the Fennoscandian Shield outcrop along the northern coasts, partly obscured by glacial drift and marine deposits, they are often fringed by the low, rocky, islands known as a skerry guard. These are most numerous in the Saltsjön between Stockholm and the open waters of the Gulf of Bothnia and, again, off the southwest coast of Finland. Off southeast Sweden, the narrow, elongated island of Öland and also Gotland formed of ancient limestone, partly covered by sandy drift deposits. Off southern Sweden, the rectangular island of Bornholm is formed of a detached fragment of granite, and its high cliffs were shaped by faulting and shearing of the rock strata.

Glaciation and sea-level changes

The coastal features of eastern Denmark are the outcome of Ice Age glaciation and of subsequent changes in sea level. The east coast of Jutland, north of the Djursland peninsula, is smooth and low-lying. To the south are shallow bays, or *viks*, divided by low promontories. In the area around Schleswig, shallow, straight-sided inlets known as *Förden* occur, and the Flensborg Fjord (Flensburger Förde) forms part of the boundary between Denmark and West Germany. The islands of the Danish archipelago have a broken coastline, with a number of shallow inlets and also bars, notably, the Odense bar on the island of Fyn. Where terminal moraines (deposits marking the farthest extent of glaciers) reach the sea, low promontories are formed. Solid rock seldom outcrops except along the coast of Møn Island where chalk rock forms a moderate cliff feature known as Møns Klint.

The short length of Baltic coastline within West Germany is one of shallow *Förden* and bays. Kiel lies at the head of one such inlet, south of the entrance to the Kiel (Baltic–North Sea) Canal, which runs through German territorial waters. To the east is the Lübecker Bucht (Bay of Lübeck) at the head of the Trave estuary, where Travemünde—a ferry port for Copenhagen and Sweden—combines the functions of a seaport and tourist resort. Beyond the frontier of East Germany, the coast of Mecklenburg is flat and low-lying. A series of long shingle bars, capped by moving sand dunes, has been built up here, cutting off the distinctive shallow lagoons from the open sea. Distinctive examples are the west–east spit of Darsser-Ort, on the island of Rügen, and the link (near Świnoujście) between the islands of Usedom and Wolin, which isolate the lagoon of Das Haff from the open sea. Beyond the Polish frontier, the port of Szczecin lies at the mouth of the Oder River. Solid rock outcrops conspicuously only on the island of Rügen, where the remarkably irregular coastline includes the chalk cliff known as the Königstuhl.

East of Szczecin, the coast of Pomerania (Pomorze) is generally flat and featureless, with sand dunes and spits bounding brackish lagoons. Beyond the Ikin spit of the Mierzeja Helska (Helska Peninsula) the Vistula Delta drains into the Baltic through a number of distributaries, and the historic city of Gdańsk lies on the most westerly of these, the Motława. To the east, spectacular *Haff* and *Nehrung* ("lagoon" and "shingle bar") features have developed: the region is well-known for the classic development of these features, the names of which have been applied to similar areas throughout the world. Sand dunes, covering an elongated shingle spit, almost enclose the famous brackish Frisches Haff, at the northeastern end of which lies Königsberg, the historic German outpost founded by the Order of the Teutonic Knights. Once the chief city and port of East Prussia, it is now the Russian port of Kaliningrad. Northward, the cliff coast of Uzlovoye is noted for its amber, a fossilized resin that formed a valued item of medieval trade throughout the Baltic and as far afield as Venice, which was reached by the "amber route" via Kraków and Vienna. At the northern end of the triangular inlet of the Kurisches Haff, at the mouth of the Neman, lies Klaipėda (German Memel), the most northeasterly city of Germanic origin in the Baltic. Cutting off the lagoon from the Baltic is the 60-mile-long shingle bar (the classic Nehrung), capped by low fixed dunes fringed by high moving dunes of white sand.

Haff and *Nehrung* features

The Soviet Baltic coast consists of rather monotonous features. Glacial deposits cover solid rock, and the coast is broken by broad bars, such as those on which the Latvian port of Riga lies. At the head of the Gulf of Finland is Peter the Great's "window on the Baltic," the city of St. Petersburg, renamed Leningrad in 1924, with its white buildings in classical style lining the waterfront of the Neva.

Economic resources. The Baltic Sea is no longer the major highway of trade that it was in the Middle Ages, when it flourished as the main means of communication between the ports (Lübeck, Rostock, Visby, and Gdańsk) of the Hanseatic League. The German Hansa merchants traded mainly in fish, notably, salted herring and stockfish (the dried cod from Norway and Iceland), and also in softwood timber for shipbuilding, hemp for ropes, flax for sailcloth, and grain. Forest products traded included honey and furs, notably, from Russia and Finland, as well as Stockholm tar, while amber was a semiprecious commodity. Overfishing of the herring, the opening up of trade to the New World following the Age of Discovery, and the increase in the size of sailing ships led to the decline of the Hanseatic League. Copenhagen, however, continued to prosper on the profits from tolls exacted from passing shipping during the period when both shores of The Sound were under Danish rule. Following the Reformation and the rise to political ascendancy of Sweden as the dominant Baltic power of the 17th century, Stockholm began to supersede older Swedish trading centres such as Kalmar and Visby. The later rise of Russia under Peter the Great and the empress Catherine, however, saw a corresponding decline in the economic importance of both Sweden and Denmark. Finland became a grand duchy under Russia by the Treaty of Hamina (1809), and its independence was only proclaimed in 1917. Poland ceased to exist as an independent state until resurrected in 1918, though Gdańsk (formerly German Danzig) thrived as a port. The present political and economic structure of the states bordering the Baltic in the south and east is largely a reflection of post-World War II changes, and traditional trading patterns have been amended accordingly. Timber and fish are the main items in modern Baltic trade. Softwood timber is Finland's "green gold," and it is also a major export from a large number of Swedish ports, as well as from the Soviet Union. Processed wood—including woodpulp, cellulose, paper, and hardboard from Sweden and Finland—is of growing importance. Sweden exports iron ore during the summer months from the Arctic

port of Luleå, as well as from the older port of Gävle in the south. Both Finland and Denmark have developed shipbuilding industries and marine engineering, especially since the end of World War II, Finland paying reparations to the Soviet Union in the form of ice breakers. Copenhagen, Helsingør, and Odense build motor ships. The Swedish port of Göteborg now manufactures cars and has a number of light engineering industries using high-grade steel, notably, in the manufacture of ball bearings.

Fisheries

The fisheries of the Baltic, once so rich, no longer compare in value with those of the North Sea, which are far more abundant in marine life. The shipment of fish on ice and as factory-packed frozen fish is nevertheless of growing significance to both Sweden and Denmark, the fisheries of the Kattegat, notably, those for plaice, cod, and herring, yielding the greatest returns. The traditional trade in smoked and salted herring remains important, especially with Germany. On the island of Bornholm, the traditional method of smoking the herring is over open wood fires. In Denmark, the island of Falster has offshore oyster beds, and in summer crayfish and prawns are caught off the coast of southern Sweden. In the warmer waters of the southern Baltic, herring, cod, and a variety of flatfish make up the catch off the Polish and German coasts. Gdynia (part of greater Gdańsk), Szczecin, and Elbląg are also commercially important, as are the German ports of Travemünde and Rostock.

BIBLIOGRAPHY. W.R. MEAD, *An Economic Geography of the Scandinavian States and Finland* (1958), provides economic coverage of the northern Baltic states. W.R. MEAD and HELMER SMEDS, *Winter in Finland* (1967), gives a detailed, illustrated account of the Finnish winter and its effect in the Baltic Sea. A.C. O'DELL, *The Scandinavian World* (1957), includes descriptions of the Baltic Sea and its surrounding lands. AXEL SOMME (ed.), *A Geography of Norden*, rev. ed. (1968), provides a full and illustrated account of the geography of the Scandinavian states and Finland with reference to the Baltic Sea. JOHANNES HUMLUM and KNUD NYGARD, *Danmark-Atlas* (1961), is a useful compilation of maps. See also the national atlases of Denmark, Finland, and Sweden.

(A.F.A.M.)

Baltic States, History of the

The Baltic states comprise the present-day Soviet republics of Lithuania, Latvia, and Estonia, on the eastern shores of the Baltic Sea. While, in some respects, they have a common history, they are ethnically and linguistically diverse. The Lithuanian and Latvian languages belong to the Baltic branch of the Indo-European linguistic family. The Estonian people, on the other hand, belong to the Finno-Ugric family of peoples, and their ancestors migrated to the Baltic probably as early as the 3rd millennium BC, coming from the wooded regions west of the Ural and from the middle Volga.

Diverse backgrounds of the Baltic peoples

FROM EARLIEST TIMES TO THE 18TH CENTURY

In prehistoric times these lands were inhabited by many different tribes. The Estonians and Livs lived in the northern and western areas bordering on the sea. The Latvians and Lithuanians, consisting of several related but independent tribes, inhabited the southern and inland areas. The western part of what is now Latvia, bordering the sea, was settled by seafaring people known variously as Kurs, Cours, or Couronians. During the 10th and 11th centuries these tribes were subject to a double pressure: Russian penetration from the east, and a Swedish push toward the shores of Courland and Estonia from the west.

The political-social structure of the various tribes showed certain differences. The Estonians and Cours of the coastal regions were peasant and seafaring peoples led by "elders." They were influenced by the Scandinavian vikings. The Latgallians, Semigallians, Selonians, Livs, and Lithuanians, who had connections with the Russo-Varangian principalities (Polotsk, Pskov), had a more developed, almost feudalistic social structure. The Lithuanians, who very early developed a warrior class, were far superior to their neighbours in military prowess. At the beginning of the 13th century all of these tribes were still entirely rural; there were markets, seaports, and castles, but cities in the European sense did not yet exist. Some trades, particularly that of the smith, were remarkably advanced, but building with stones and mortar was still unknown. There was no writing. The religion of these tribes was nature worship, with dimly anthropomorphic figures representing the sky, sun, moon, thunder, and other forces of nature.

The German conquest of Latvia and Estonia. The Latvians and Estonians were conquered and made Christians by the Germans. Traders began visiting the estuary of the Daugava River in the latter part of the 12th century. Coming in contact with the Livs first—they are today absorbed by the Latvians—the Germans named the country Livland, a name rendered in Latin as Livonia. Missionaries followed the traders, but their efforts had little success. Berthold, appointed bishop of Livonia, decided to use the sword and was killed in battle in 1198. His successor, Albert of Buxhoevden, was more successful. With the permission of Pope Innocent III he organized a crusade against the "treacherous Livs." In 1201 he founded the city of Riga, and in 1202 the Order of the Knights of the Sword was formed.

The Knights of the Sword

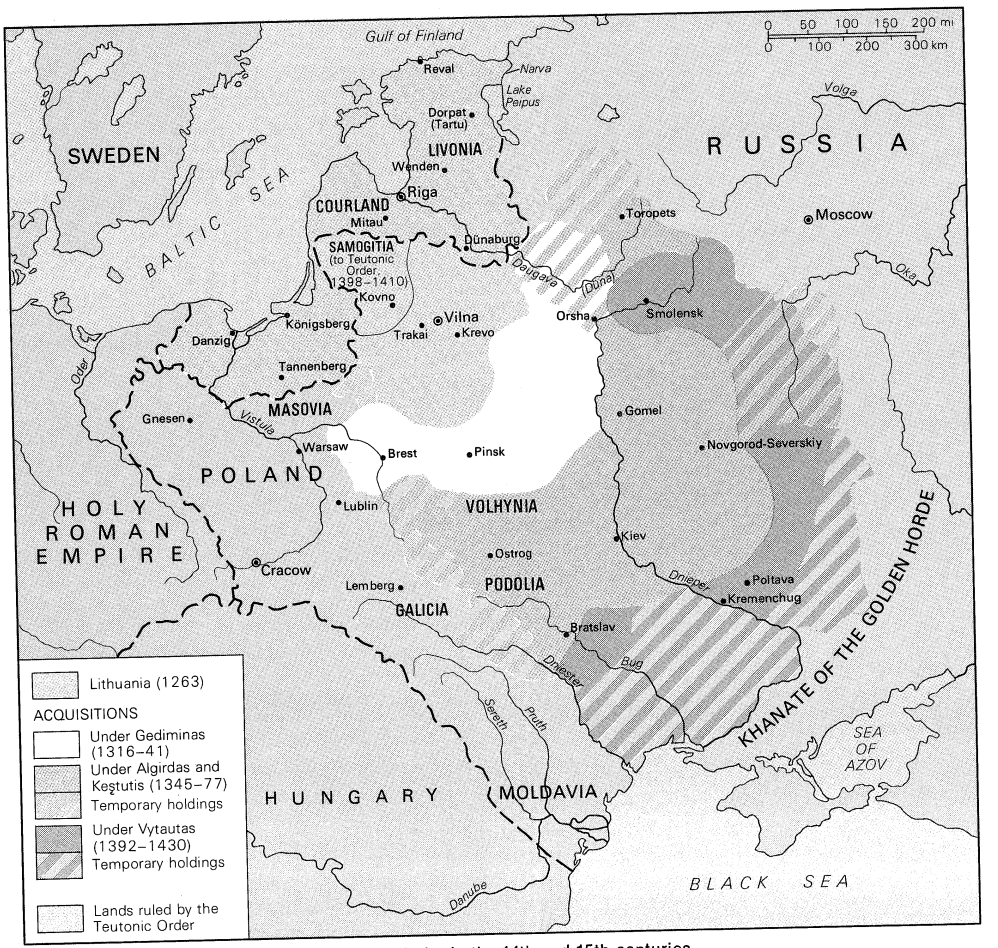
By 1208 the Knights were firmly established on both banks of the Daugava, and Albert proceeded northward to the conquest of Estonia. In a major battle in 1217, the crusaders defeated the Estonians; their commander Lembitu was killed. Albert then concluded an alliance with King Valdemar II of Denmark, who in 1219 landed with a strong army on the northern coast, on the site of present-day Tallinn. By 1227 the conquest was complete.

In 1237 the Knights of the Sword joined the Teutonic Order, which now assumed control of Livonia. Northern Estonia was under Danish rule; the greatest part of Livonia—that is, Southern Estonia and Latvia—was shared between the Teutonic Order and the bishops.

The conquered tribes, particularly the Estonians, were restive; a number of revolts occurred. The imposition of a new class of feudal overlords upon them reduced the Estonians, the Latvians, and the Livs to a state of serfdom. Their superficial conversion to Christianity was of little help; occasional attempts by the papal curia to intercede on behalf of their rights were of no effect. Personal freedom was greater in the cities, such as Riga, Dorpat (Tartu), and Reval (Tallinn), where men could support themselves in the trades. Much of the urban population consisted of German settlers, but the countryside remained native. The various indigenous nationalities retained their folklore and their language, while the German overlordship brought the Latvians and Estonians as a whole into the cultural sphere of the West. The borderline along the Narva River and Lake Peipus (Peipsi) became a cultural dividing line between the Roman Catholic West and the Russian Byzantine East.

Independent Lithuania. While the Latvians and Estonians were losing their independence in the 13th century, the Lithuanians succeeded in retaining theirs and in subsequently establishing a state that extended far beyond the ethnographic borders of the original nation. The Lithuanians in the early 13th century were ruled by a lower warrior nobility and the boyars (higher nobles). The first leader to unite the nation was Mindaugas, who in 1236 succeeded in bringing Lithuania proper, Samogitia, and the adjoining Belorussian territories under his single rule. In order to consolidate his regime he declared himself ready to accept Christianity and to relinquish his claim upon the contested province of Samogitia in the northwest in return for peace with the knights of the Livonian branch of the Teutonic Order. Mindaugas was crowned king by the authority of Pope Innocent IV in 1253. His peace with the Livonian knights did not last, however, because of continuing conflict in Samogitia. He and his two sons were murdered in 1263, and Lithuania returned to paganism.

The country was not unified again until 1290, when the grand duke Vytenis was recognized as absolute ruler. He was succeeded in 1316, by his younger brother Gediminas, who ruled until 1341. Gediminas extended Lithu-



Lithuania and the lands ruled by the Teutonic Order in the 14th and 15th centuries.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*; Georg Westermann Verlag, Braunschweig

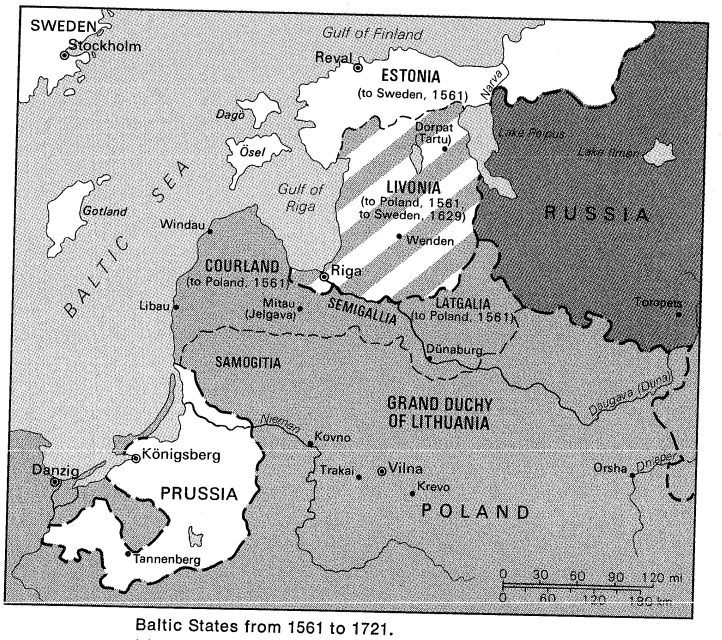
The Lithuanian Empire

ania's territories from the Baltic Sea southward almost to the Black Sea and eastward to the Dnieper. Lithuania became a major power. Under Gediminas' sons Algirdas and Kęstutis the Grand Duchy expanded to include Kiev in the east. The expansion was a product of political conditions in eastern Europe. The Tatar conquest of Kiev had destroyed the influence of the Kievan state over the other Russian principalities, which then tended to gravitate toward the west. This gave Lithuania an opportunity to expand to the east and southeast, while at the same time fighting off the Teutonic Order on its western frontiers. That a small non-Christian state was able to conquer and maintain control over such an extended area was partly the consequence of skillful diplomacy (including political marriages). The conquered Russian principalities were allowed to keep their autonomy and their Orthodox religion. The business of the state was conducted in a Slavic language (a kind of Belorussian Church Slavonic). Germans were brought into the east to build trade and commerce, and many Jews also settled there.

In 1385 Jogaila, the son of Algirdas, concluded an agreement to unite Lithuania and Kievan Russia with the Polish crown if he could marry the 12-year-old Queen Jadwiga of Poland and become king of Poland himself. He went to Cracow, was baptized on February 15, 1386, receiving the name Władysław II Jagiełło, married Jadwiga, and on March 4 was crowned king of Poland (see POLAND, HISTORY OF). The Lithuanians were baptized in 1387, and Władysław granted the Lithuanian boyars, or gentry, great privileges. Władysław's cousin Vytautas became grand duke of Lithuania. In 1410 the Polish-Lithuanian forces inflicted a crushing defeat on the Teutonic Order at Tannenberg (Grünwald). Samogitia returned to Lithuania. The grand duke Vytautas renewed the policy of eastward expansion under the sovereignty of his cousin King Władysław, and during this time Lithuania reached its largest expansion. When Vytautas, often called "the great," died at the age of 80 in 1430, the heroic epoch of medieval Lithuania ended.

With the acceptance of Roman Catholicism, Lithuania was drawn culturally toward the West. The Teutonic Order ceased to be a menace, but in the East the rise of Muscovy posed a threat to Lithuania's Belorussian conquests. The federal union between Lithuania and Poland was of no advantage to the Lithuanian and Belorussian peasantry. The culturally more advanced Polish nation tended to assimilate the Lithuanians, so that the legal equalization of the aristocracy of both countries resulted in the Lithuanian and Belorussian nobility becoming polonized; they began to consider themselves an integral part of the Polish nation. The Lithuanian and Belorussian peasantry now met with a fate similar to that of the Latvians and Estonians before them; isolated under the rule of an aristocracy that did not speak their language, they were restricted to farming and a rustic mode of life. Lithuanian and Belorussian became the language of peasants. In 1569 the personal union between the dynasties of Poland and Lithuania was changed to a union of the two countries, and Lithuania was reduced to a subsidiary land under the Polish crown. In the 18th century, when Poland was partitioned among Prussia, Austria, and Russia, Lithuania was annexed by Russia.

Livonia, Estonia, and Courland from the 16th to the 18th century. Medieval Livonia was a loose confederation of feudal ecclesiastical states, whose sovereigns, the bishops, were princes of the Holy Roman Empire. The responsibility for the defense of the confederation against the growing power of the Grand Duchy of Moscow was assumed predominantly by the Master of the Teutonic Order. While the power of the sovereigns decreased, the importance of the corporations of the landed nobility (*Ritterschaften*) and of the free cities increased. The towns enjoyed prosperity through the commercial



Baltic States from 1561 to 1721.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*; Georg Westermann Verlag, Braunschweig

activity of the Hanseatic League, predominantly in the trade with Russia (Novgorod). As early as 1522–24 the teachings of Martin Luther began to gain ground among the German ruling class of Livonia. This was to prove important in the history of the non-German population, for the evangelical ministers did much to foster written literature in the Estonian and Latvian languages. With the establishment of Lutheranism, the ecclesiastical states became anachronisms.

Partition of Livonia in the 16th century

When the Russian tsar Ivan IV the Terrible advanced claims on Livonia in order to secure access to the Baltic Sea, the Livonian confederation broke down before the violent onslaughts of the Russian and Tatar troops. The sovereigns, the nobles, and the magistrates of the cities were forced to apply for protection to the kings of Sweden and Poland-Lithuania. Livonia broke up into three "duchies" of Livonia. Estonia, and Courland, the borders of which were different from the original Estonian and Latvian settlements; they remained the administrative divisions until 1917. Estonia, with its capital, Reval, came under Swedish rule; Livonia, with its capital, Riga, became a part of Lithuania; while Courland became a hereditary duchy to be held as a Polish fief. The nobility and the magistrates of the free cities retained their privileges; German was recognized as the official language; and German law and German administration remained.

The division of Livonia between Poland-Lithuania and Sweden did not ensure a lasting peace. In 1592 the lands became the object of the first Swedish–Polish war and of the struggle between Protestantism and Catholicism. In 1629 Poland was forced to cede Livonia with Riga to Sweden, retaining only the southeastern province of Latgale.

The Swedish kings, particularly Gustavus II Adolphus (reigned 1611–32) and Charles XI (reigned 1660–97), were accustomed to a free peasant class in their own country. They sought to raise the Estonian and Latvian peasants from the serfdom into which they had been thrust. Under Charles XI the levies and payments of the peasants were regulated in accordance with a general land registration. The Bible was translated into Latvian and Estonian. There were numerous administrative, judicial, educational, and ecclesiastical reforms in favour of the peasantry, but they fell short of their intended accomplishments because of the frequent, devastating wars. The Swedish crown needed the support of the landed nobility and rewarded their services with estates and with concessions of autonomy. Despite this, the association with the Swedish Empire has traditionally been regarded by the Estonians and Latvians as one of the better periods of their history.

In Courland the last master of the Teutonic Order of Livonia had preserved its independence as a duchy under the formal suzerainty of Poland. His grandson Jacob, the ablest of all the dukes of Courland (sole ruler 1642–82), developed local industry, fostered foreign trade, and created not only a sizable merchant marine but also a navy of warships. He also acquired two colonies: the island of Tobago in the West Indies, and Gambia on the West African coast at the mouth of the Gambia River. His ambitions were cut short because the country was unavoidably drawn into the European power struggle. In the third partition of Poland in 1795, Courland became a province of Russia.

THE BALTIC COUNTRIES UNDER RUSSIA (UNTIL 1918)

With his victory over Sweden in the Great Northern War (1700–21), the Russian tsar Peter I the Great gained both Livonia and Estonia. He thus succeeded in his aim of "opening the window to Europe" for his country. The other Baltic lands passed into Russian hands at the first and third partitions of Poland in 1772 and 1795. For the Germans of the Baltic lands—particularly the nobility—their incorporation into Russia opened up great opportunities for advancement in the service of the tsar. For the peasants it brought a deterioration of legal status and increased exploitation.

Social changes under the tsars

Not until the 19th century did a process of social and national emancipation begin. Under the tsar Alexander I the Estonian and Latvian peasants were given their personal freedom (1816–19), but without the right to own land. By the middle of the century, however, they were allowed to acquire leased land as their personal property. The Estonian and Latvian provinces thus began to develop an agrarian structure quite different from that in Russia. However, as the big landed estates remained untouched, most of the peasants were not able to acquire enough land to be self-supporting. In consequence, many Estonians and Latvians migrated to the Russian interior where land was available for settlement. In Lithuania the peasantry was not liberated before 1861—when the emancipation of the Russian serfs took place—and there was no consolidation of individual holdings, before the beginning of the 20th century; the Lithuanian emigration of the late 19th century went westward to Canada and the United States.

Advances in education

Some progress was made in education, with the development of elementary schools teaching in the native tongues. By the end of the century there was almost no illiteracy in Estonia and Latvia. A German-language university was established in Dorpat (Tartu) in Estonia in 1802. By the middle of the 19th century it had become a focal point for national revival among the Estonians and Latvians. A Polish-language university founded at Vilna (Vilnius) in 1803 served in a similar way for the Lithuanians. Educated Latvians, Estonians, and Lithuanians began to grow conscious of their national origins, whereas previously they had been germanized or polonized. Starting with an interest in the past and with the study of the national languages, the movement inevitably developed a political tendency.

The Lithuanian national movement had to fight particularly hard for its aims. The Lithuanians took part in the Polish rebellions of 1830–31 and 1863 and suffered the same repression afterward. From 1864 to 1905 the policy of russification extended to every part of Lithuanian public life: it was forbidden to publish newspapers, periodicals, or books in Polish or Belorussian, while books in Lithuanian could be printed only if the Russian alphabet was used; Russian was the only language of teaching in the schools; and the Roman Catholic religion was persecuted. The Lithuanian resistance was able to capitalize on the fact that Lithuanian was also spoken in the eastern part of East Prussia, and the national movement flourished there under the leadership of Lutheran clergymen and teachers. It was there, on Germany territory, that the first Lithuanian daily newspaper was published, and Lithuanian books were printed to be smuggled into Russia.

Beginning in the 1880s the Lithuanian resistance to russification also received strong support from the Catholic clergy, which sided with the national middle class and the growing class of intellectuals. Both of these classes had developed out of the peasantry and sought to emancipate themselves from the lower nobility and the owners of the large estates, who were naturally interested in maintaining their association with the Poles and the Polish language.

In the Baltic provinces the Russian government introduced a series of liberal reforms during the 1860s and 1870s, but after the assassination of the tsar Alexander II in 1881 a general strategy of systematic russification began that lasted until 1905. The russification extended to the whole educational system, the courts, and governmental administration. At the same time, however, it did much to strengthen the Baltic nationalities, particularly the Latvians and Estonians. It released many of them from their provincial, patriarchal environment and gave them careers as civil servants in Russia; it also brought them into contact with the developing Russian revolutionary movement. The provinces became part of the giant Russian Empire. Railroad lines were built from the Baltic seaports to the Russian hinterland. Riga became a world port, its population growing to 290,000 in 1900 and to 530,000 in 1914. Riga, Tallinn, and Narva became important industrial centres. These developments changed the character of the city populations. The Baltic Germans, who had never comprised more than 10 percent of the population, declined in number and importance. Although the German influence remained strong in the sciences (through the University of Dorpat), as well as in the Lutheran Church, in the large landed estates, in wholesale trade, and in industry, banking, and the professions, the advancing Estonians and Latvians crowded the Germans out of the middle class trades, business, and the civil service. Many of the German academicians emigrated to Germany, where they were successful as scientists and writers, while the number of Estonian and Latvian academicians grew steadily. The percentage of Estonians in the population of Tallinn rose from 51.8 in 1867 to 88.7 in 1897, and the percentage of Latvians in the population of Riga rose in the same period from 23.5 to 41.6.

Marxism appeared in the Baltic provinces at the turn of the century. The Latvian Social Democratic Party, founded in 1904, joined forces with the Russian Social Democratic Workers' Party in 1906. An independent Estonian sister party was established in 1906. Both parties maintained connections with the German Social Democratic Party. A Lithuanian Social Democratic Party was founded in 1895, but in that agrarian, Catholic country it could not flourish.

The Russian Revolution of 1905 was felt strongly in the Latvian and Estonian provinces. Bourgeois politicians, together with radical revolutionaries, raised the demand for national autonomy. When revolutionary bands spread into the countryside, looting and burning the manor houses, the government in St. Petersburg sent troops to put down the uprising. A total of 908 Latvians and Estonians were court-martialed and shot, and a number of leading politicians fled abroad. As elsewhere in Russia, the Revolution was followed by concessions from the tsar in the way of liberal reforms, particularly in education and cultural and economic life. The Baltic regions were allowed to send elected representatives to the new imperial Duma (legislature).

LIBERATION AND INDEPENDENCE (1917–40)

Liberation (1917–20). At the end of World War I a chaos of events occurred in the Baltic states. Allowing for many differences, there was a common pattern. Following the Bolshevik Revolution of October 1917 (November in the present calendar), Soviet regimes were proclaimed in Latvia and Estonia; they were swept away with the advance of the German armies. After the defeat of Germany independent non-Communist (bourgeois) governments were set up with the support of the Allies. The government of Soviet Russia tried in vain to overthrow them and in 1920 signed peace treaties with independent Estonia, Lithuania, and Latvia, which survived as sovereign states until World War II.

The rise of bourgeois governments

Estonia. The Russian Revolution of March (February, old style) 1917 overthrew the tsar and brought a brief period of political autonomy for Estonia. On April 12 the Russian provisional government passed a decree providing that all districts with an Estonian majority were to be united into one province; Jaan Poska, the National Liberal mayor of Tallinn, was appointed commissioner for Estonia and an Estonian national council was planned. Elections to the council (*maapäev*) took place in June.

The October coup d'etat by the Bolsheviks in Petrograd was immediately felt in Estonia. The bourgeois majority parties of the *maapäev* decided to break away from the crumbling Russian Empire, but the Bolsheviks appointed a Communist government for Estonia. In February 1918 German forces advanced to Estonia. The Communists fled from Tallinn, and on February 24 the *maapäev* declared Estonia's independence and formed a provisional government. This collapsed the following day when German troops entered Tallinn. On March 3, 1918, the Brest-Litovsk Treaty was signed, effectively transferring sovereignty over the Baltic countries to Germany.

Germany capitulated on November 11, 1918. The Estonian provisional government once more proclaimed the independence of Estonia. But the Soviet government declared the Treaty of Brest-Litovsk null and void. On November 28 the Red army took Narva and started the invasion of Estonia, which had been denuded of all arms by the retreating Germans. The government of Konstantin Päts was successful in obtaining weapons and war materiel from the Allies. With the aid of a British naval squadron and a Finnish voluntary force of 2,700 men, the commander in chief Col. (later General) Johan Laidoner was able to open a counteroffensive in January 1919. By the end of February all of Estonian territory had been freed, and the Estonian Army penetrated into Soviet and Latvian territory.

Repulse of the Red army

Latvia. In Latvia the struggle for independence was even more difficult than in Estonia. The Latvian People's Council, representing peasant, bourgeois, and Socialist groups, proclaimed independence on November 18, 1918. A government was formed by the leader of the Farmers' Union, Kārlis Ulmanis. The Soviet government established a Communist government for Latvia at Valmiera, headed by Pēteris Stučka. The Red army, which included Latvian units, took Riga on January 3, 1919, and the Ulmanis government moved to Liepāja (Libau), where it was protected by a British naval squadron. But Liepāja was still occupied by German troops whom the Allies wished to defend East Prussia and Courland (Kurzeme) against the advancing Red army. Their commander, Gen. Rüdiger Graf von der Goltz, demanded the control over the Latvian units as well. He intended to build a German-controlled Latvia and to make it a German base of operation in the war against the Soviets. This intention caused a conflict with the government of independent Latvia supported by the Allies. Von der Goltz had at his disposal—besides his German troops—the *baltische Landeswehr*, a combat-ready unit of predominantly Baltic-German volunteers including also Latvian units. On May 22 these forces took Riga. Pushing northward, the Germans were stopped near Cēsis by the Estonian army which included 2,000 Latvians. The British general, Sir Hubert de la Poer Gough, head of the Allied military mission, negotiated an armistice. The Germans had to abandon Riga, to which the Ulmanis government returned in July. In the meantime, the Red army, finding itself attacked from the north by the Estonians, had withdrawn from Latvia.

In July Gough demanded that the German troops should retreat to East Prussia. But von der Goltz now raised a "West Russian" army, systematically reinforced by units of German volunteers. These forces, headed by an adventurer, Col. Bermondt-Avalov, were to fight the Red army, co-operating with the other "White Russian" armies of Kolchak, Denikin, and Yudenich, supported

by the Allies. But on October 8 Bermondt attacked the Latvian troops and occupied the suburbs of Riga south of the river. By November 10, however, the Latvians, helped by the artillery of an Anglo-French naval squadron, cooperating with Estonian forces, defeated von der Goltz's and Bermondt's troops, attacked finally also by the Lithuanians. Until December 1919 all German troops had abandoned Latvia and Lithuania. Only Latgale remained in Red hands; but this province was cleared by 33,000 Latvians under Gen. Jānis Balodis, 20,000 Poles under Gen. Edward Śmigły-Rydz, and 6,000 men of the *Landeswehr*, which had been put under the command of the British Lieut. Col. H.R.L.G. Alexander (later Earl Alexander of Tunis).

Lithuania. During World War I the Germans occupied a great part of historic Lithuania. An independent state of Lithuania was recognized by Germany after Brest-Litovsk in March 1918, with the proviso of a "perpetual alliance" of Lithuania with Germany. On November 11, 1918, a government was formed by Augustinas Voldemaras. As the German armies withdrew, the Red army occupied Vilnius on January 5, 1919, and installed a Communist government. The Germans remained in western Lithuania until December 1919. The Lithuanian Army took the offensive against the Reds in February 1919, and by the end of August the country had been cleared of Soviet troops.

A dispute with Poland had developed, however, over the possession of the capital city of Vilnius and the district of Vilna. The city was largely Polish in population, while the district of Vilna was predominantly Lithuanian. Józef Piłsudski the head of the restored Polish state, proposed two alternatives: an independent state without Vilnius and its region or a larger state including Vilnius but federally linked to Poland. On April 20, 1919, the Polish Army led by Piłsudski took Vilnius from the Red army, which enabled the Lithuanians to re-enter Kaunas. Even though the Lithuanians were able to regain the region in July of 1920, during the war between the Soviet Union and Poland, the territorial issue remained in dispute for some time and was revived by Poland in 1938.

Consolidation. The Estonians signed the very favourable Treaty of Tartu with the Soviet Union on February 2, 1920, in which the Soviets recognized "without reservations" the independence of Estonia and "relinquished for all time" all rights of sovereignty over Estonian territory. Peace treaties between the Soviets and Lithuania and Latvia were signed later in the same year, on July 12 and August 11, respectively. The province of Latgale returned to Latvia, and the Soviet Union recognized the claim of Lithuania to the region of Vilna.

Political and social reforms

The internal problems faced by all three republics were largely the same: to reorganize their semifeudal agrarian structures, to adapt their economies to the new conditions, and to establish constitutions. In Estonia and Latvia the governments had promised the distribution of land parcels to the combatants during the war. Now both republics solved their agrarian problem with the expropriation of all the holdings of large landed estates (10,600,000 acres), thus destroying the economic and political power of the Baltic-German nobility whose corporations were dissolved. Tens of thousands of the rural proletariat were given land. The expropriated forest lands remained the property of the state and became an important source of income from lumber exports.

In Lithuania the large estates were mainly in the hands of Poles. The land reforms, less radical than in Latvia and Estonia, left the owners with maximum holdings of 370 acres. The government sponsored cooperatives to handle the collection and marketing of farm produce.

Constitutional reform. The constitutional reorganization in all three countries was radically parliamentarian in character, the legislative body clearly predominating over the executive branch. In Estonia, for example, there was a single-chamber parliament with a system of proportional representation, and the prime minister was also the chief of state.

Estonia and Latvia had a three-party structure. There were the conservative Farmers' Party, the National-Liberal Centre Party, and the Social-Democratic Party. The Farmers' Party was of particular importance since the two countries were predominantly agricultural, and it participated in almost all of the government coalitions. In Estonia the National Centre Party had its power base in the university city of Tartu. The Communist Party was outlawed in Estonia after an attempted coup in Tallinn on December 1, 1924. The trials and executions that took place before and after the coup did not improve Estonia's already delicate relationship with the Soviet Union.

Lithuania also had a three-party system. The conservative Christian Democratic Party was the strongest. The second strongest party was the left-liberal People's Socialists (Liaudininkai), and the third was the Social Democratic Party.

The economy. In seceding from the Russian Empire, the Baltic states had lost their economic hinterland. The result was unemployment, particularly in the metal-processing and textile industries. Estonia developed an entirely new industry with the opening up of the rich oil shale fields in the northeast. The timber and related industries increased slowly in importance, as did the export of meat, dairy, and poultry products. Great Britain became the principal market for all three countries, with Germany a close second. Trade with the Soviet Union remained slight. By 1930 employment in manufacturing accounted for 17.4 percent of the labour force in Estonia and 13.5 percent in Latvia, while Lithuania, with only 6 percent, remained almost entirely agricultural.

Education and culture. Freed from outside restriction, cultural life expanded. Schools of all kinds increased. The russified University of Dorpat (Yurev) in Estonia became the State University of Tartu; in Latvia the Riga Polytechnic Institute was enlarged to become a complete university; in Lithuania, since Vilnius was in Polish hands, the government established the Vytautas University in Kaunas in 1922. Literature, music, and the fine arts reached the level of the rest of Europe. Cultural policy was strongly Western in its orientation; English was the first foreign language taught in the schools. The problem of national minorities was settled best in Estonia by means of the law on cultural self government in 1925.

Political tendencies in the 1920s and 1930s. For many Estonians, Latvians, and Lithuanians the independent, democratic, national state had long appeared to be the goal of their historical development. When it was finally attained, however, some serious problems appeared. Political experience and democratic traditions were lacking, as well as institutions that would have protected the interests of the state against those of particular groups. The radical parliamentary constitutions drawn up in the first hours of independence hampered the creation of stable governments. By the end of 1926 an authoritarian presidential regime had been established in Lithuania, similar to that of Piłsudski in Poland. This happened when the small Nationalist Party, with the consent of the Christian Democratic Party, overthrew a left-democratic regime in December 1926.

Problems of establishing democratic governments

The numerous political parties in Estonia and Latvia prevented the formation of stable coalitions and led to frequent governmental crises during the 1920s. The lifespan of the governments of Estonia during the years 1919–1933 averaged eight months and 20 days. The political problem became even more pronounced in 1930, when the world economic crisis brought financial difficulties and unemployment that emphasized the need for stable government. Voices demanding constitutional reform were heard in both countries. In Estonia the movement was led by the "Vaps" (Vabadussõjalaste Liit, or League of Freedom Fighters), which had grown from a group of war veterans into an anti-Communist and anti-parliamentary mass movement. The proposal of the Vaps won a majority of 72.7 percent in a referendum of October 1933. The acting president Konstantin Päts was expected to prepare for the election of a new president. Instead, he declared a state of emergency on March 12, 1934; the Vaps was dissolved, its leaders were arrested, and the parliament was soon also dissolved. After that Päts ruled by decree until 1938.

In Latvia a similar development occurred on May 15th, 1939. After attempts at constitutional reform had failed and the country had become increasingly polarized between the far right and the far left, the prime minister Karlis Ulmanis declared a state of emergency. He formed a government of national unity from representatives of almost all the important parties. From then on he governed without the parliament.

In neither Estonia nor Latvia was there any significant resistance to the suppression of parliamentary government. Quite a few party representatives were even ready to support the authoritarian regimes. These regimes drew their main support from the well-to-do and the peasants, from the army and the home guard. Both heads of state based their coups d'etat on the need to prevent the interference of foreign powers in state affairs. Both were also successful in diminishing the power of the radical right. Both strove to reorganize the society by setting up representative bodies of the professions.

But there were marked differences in their styles of leadership. The moderate Estonian president regarded his authoritarian regime as a "regency for the restoration of the endangered democracy" and worked for a conservative reform of the state. He had his regime legalized by a plebiscite in 1936 in order to elect a constituent assembly to draft a new constitution. The candidates were chosen from the ranks of the Patriotic League that he had founded in February 1935. The new parliament that convened in 1938 had in its lower chamber 63 members of the Patriotic League and a token opposition of 17.

In Latvia the energetic "leader of the people" K. Ulmanis did not bother to legalize his regime by popular referendum or even to organize a unified following. In 1936 he combined the office of prime minister with the office of president of state and adopted the nationalistic theme of "a strong and Latvian Latvia." He also enlarged the state-run sector of the economy, predominantly at the expense of the German minority. In consequence, the younger generation of Germans felt themselves to be losing out to the Latvian majority, and German National Socialism found an increasing number of supporters among them. In both Latvia and Estonia the liberal intellectuals chafed under the restraints put upon speech and press. The rural population and the business interests, on the other hand, favoured the authoritarian regime because it brought prosperity; foreign trade showed a steady increase.

Rise of the one-party state in Lithuania

In Lithuania a nationalistic one-party state emerged. The dictatorial tendencies of prime minister Voldemaras aroused opposition among conservative-ecclesiastical circles, which led to his removal by the president, Antanas Smetona, in 1929. Smetona now cast himself as a "people's leader" with the Nationalist Party in full control of the state, supported by some of the Christian Democrats. His regime also had the support of the army, the home guard, and the state-sponsored youth organization, Young Lithuania. The obvious model for the regime was Fascist Italy.

A continuing issue in Lithuanian politics was the city of Memel (Klaipėda) and the surrounding region on the right bank of the Memel River (Nemunas), which had been taken from Germany after World War I and given to Lithuania. The Lithuanian government had been required to grant autonomy to Memel. The nationalist leaders in Lithuania aspired to Lithuanianize the territory, whose inhabitants were mostly of Lithuanian descent, although the sentiment in Memel was for reunion with Germany. Following a German ultimatum, Lithuania was forced to restore Memel to Germany on March 23, 1939. (After World War II, Memel was absorbed by the Lithuanian Soviet Socialist Republic.)

The end of independence. The Baltic states had won their independence at a time when both Russia and Germany were defeated in war. They retained it as long as the two powers remained weak. Endeavouring to remain neutral, the Baltic republics refrained from aligning themselves with political blocs in the 1930s. The likelihood of reaching an understanding with the Germany of Hitler or the Soviet Union of Stalin seemed so remote that the governments of the Baltic states never considered it in their plans. Proposals for ties with Finland and Poland ran aground on the unreconcilable differences between Lithuania and Poland and on the refusal of Finland to engage in affairs south of the Gulf of Finland. An Estonian–Latvian defense alliance was signed in 1923 and renewed in 1934, which also provided for cooperation in foreign policy. When Lithuania joined the alliance in 1934 it became known as the Baltic Entente. All three of the states signed nonagression pacts with the Soviet Union. They differed, however, as to whether the greater danger lay in the West or in the East. While influential circles in Latvia inclined toward the idea that Germany was the graver danger, the Estonians were clearly preparing to meet a possible attack from the Soviet Union. In the summer of 1939 the Baltic question was one of the issues in the ill-fated Anglo-French negotiations with Moscow.

The geographic plight of the Baltic states

In a secret protocol to the German–Soviet pact of August 23, 1939, Estonia and Latvia were recognized as belonging to the Soviet sphere of interest. In September, after the German victory over Poland, Lithuania was put in the same category. Moscow then demanded that the Baltic states should sign mutual assistance pacts with the U.S.S.R. and allow the construction of Soviet military bases on their territory. Completely isolated, the governments of the Baltic states realized that military resistance was useless. The agreement between Berlin and Moscow for a resettlement of Baltic Germans in Germany made it abundantly clear to the Balts that Hitler had left their states at the mercy of Stalin. The alacrity with which the ethnic Germans of Estonia, Latvia, and Lithuania departed for the *Reich* showed that they had little doubt as to the fate impending for those states. On October 10 the Soviet Union returned the Vilnius region to Lithuania.

The Soviets were at first satisfied to observe the limits of their bases, watching with growing distrust the economic infiltration of Germany into the Baltic states as well as attempts to activate the Baltic Entente. When the Germans took Paris in June 1940, Stalin demanded that the governments of the Baltic states admit more Soviet troops. After the Soviet Army had established itself in the countries, pro-Soviet governments were set up. These held rigged elections on July 14–15 in which only single lists of Soviet-sponsored candidates were allowed to stand. The new parliaments immediately voted for the incorporation of their countries into the U.S.S.R.; the proposals were officially accepted by the Supreme Soviet on August 3, 5, and 6.

Arrests and deportations

Many political leaders were arrested or fled to the West. In the first year of Soviet occupation mass deportations took place in June 1941 and many people were arrested and executed. Estonia lost more than 60,000; Latvia about 35,000; and Lithuania about 45,000. Those deported included politicians, military personnel, higher officials, writers and publishers, and businessmen, together with their families.

After the German attack on the U.S.S.R. in June 1941, the Baltic states became part of a larger Ostland in which Belorussia was also included. Many Estonians, Latvians, and Lithuanians were by that time ready and willing to cooperate with the Germans. But they found that it gained them neither their national independence nor the return of their nationalized property. Whereas the Soviets had sought to extirpate the "class enemy," the Nazis tried to wipe out the Jews; they killed about 190,000 Jews in Lithuania, about 90,000 in Latvia, and about 4,500 in Estonia.

In the fall of 1944, as the Germans retreated and the Soviets returned, large numbers of people fled before the advancing Soviet Army. About 30,000 Estonians escaped by sea to Sweden and 33,000 to Germany. About 115,000 Latvians fled to Germany and Sweden, and about 70,000 Lithuanians to Germany. After the Soviets had restored Communist governments in the three countries there were new deportations. Estimates of the numbers deported in the years 1941–49 run to about 500,000, including large numbers of peasants who were deported for

resisting the collectivization of their farms. About 25–30 percent of these persons are said to have returned to their native countries after Stalin's death (1953).

THE BALTIC STATES AS UNION REPUBLICS OF THE U.S.S.R.

Soviet policies

After the victory over Germany the Soviet authorities resumed their previous efforts to integrate the Baltic states into the U.S.S.R. Most of the Communist leaders of the early days were replaced by officials who had grown up in the U.S.S.R. or been trained there.

Collectivization and industrialization. The rural population was forced into the *kolkhozy*, or collective farms, without regard for the consequences to agriculture. Resistance by partisans or guerrillas, which persisted longest in Lithuania, was ultimately broken by special forces of the security police. Collectivization eliminated the independent farming class, which had been the political basis of the Baltic states.

The *kolkhozy* as a rule included about 150 families and had an average area of from 5,000 to 6,000 acres. They were run, like all Soviet collective farms, on the principle of common cultivation; members were paid out of the proceeds of the farm according to the number of work units they had performed. Every household was allotted a plot of land for its own use. There were also state farms, or *sovkhozy*, in which workers were paid wages. The Baltic regions concentrated on dairy farming and cattle breeding; in the 1950s and 1960s Estonia and Latvia held first place among the union republics of the U.S.S.R. in milk production per cow.

The economies of the Baltic republics were integrated into the Soviet system of economic planning and development. This resulted in considerable growth in production, as a result of Soviet investment in the Baltic region. Some outstanding projects of the 1950s and 1960s included the development of the Estonian oil shale industry, which supplies gas for Leningrad and Tallinn. The oil shale also serves as fuel for the giant power stations, to generate electricity, as do the dams built on the Narva and Daugava rivers. Living standards remained relatively low. One reason was the preference given to industrial goods over consumption goods in economic planning; another was that the Baltic republics, which were far more productive than the Soviet average, were expected to share their surplus with other republics. The standard of living nevertheless remained considerably higher than in other parts of the Soviet Union.

Demographic changes. These industrial and agricultural policies worked a fundamental change in the social structure of the Baltic republics. From predominantly rural societies they became predominantly urban. In 1939, 65 percent of the Latvians lived in rural areas, as did 66 percent of the Estonians; but by 1966 the ratio was reversed, and more than 60 percent of the Latvians and Estonians were urban dwellers. The change in Lithuania was not pronounced: from 61 percent rural in 1959 to 55 percent in 1966. The larger cities grew as follows: Riga from 347,800 in 1939 to 667,800 in 1966; Tallinn from 143,384 in 1939 to 327,800 in 1965; and Vilnius from 236,100 in 1959 to 304,700 in 1966.

Russian immigration

Another demographic change was the immigration of Russians. In 1970 more than half the population of Riga was estimated to be Russian; Tallinn was 40 percent Russian; and in Liepaja, an important military base, the Russians comprised about 65 percent of the population as early as 1960. Many Russians were brought in to work in the Estonian oil shale industry and in the Narva region. The proportion of Latvians in the total population of Latvia declined from 75 percent in 1935 to 62 percent in 1959, and to 56.8 percent in 1970; Estonians in Estonia decreased from 88 percent in 1934 to 75 percent in 1959, and to 68.2 percent in 1970; and Lithuanians in Lithuania from 84 percent to 79 percent between 1923 and 1966. But according to the results of the census of 1970 the proportion increased anew to 83 percent.

An important part of the Baltic peoples now lives in countries outside the Soviet Union, most of them in Sweden, the United States, Canada, and Australia. Their total numbers in 1970 were estimated at 102,800 Estonians, 180,000 Latvians, and almost 800,000 Lithuanians. The population of Estonia in 1959 was 1,197,000; in 1970, 1,356,000; the population of Latvia in the same years, 2,093,000 and 2,363,000; the population of Lithuania rose from 2,711,000 to 3,128,000.

Religion, education, and culture. Under Soviet rule, the activities of the formerly influential Lutheran and Roman Catholic churches have been severely limited. They are forbidden to print and disseminate religious literature, and the training of new clergymen is discouraged. Church attendance has declined markedly.

Education and culture in the Baltic republics have been "national in form, socialist in content." The native languages and literature, theatre and music, popular customs and national histories have all been promoted, but as part of a multinational Soviet culture and in terms of Soviet ideology. The severance of ties between the Baltic states and Russia after 1918 is regarded as the work of Western imperialism, and the forced integration into the Soviet Union as a "liberation from the yoke of imperialism."

In 1971 the incorporation of the Baltic states into the Soviet Union had still not been recognized by the United States and many other countries, although some governments have accorded it de facto recognition.

BIBLIOGRAPHY

General works: G. VON RAUCH, *Geschichte der baltischen Staaten* (1970), for the period of independence 1918–40, with bibliography; R. WITTRAM, *Baltische Geschichte* (1954), for the period 1180–1918 in Latvia and Estonia, with emphasis on the role of the German ruling class (with a complete bibliography); W. KIRCHNER, *The Rise of the Baltic Question* (1954), on the collapse of the Livonian confederation and the Baltic policy of Tsar Ivan IV; S.W. PAGE, *The Formation of the Baltic States* (1959); M.W. GRAHAM, *The Diplomatic Recognition of the Border States* (1939); A.N. TARULIS, *American-Baltic Relations, 1918–1922* (1965), and *Soviet Policy Toward the Baltic States, 1918–1940* (1959); J. VON HEHN, *Die Entstehung der Staaten Lettland und Estland, der Bolschewismus und die Grossmächte* (1956); J. VON HEHN *et al.*, *Von den baltischen Provinzen zu den baltischen Staaten, Beiträge zur Entstehungsgeschichte der Republiken Estland und Lettland 1917–1918* (1971); B. MEISSNER, *Die Sowjetunion, die baltischen Staaten und das Völkerrecht* (1956); EBBA CEGINSKAS, "Die baltische Frage in den Grossmächteverhandlungen 1939," in *Commentationes Balticae*, 12–13:3–73 (1967), a critical review of various treatises on the Baltic question from the viewpoint of the Western democracies.

The separate states: (*Estonia*): E. UUSTALU, *The History of the Estonian People* (1952); E. LAAMAN, *Eesti Iseseisvuse Sünd* (1936, reprinted 1964), a substantial, well-written account of the emergence of an independent Estonia, from the national-democratic viewpoint, including many illustrations; A. KAELAS, *Das sowjetisch besetzte Estland* (1958), from the viewpoint of Estonians in exile; *Eesti Entsüklopeedia*, 8 vol. (1932–37); E.F. VAREP and V.Y. TARMISTO, *Estoniya* (1967). (*Latvia*): E. ANDERSONS (ed.), *Latvia: Past and Present, 1918–1968* (1968); A. SPEKKE, *History of Latvia* (1957); E. DUNSDORFS *et al.*, *Latvijas Vesture*, 4 vol. (1958–67), an authoritative work, written by specialists; J. VON HEHN, *Lettland zwischen Demokratie und Diktatur* (1957); *Latvju Enciklopedija* (1950–55); V.R. PURIN and A.A. BRED, *Latviya* (1968). (*Lithuania*): C.R. JURGELA, *History of the Lithuanian Nation* (1948), with illustrations; A. SAPOKA (ed.), *Lietuvos istorija* (1936); T.V. PASHUTO, *Obrazovaniye Litovskovo gosudarstva* (1959), a scholarly work with a complete bibliography; M. HELLMANN, *Grundzüge der Geschichte Litauens* (1966); J.J. STUKAS, *Awakening Lithuania* (1966); A.E. SENN, *The Emergence of Modern Lithuania* (1959), with complete bibliography, and *The Great Powers: Lithuania and the Vilna Question 1920–1928* (1966); V. STANLEY VARDYS (ed.), *Lithuania Under the Soviets: Portrait of a Nation, 1940–65* (1965), with a description of the German occupation, 1941–44, and the national resistance against the returning Soviets; M.J. ROSTOVTSEV (ed.), *Litva* (1967); *Lietuviu Enciklopedija*, vol. 1–34 (since 1951); Albert Gerutis, *Lithuania: 700 Years* (1971).

For current information, see *Acta Baltica*, a periodical containing articles on political, economic, social, demographical, and cultural development in the Baltic countries under Soviet rule; *Baltic Review;* and MID-EUROPEAN LAW PROJECT, *Legal Sources and Bibliography of the Baltic States* (*Estonia, Latvia, Lithuania*) (1963).

(A.v.T./Ed.)

Baluchistan

Baluchistan (Balūchestān) is the westernmost province of Pakistan. It is bordered on the west and northwest by Iran and Afghanistan; on the north and east by the Pakistani provinces of Northwest Frontier, Punjab, and Sind; and on the south by the Arabian Sea. With an area of 134,050 square miles (347,188 square kilometres), it is the largest Pakistani province, but it is the most sparsely populated, with an estimated 2,409,000 inhabitants. The provincial capital of Quetta is located in the north.

The name Baluchistan is derived from the words Balūch and *estān* and literally means "place, or abode, of the Baluch people," who inhabited most of the contemporary provincial area by the end of the 15th century. Baluchistan as the name of the area was in use in the 15th century, but it was not the official designation of the region until the 19th century.

The least developed of the Pakistani provinces, Baluchistan is primarily a pastoral region. Its rich mineral wealth awaits exploitation, and the expansion of agriculture depends upon increased irrigation projects. Industrial activity is minimal, and development is hampered by the lack of adequate transport and communications facilities.

History. Archaeological remains have confirmed that prehistoric Baluchistan passed through the Stone and Bronze ages. The materials of the Quetta, Togho, Kulli, and Nal sites represent a fairly widespread level of cultural achievement and indicate that Baluchistan served as an intermediary link between the cultures of South Asia and the Middle East. Little is known of the region's early history, but it may have been subject to the Assyrians and the Medes. It was subjugated and annexed as part of the 14th satrapy of the Persian Empire under Darius I (522–486 BC); but, with the defeat of Darius III by Alexander the Great in 330 BC, the area came under the Greek supremacy.

The chronology of the succeeding powers suggests that, after Alexander's death, the territories of Baluchistan became a part or were under the political influence of the empires of Seleucus I Nicator and Candragupta Maurya (305 BC), the Indo-Greeks and the Parthians (3rd–2nd century BC), the Scythians (100 BC–AD 200), and the Sāsānids (3rd–7th century AD). The Hephtalite Turks controlled central and northeastern Baluchistan from AD 470 to 520, leaving the southern coastal area to the Sāsānids. As Sāsānid power weakened during the 7th century, the Brahmin rulers of Sind extended their influence into western Baluchistan.

Arab invasions of the 7th century

Baluchistan emerges into recorded history with the advance of the Arab armies in the 7th century. With the final conquest by Muḥammad ibn al-Qāsim in 711, most of Baluchistan became part of the Sind province of the Umayyad and the ʿAbbāsid empires. From the 11th century, the region fell under the control of various powers and formed part of the Mughal Empire from about 1595 to about 1638.

The first Baluch people to arrive in Baluchistan were mainly of the Brahui group. They set up the principality of Kalāt in the central portion of the region in the 14th century. In the 15th century, Kalāt was overrun by the last great migrating body of the Rind–Lāshār Baluch. Naming the new ruler of Kalāt, the Rind and the Lāshār moved onward; the former founded the principality of Sibi with its capital at Fatehpur and the latter the principality of Kachhi with its capital at Gandāvá. As a result of the 30-year Rind–Lāshār War (*c.* 1490–1520), most of the Rind and Lāshār migrated further to the Punjab, Sind, and Gujarāt.

The Brahui regained control of their principality, and the khanate of Kalāt, which became the future nucleus of Baluch power, was founded in 1666. Nasīr Khān (1750–93) welded together the region's different ethnic stocks, organized the military and socio-political institutions of the Baluch, aligned himself with Nāder Shāh of Iran and Aḥmad Shāh Durrānī of Afghanistan, and succeeded in creating a political unit independent of neighbouring Sind, Iran, and Afghanistan.

The British influence in Baluchistan commenced with the mission of the British administrator Sir Robert Sandeman to Kalāt in 1875 and the subsequent occupation of Quetta in 1877. By the Treaty of Gandamak with Afghanistan in 1879 and other treaties with the *khāns* of Kalāt, the territories acquired were constituted into British Baluchistan Province and Tribal Areas by 1896, while Kalāt became a protected princely state. Baluchistan Province became part of Pakistan in 1947, and Kalāt state acceded to Pakistan one year later. The various parts of the former British Baluchistan Province (settled districts, native states, and tribal areas) became a more integrated single administrative unit after 1947. In 1955, Baluchistan was merged into the newly created "Province of West Pakistan" which was abolished on the 1st of July 1970 when Baluchistan was reestablished as a separate province in its present form.

The natural environment. *Relief, drainage, and soils.* Baluchistan can be divided into four physical regions. The upper highlands of the central and northeastern regions are traditionally known as the Khorāsān country. The area is bounded by the Sulaimān Range to the east and the Toba Kākar Range to the northwest. To the north of the central Sulaimān Range is an oblong massif known as Kaisargarh (Kasi-barh) that reaches an altitude of 11,290 feet (3,441 metres). A series of hills curve westward from the Sulaimān Range to the broad mountain arc that abuts the Quetta-Pishīn uplands. Peaks in this region include Khalifat, which rises to 11,434 feet (3,485 metres) south of Ziārat, and Zarghūn, which attains 11,738 feet (3,578 metres) northeast of Quetta.

Four physical regions

The great break in the mountains formed by the Bolān and Khojak passes separates the upper highlands from the lower highlands. The lower highlands include the eastern slopes of the Sulaimān Range; the lower ranges of Makrān, Khārān, and Chāgai on the west; and the Pab and Kīrthar ranges on the southeast.

The third region consists of the extensive flat plains along the coast that extend northward into the mountains. In the east, the plain stretches along the eastern slopes of the Kīrthar Range as far north as Sibi. In the northwest are the arid deserts of Chāgai, Khārān, and Makrān and the swamps of Lora, near the Afghanistan border; and Māshkel, near the Iran frontier.

The upper highlands drain into the Indus River, while the lower highlands drain northward into the swamps or southward to the Arabian Sea. The main rivers are the Zhot in the northeast, Nāri in the southeast, Dasht in the extreme southwest, and Hingol, Porāli Nai, and Hab in the western Las Bela district.

The variety of soils is related to the region's physical features. Saline soils are found along the coast, around the swamps and lakes, and in some of the low-lying arid areas. Sandy soils occur in the three main desert areas and in parts of the plains. The valleys of the lower and upper highlands contain soft rocky soils, while stony soils dominate the vast waterless flats between the hills. There are deep deposits of alluvial soils in the plains of Sibi, Kachhi, and Las Bela.

Climate, vegetation, and animal life. There are three climatic zones. The climate is temperate along the coast, with a mean annual temperature of 86° F (30° C). The inland deserts and arid zones are hot, with a mean annual temperature of 98° F (37° C), while the submountainous region is cold, with a mean of 76° F (24° C). Outside the influence of the monsoon, most of the province is dry and experiences the extremes of heat and cold. The coastal plains become extremely hot during the long summer from April to September, with temperatures at Sibi rising to 122° F (50° C). The plateaus of Kalāt and Quetta-Pishīn receive some snowfall and experience severe cold during winter from October to March, the temperature falling to several degrees below freezing. Rainfall is irregular and scanty; it ranges from an average of four to 11 inches on the coast to between six and 16 inches in the hilly region.

The upper highlands of Kalāt, Quetta-Pishīn, and Ziārat support such trees as the juniper, pistachio, olive, ash, edible pine, poplar, and willow. The flats remain

Vegetation and animal life

covered with saxual and sagebrush, while the mountain valleys have a variety of fruit trees. The hillsides abound in herbaceous and bulbous plants, and the Ziārat and Kalāt uplands support such plants of economic value as cumin, hyssop (a European mint), and licorice. The perennial herb that is the source of asafetida occurs in Koh-i-Sultān, and wild rhubarb is found in the Khwāja Amrān range. Dwarf palms grow in the hilly tracts, and tamarisk, acacia, and wild caper (a low prickly shrub) occur on the plains. The valley of the Dasht River abounds in date-palm trees of the best variety, and the grasses in the Khārān region yield seeds that for centuries have been used as food grain during periods of famine.

The straight-horned markhor (a wild goat) and mountain sheep are found in the higher hills and the Sind ibex (a wild goat) in the lower highlands. Leopards and black bears are occasionally found in the western hilly regions, while the wild ass, the Persian gazelle, and the wolf occur in the deserts. The jackal, fox, and hyena are more common. Baluchistan has been a natural habitat for various breeds of sheep, cattle, and horses.

Typical game birds are the *chikōr* (a colourful bird of the partridge family) and *sisi* (of the same family, but smaller and dusty brownish in colour). The black and white *shakūk* (a larger bird with a longer tail) struts around in summer. There are waterfowl, bustards, and other migratories, while the raven, lammergeier (a large bird of prey), and golden eagle are among the permanent bird population. The deserts and plains abound in a vast variety of reptiles and insects. The horned snake, *kingarmār*, is typical of the Khārān region. The tortoise is common, and the skink (a lizard) occurs in the sandhills of Chāgai. Crocodiles are found in the Hingol River, and the coastal belt is rich in fish and mollusks (such as snails, limpets, mussels, and cuttlefish). A large variety of marine fish include shark and skate, as well as pomfret, sole, and sardine.

The population. The Baluch and Pathan peoples constitute the two major and more distinct ethnic groups; the mixed ethnic stock, mainly of Sindhian origin, forms the third major group. There are minor ethnic communities of Meds along the coastal belt, Lurs in the predominantly Baluchi areas, and Jāṭs in the Sibi, Kachhi, and Las Bela districts.

The Pathan and Baluch peoples

The Pathans, divided into 13 subgroups, are believed to be of Turko-Iranian origin. Pathan tradition claims northern Baluchistan immediately west of the Sulaimān Range to be their ancient home, and the Pathans are still concentrated in the contemporary Zhob, Loralai, and Quetta-Pishīn districts.

The Baluch are concentrated in the Kalāt Division and in the Sibi and eastern Loralai districts of the Quetta Division. Their tradition, irrespective of the languages spoken by the various subgroups, traces their origin back to the ancient Babylonian homeland of Aleppo, Syria. Baluchi tradition and scholarship support the theory that the Baluch are Chaldean descendants of Belus, identified as the Babylonian king Nimrod.

Baluchistan is a multilinguistic region; a majority of its inhabitants are bilingual, and there is a sizable trilingual population. Baluchi, Brahui, and Sindhi are the principal languages of the Kalāt Division, while Pashto, Baluchi, Sindhi, and Seraiki are the main languages of the Quetta Division. Baluchi has eastern and western dialects, and Pashto is spoken in its southern dialect as distinct from the northern Pashto of Northwest Frontier Province. Sindhi has four dialects, and Seraiki has two dialects. Persian, in the Dehwari dialect, is spoken by the Dehwar communities of Kalāt and Mastung, and a mixed dialect, the Mokâki, is spoken by the Lurs in the Kalāt Division. Punjabi is spoken by the settlers from Punjab, mainly in the urban areas, while Urdu is used as a lingua franca.

The Baluchis and Pathans are mostly Sunnī Muslims, although some communities profess the Shīʿite doctrine. Some of the Baluchi communities and peoples of Makrān and Las Bela districts profess to be the followers of the 16th-century messiah Mīrān Muḥammad Mahdī of Jaunpur and constitute a sect by themselves. Hindus, Parsis (Zoroastrians), and Christians constitute minority communities, the latter two living mainly in Quetta city.

The vast rural areas are mostly barren, and semi-arid lands are visited by nomads in the rainy season. The few green areas have pastoral concentrations, while the coastline has clusters of fishing villages. Though some mining settlements have sprung up in the interior, the bulk of the rural population remains nomadic.

Settlement patterns

Nomadism and ruralism traditionally have been the dominant features of the Baluchistan population. Since 1947, however, the process of urbanization has increased in response to industrialization and the greater educational and employment opportunities in urban areas. The major towns are Fort Sandeman, Hindubāgh, Chaman, Mastung, Mach, Sibi, Kalāt, Khuzdār, Panjgūr, Turbat, Bela, Sonmiāni, and Gwādar. Since 1951 there has been a marked increase in the urban population; in seven of the province's ten districts, the increase has been more than 75 percent. Immigration from other Pakistani provinces has also been responsible for this increase, particularly in Quetta city.

Administration and social conditions. The province is headed by a governor, who is appointed by the central government. Each of its two divisions of Quetta and Kalāt are under a commissioner. The divisions are subdivided into ten districts, which are headed by deputy commissioners, political agents (Zhob and Chāgai districts), or deputy commissioner-cum-political agents (Sibi and Loralai districts), depending upon the degree of political control exercised within their jurisdiction. Quetta-Pishīn, Sibi, Loralai, Zhob, and Chāgai districts constitute the Quetta Division, while the Kalāt Division includes Kalāt, Kachhi, Las Bela, Makrān, and Khārān districts. The districts are subdivided into tahsils headed by *tahsildārs*, and the tahsil is further subdivided for purposes of revenue into *ḥalqah* and *maḥāl*, the smallest administrative unit, headed by a revenue officer.

Health facilities, though inadequate, have been extended to the whole province. Each of the ten districts has hospitals, dispensaries, and mobile dispensaries or rural health centres. There are also two tuberculosis clinics. Medical personnel include doctors, specialists, health visitors, and nurses. The major health problems are infant mortality, smallpox, eye diseases and blindness, and stomach troubles and kidney stones due to the non-availability of clean water.

Educational facilities

Traditional education includes an indigenous community-supported Islāmic system of *maktabs* (primary schools) and *madrasahs* (secondary and higher institutions) situated in towns and larger settlements. Classes are taught in Persian and Arabic, and the religious studies are the main educational objective. Western education was introduced in 1882, when the first government school was established by the British in Quetta. In the early 1970s, there were more than 1,460 primary schools, 150 middle schools, 60 high schools, seven colleges, and five vocational, commercial, or polytechnic institutes. Enrollment is much higher in the primary and secondary schools than in the post-secondary schools, and there are more male than female students. The University of Baluchistan in Quetta was established in 1970. Government policy aims at providing more education funds to extend and improve primary schools, to expand the scope of science education, to introduce "marketable skills" in the school curriculum, and to orient the university program to technical education.

Social-welfare work commenced with the sanction of 13 urban–rural community development projects, six of which were functioning in 1970. A provincial Social Welfare Council administers the program through voluntary social-welfare agencies. Among the projects run by the agencies are homes for industrial workers, orphanages, schools for the deaf and dumb, clinics and health centres, social services, and vocational training centres.

Rural development through the Rural Works Program is geared to the overall economic goal of increasing production. It includes the expansion of employment opportunities and the creation of new sources of income.

As a measure against drought, a food-storage scheme has been devised to construct grain-storage facilities in central areas. During the early 1970s funds were provided for the construction of such facilities at Quetta, Chaman, Nushki, Mastung, Mach, and Turbat.

The economy. *Resources and their exploitation.* Baluchistan's major economic potential lies in its mineral wealth, marine fisheries, livestock, fruit farming, irrigated agriculture, and scope for industrial growth. Development costs of water and power resources and communication facilities, however, remain prohibitive.

Mineral wealth

The province has proved reserves of coal, chromite, sulfur, marble, limestone, gypsum, emery stone, magnesite (native magnesium carbonate), natural gas, barites (the barium sulfate mineral), fluorite, onyx, and manganese. There are deposits of copper, bauxite, lead, antimony, iron ore, laterite, and brucite. Several of these minerals, such as coal, chromite, marble, sulfur, limestone, gypsum, barites, emery, and magnesite, are mined for export. The natural gas found in 1952 at Sui has contributed to industrial growth by the production of almost half the nation's electric power.

The coastal belt has large fishing potential. Fishing cooperatives are supplied with motor launches, and facilities are being extended at the ports of Sonmiāni, Ormāra, Kalmat, Pasni, and Jīwani. There is a cold-storage plant and an ice factory at Gwādar, where a new fishing harbour is to be constructed by 1975.

The sheep industry engages more than 90 percent of the population and utilizes more than 95 percent of the provincial acreage. The sheep raised produce high-quality wool, most of which is exported. Cattle and other animals yield hides and skins for the local leather industry and export.

Agriculture remains limited for want of water, power, and adequate communications. Of the total cultivable area of 17,000,000 acres, less than 3,000,000 acres are cultivated, mainly in areas of scanty rainfall. With the extension of irrigation from the Indus River to Sibi and Kachhi districts, acreage and production have increased. Major crops include wheat, rice, barley, millet, fruit, and vegetables.

The province contains less than 1 percent of the nation's industry. In the early 1970s, large-scale industry consisted of one pharmaceutical plant, one cotton-spinning mill, one woollen mill, one distillery, and a charcoal plant. Small industry comprises fruit and vegetable processing, grain milling, light-engineering workshops, and hand-embroidery units. Two large industrial estates in the Quetta area are to be established by 1975 at Usta Muhammad in Sibi District and Uthal in Las Bela District.

Transport. Baluchistan has almost 1,000 miles of railway and about 7,000 miles of roads, of which 950 miles are metalled. Modest fair-weather roads connect the important towns. The main highways are the Quetta–Mirjāveh road leading to Iran, the Quetta–Chaman road to Afghanistan, and the Quetta–Sibi road to Karāchi and Lahore. A shorter road from Quetta to Karāchi was under construction in the early 1970s as part of an international highway connecting Pakistan, Iran, and Turkey.

The railway system

The railway system radiates from Quetta to Zāhedān, Iran; Chaman on the Afghanistan border; Sibi through the Bolān Pass to Karāchi and Lahore; Fort Sandeman in the north; and Loralai in the northeast. Pakistan International Airlines provides daily domestic services between Quetta, Karāchi, Lahore, and Rāwalpindi. Small ships, launches, and boats ply along the coast, linking the ports of Sonmiāni, Jīwani, Pasni, and Gwādar with Karāchi and the Persian Gulf.

The expansion of transport facilities is basic to future economic development. There are proposals to link the mining area centre of Sepzand with Quetta and to link the town of Uthal—a proposed industrial site and headquarters of Las Bela district—with Karāchi by rail.

Cultural life. The Baluch and Pathan peoples are organized on the basis of *qawm* or *qabīlah*, "community" or "tribe." Depending upon its strength and standing, a community headed by a *sardār* or *khān* may be composed of various divisions and groups, each with its own headman. The nucleus of a Pathan community is characterized by blood bonds and homogeneity, while that of a Baluch community is characterized by "common weal and woe" and heterogeneity. In the Marri community of the Baluch, land is the common property of all and is divided among the three main clans every tenth year. *Riwāj* ("custom"), specific to each community, has served the function of unwritten law. The Balūchī Dīwān serves as a legislative assembly for the Baluch, while the *jirgah* serves as a judicial tribunal among both the Baluch and Pathan communities. *Mayār* or *nang* is the traditional Baluch code of honour, based on blood feud and chivalry. Marri society includes the *rāhzan*, who conducts the attack in war, and the *rhezwār shāʿir*, or community poet laureate, who ennobles their deeds and defends their actions.

Horse racing, wrestling, religious holidays, *melās* (fairs), and marriage feasts are the main festivals.

Baluch communities have professional minstrels who sing from *Balūchī daptar*, known as the great epic of the 30-year Rind–Lāshār War. Stories of valour and romance such as "Shay Murīd and Ḥānī" and "Bivargh and Gran Naz" originated from this epic. Other well-known romantic stories indigenous to Baluchistan include *Sammi-Dodā*, *Shīrīn-o*, *Dūstīn*, *Shadad-Mahnāz*, and *Lallah-Gran Naz*.

Zahīrūk songs of the western Baluchi area, *Lailee-Moar* of the central Brahui-speaking area, and *dehis* and *dastānaghs* of the eastern Baluchi area are the typical regional folk songs. Others are the *balūchī shiʿar*, ballads pertaining to the themes of war and valour; *līkū*, the traditional caravan song of the cameleer; *hallū-hālū*, *laylaṛṛū*, *lādū*, *lāṛṛū*, *saʾot*, and other marriage songs; cradlesongs and lullabies such as *nāzīnk* for girls and *lolī* for boys; and *mūdik* or *muhthik*, elegies and dirges over the dead. Among the Pathans, the *landai* form of folk song in the *kākarrī* style is the most popular singing style. The Kākar community has a rich variety of folk songs; *babū lāla* and *de shīn khāl-o naare* ("song of the blue mole") are their typical marriage songs, while *de atanṛ naare* are common songs for women.

The Baluchi Academy, the Brahui Adabī Dīwān, and the Pashto Academy—all centred in Quetta—are engaged in the investigation and preservation of the Baluchi, Brahui, and Pashto cultural and literary traditions of Baluchistan.

BIBLIOGRAPHY. ARRIAN, *Anabasis*, ed. by A.G. ROOS (1907) and *Indica*, ed. by RUDOLF HERCHER (1885), contain an account of Alexander's conquests in India; the *Futūḥ al-Buldān*, by AHMAD IBN YAHYA, AL-BALADHURI, is available in translation but the original Arabic text edited by M.J. DE GOEJE (1866), gives an early and most authentic account of the Arab conquest. For discussions of the origin and history of the Baloch people, see M.K.B. MARRI BALOCH, *The Balochis Through Centuries: History Versus Legend* (1964); M.S. KHAN, *History of Baluch Race and Baluchistan* (1958) and *The Great Baluch* (1965). An account of the Brahui and their language is contained in SIR DENYS DE SAUMAREZ BRAY, *The Brahui Language*, pt. 1, *Introduction and Grammar* (1909), pt. 2, *The Brāhūī Problem*, and pt. 3, *Etymological Vocabulary* (1934). The *Census of India 1901*, vol. 5 and 5-A, *Baluchistan*, contains statistical data and an informative introduction on all aspects of the country. The *Population Census of Pakistan, 1961*, vol. 1 (1962), contains the latest comparative tables for the country, and the *Baluchistan District Census Reports*, present the latest descriptive and statistical information about each district of Baluchistan. For economic material see the various reports of the government of Baluchistan.

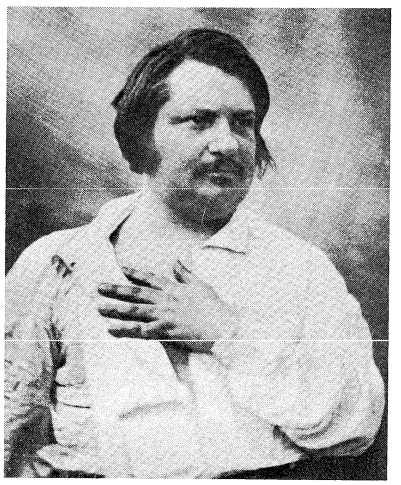
(N.A.B.)

Balzac, Honoré de

Recognized as one of France's greatest writers and as a genius in the novel, Honoré de Balzac drew inspiration from models as diverse as the French writers François Rabelais, Molière, Pierre-Augustin Caron de Beaumarchais, Alain-René Lesage, Denis Diderot, and Jean-Jacques Rousseau; the English writers Tobias Smollett, Henry Fielding, Samuel Richardson, and Laurence Sterne; and the U.S. novelist James Fenimore Cooper;

but, above all, the Scottish Romantic novelist Sir Walter Scott. He converted what had been styled "romance" into a genuine and convincing record of human experience. His vast life work, *La Comédie humaine* ("The Human Comedy"), written from 1829 to 1847, is exceptional in the history of fiction, so much the more so because he claimed at once to be a philosopher—even a seer—who explained man to himself, a historian (a "secretary" of society), a sociologist, and a psychologist.

J.E. Bulloz



Balzac, daguerreotype, 1848.

Balzac was of southern peasant stock, though in fact he was born in Tours on May 20, 1799. His father had made a career in the civil service, mainly based in Paris, but he was in Tours from 1798 to 1814. He had no right to the aristocratic particle *de*, which first he, then his son, assumed. Honoré's mother was of bourgeois stock, her family being cloth makers. Of neurotic temperament, she never understood her son, but it was probably thanks to her that he became attached to the various kinds of pseudo-science and occultism—mesmerism, magnetism, somnambulism, physiognomy, phrenology, and Illuminism (a belief in a special, supernaturally derived wisdom) —that permeated his thought. His sister Laure (later de Surville) was his only childhood friend, and she became his first biographer.

He spent nearly six years at the Collège des Oratoriens at Vendôme. This experience provided material for the novel *Louis Lambert* (1832–35). At Napoleon's downfall his family moved to Paris, where Honoré went to school for two more years and then spent three more as a lawyer's clerk, thus acquiring experience of the law and its victims, to be drawn upon in *Le Colonel Chabert* and *Un Début dans la vie* (*A Start in Life*). He aimed, however, at a literary career: as a writer of tragedy (*Cromwell*, 1819) he had no success. He turned to the novel: the results were sentimental, mystic, reminiscent of Rousseau. In addition, he linked up with hack writers and turned out potboilers—Gothic, humorous, historical novels—written under composite pseudonyms. He was learning his trade, but scholars and critics have found promise of his later achievements in what he himself called "this literary hogwash." He turned his attention to social skits—"physiologies" and "codes," parodies of scientific and legalistic compilations. Then he tried a business career as publisher, printer, and typefounder. By 1828 he was on the verge of bankruptcy and pulled out—the beginning of a lifetime of debt. He returned to writing, and his literary apprenticeship was over.

Early career, 1829–33. Two works of 1829 brought Balzac to the brink of success. *Les Chouans*, the first novel published under his own name, is a historical novel about Breton peasants called Chouans and their part in royalist guerrilla warfare in western France in 1799; in it he showed that women can be more full-blooded than Sir Walter Scott's characters Rebecca and Rowena. The other, *La Physiologie du mariage* (*The Physiology of Marriage*), was anonymous and, on the surface, humorous and satirical: the subject was cuckoldry, encompassing both its causes and cure. It betrayed a fundamental sympathy for and understanding of women that he was immediately to unfold in fiction and thereby began to establish his reputation.

Balzac's parents had retired to Versailles, while he himself spent most of his time in Paris: the man from Tours was already a Parisian and settled in a flat in the rue Cassini. A boisterous, somewhat vulgar person, avid for fame, fortune, and love, but above all conscious of genius, he decided to conquer not only the world of letters but also fashionable and artistic society. His knowledge of women was no longer secondhand. The almost brotherly affection he had for Zulma Carraud, a friend of his sister, lasted his whole life: an affectionate and admiring counsellor, critic, and hostess, she often welcomed him at her home, first at Angoulême (where he gathered material for one of his greatest novels, *Illusions perdues* [*Lost Illusions*]), then at Frapesle, near Issoudun. In 1822 he became enamoured of Madame Laure de Berny, a lady whom he assiduously wooed and won, who gave him his first experience of passionate love—and also a tender affection that, because she was old enough to be his mother (there was almost always a mother-oriented tendency in Balzac's loves), became more and more maternal until her death, in 1836. Between 1828 and 1831 Laure Junot, Duchesse d'Abrantès, widow of Andoche Junot, a Napoleonic general, partly occupied his attention.

Determination to conquer fashionable and artistic society

Other loves were to follow. Madame de Berny—"la Dilecta" ("The Beloved")—was, however, the most important. No doubt it was she who helped him come to an understanding of the mature woman, which inspired *La Femme abandonnée* (*The Deserted Woman*), *La Femme de trente ans* (*A Woman of Thirty*), and *Le Lys dans la vallée* (*The Lily of the Valley*). In addition, he helped Laure d'Abrantès with her memoirs and perhaps drew much information from her about the Napoleonic period, which he used in such works as *Une Ténébreuse Affaire* (*A Murky Business*).

Indeed, between 1828 and 1834 Balzac led a characteristically tumultuous existence, spending his earnings in advance as a dandy and man-about-town, exciting some general hilarity by his sartorial extravagances, his two-wheeler and groom, his gaudy walking stick, and other adornments. As a fascinating raconteur, he was fairly well received in society. But social ostentation then and later was, above all, a relaxation from phenomenal bouts of work—14 to 16 hours at his table in his white, quasi-monastic dressing gown, with his goose-quill pen and his exorbitant drinking of coffee. He was keen to make money but was already upbraiding his age for worshipping it ("money, the only god we now believe in"). He was exigent with editors and publishers. Everything, however, was genuinely subservient to a tremendous creative urge within him and a desire to put his century to rights —to be a Napoleon in the literary sphere. It is in his activity during these years that the key to his purpose can be found.

The journalist. The reigns of Louis XVIII and Charles X witnessed an upsurge of polemical and satirical journalism. Cheap dailies, such as *La Presse* and *Le Siècle* (to which Balzac was to contribute many serial novels), did not begin to appear until 1836, while the great organs, such as *Le Journal des Débats* and *Le Constitutionnel* were not so much involved as smaller and generally scurrilous sheets bent on attacking the increasingly reactionary ministries. Balzac wrote frequently for some of these between 1829 and 1831 and even helped to found one, *La Caricature*. They were mostly liberal, but by this time Balzac was no longer liberal-minded. He moved over rapidly to the absolutist view and wrote in 1832 for the royalist *Le Rénovateur*. Furthermore, he began to restrict his journalistic writing to such reputable periodicals as *La Revue de Paris*. Even with these, however, he quarrelled repeatedly, just as he did with his publishers. He thus began to nourish a rancour against the press that was later to burst into flame in, for exam-

ple, *Un Grand Homme de province à Paris* (*A Distinguished Provincial in Paris*), Part II of *Illusions perdues*. Balzac's scathing attacks on the periodical press are truly memorable.

Growing contempt for parliamentarian government

The politician. His collaboration with the *Rénovateur* came about chiefly through his friendship with a royalist leader, the Duc de Fitz-James; his growing contempt for parliamentary government was no doubt stimulated by his infatuation for the duke's niece, Henriette-Marie, marquise (later duchesse) de Castries. He was on very affectionate terms with her until she refused to become his mistress in the summer of 1832—a humiliation for which he avenged himself in *La Duchesse de Langeais*, which forms the second part of a trilogy called *L'Histoire des Treize* (*The Thirteen*). Even without this rebuff, however, his withdrawal from the unintelligent aristocratic party would have been inevitable. He gave up a whim he had entertained of standing for the Chamber of Deputies; *Le Médecin de campagne* (*The Country Doctor*), expresses his by then quite independent views in favour of authoritarian government, to which he henceforth consistently adhered.

The social historian. Balzac worked at high pressure in his rue Cassini flat until in 1835 the pursuit of creditors drove him to a house in Chaillot, a suburb of Paris, where he adopted a series of ingenious schemes for evading writ servers. He looked upon the novel as a variety of drama, and between 1829 and 1830 he produced his first six *Scènes de la vie privée* ("Scenes from Private Life"). They are classified as *nouvelles*, long short stories of predominantly psychological import, a genre that, thanks to his peculiar method of composing, expanding on and emending proofs, he was apt to develop into full-blown novels. These first *scènes* were mainly about girls in conflict with parental authority. As, for instance, in *La Maison du chat-qui-pelote* (*At the Sign of the Cat and Racket*), the minute attention he gave to domestic background already had earned him a reputation as an emulator of the Dutch painters and had also given promise of his later Parisian studies. Further items in this series (1831–32) reveal his compulsive interest in women of maturity.

New productions of this period of special importance include *Le Curé de Tours* (*The Vicar of Tours*) and *Eugénie Grandet*. He paid many visits to the country, particularly to the château de Saché near Tours, the home of Jean de Margonne, a friend of the family (where he often retired to write), and also to the house of his friends the Carrauds. The two novels show him working toward a second kind of *scène*—that of provincial life. At the same time, *Le Colonel Chabert* also brought him nearer to Parisian life; *e.g.*, *Ferragus* (Part I of *L'Histoire des Treize*) shows him very interested in Parisian topography and types of citizens. *La Fille aux yeux d'or* (*The Girl with the Golden Eyes*—Part III of *L'Histoire des Treize*), besides being a study of lesbianism, contains, in addition, an evocation of Paris as a kind of Dantesque inferno.

The philosopher. Simultaneously Balzac was staking his claim as a thinker, in a series of *Romans et contes philosophiques* ("Philosophical Novels and Tales"). Many of these are fantastic and fascinating, but the philosophic novels of this period are all-important. *La Peau de chagrin* (*The Wild Ass's Skin*) denounces the unscrupulous acquisitiveness and anarchical individualism of his age and tells the story of a magic talisman, possession of which confers upon its owner the fatal power of having all his wishes granted. In *Louis Lambert*, Balzac, making his most intensive effort to justify an enduring belief in the oneness of mind and matter, attempts to reconcile positive science with occultism. *Séraphîta* tries to conciliate the "two worlds" theosophy of Emanuel Swedenborg with Catholicism, by then Balzac's official creed. Of these three novels, *La Peau de chagrin* relates most closely to his future work. The talisman symbolizes an idea, basic to his psychology, of the destructive power of thought (a term embracing not only cerebration but also emotion, passion, imagination, and exercise of the will)—destructive both of the individual and of society: "the blade wears out the scabbard," as Lord Byron had succinctly observed.

With this idea is linked the notion of a vital "ethereal" fluid, a store of energy—concentrated inside the man—that he may husband or squander as he will, thereby lengthening or shortening his vital span. A supremely important feature in Balzac's characters—and in himself—is that most are spendthrifts of this vital force, a fact that explains the monomaniacs who abound in his work from the very beginning: Gobseck, in the *nouvelle* of that title, a usurer gloating over his sense of power; a miser obsessed with gold in *Eugénie Grandet;* a fanatical chemist in *La Recherche de l'absolu* (*The Quest of the Absolute*); an idolatrous father, the King Lear of *Le Père Goriot;* and many others until his life's end.

"La Comédie humaine." The year 1834 marks a climax in Balzac's career, for by then he was totally conscious of his great plan to group his work so that it should form one whole. There were to be three general categories: *Études analytiques* ("Analytic Studies"), dealing with the principles governing human life and society; *Études philosophiques* ("Philosophical Studies"), revealing the causes determining human action; *Études de moeurs* ("Studies of Manners"), showing the effects of those causes, and now to be divided into six kinds of *scènes*—private, provincial, Parisian, political, military, and country life. The *Études analytiques* were unfortunately to remain undeveloped (*La Physiologie du mariage*, *Petites misères de la vie conjugale* [*Conjugal Life*], some fragmentary and projected works); even in the *Études de moeurs*, political and military life were ill catered for.

This entire project resulted in a total of 12 volumes (1834–37), the first volume of *Études philosophiques* (December 1834) being preceded by an explicative preface, of primary importance, which was written by a friend, Félix Davin. By 1837 Balzac had, naturally, written much more, and by 1840 he had hit upon a Dantesque title for the whole: *La Comédie humaine*. He negotiated with a consortium of publishers for an edition under this name, 17 volumes of which appeared between 1842 and 1848, including a famous foreword written in 1842. In 1845, having new works to include and many others in project, he began preparing for another complete edition. The 17 volumes were republished posthumously, with three supplementary volumes, in 1855. A "definitive edition" was published, in 24 volumes, between 1869 and 1876.

Also in 1834 the idea of using "reappearing characters" matured. He was to establish, as it were, a pool of characters from which he would constantly draw, thus adding a sense of solidarity and coherence to the imaginary world he was superimposing on the real world. A certain character would reappear—now in the forefront, now in the background, of different fictions—in such a way that the reader could gradually form a full picture of him. He first applied the device in *Le Père Goriot*, a masterpiece of realism. It may perhaps be considered an artificial technique, but its effect is to give the reader a convincing sense of being in contact with human experience. Furthermore, Balzac's use of this device places him among the originators of the modern novel cycle.

The middle and final years, 1834–50. "My life-story," Balzac wrote in 1841, "is the story of my work." He must then be pictured continuing his threefold activities as writer, as social lion, and, more specially still, as collector of emotional experience. In 1832 he became friendly with Éveline Hanska, a Polish countess who was married to an elderly Ukrainian landowner. She, like many other women, had written to Balzac expressing admiration of his writings. Thus began a lifelong liaison, born of romantic dreams and intensified by passionate ambitions. They met twice in Switzerland in 1833, the second time in Geneva, where they became lovers; then again in Vienna in 1835. They agreed to marry when her husband died, and so they continued to correspond; the *Lettres à l'étrangère* ("Letters to a Foreigner"), which appeared posthumously (4 vol., 1889–1950), are an important

Use of specific people as characters

source of information for the history both of Balzac's life and work. A striking feature of his work is the way he so ostensibly yet so subtly used his knowledge of specific people, especially women, in conceiving his characters. In 1834 he met another woman who, notwithstanding his increasing devotion to Madame Hanska, was to be his mistress, friend, and patroness for a number of years: Sarah Frances Lowell, the English wife of an eccentric Italian, Count Guidoboni-Visconti . . . "la contessa." In 1835 he also made momentary contact with the beautiful Jane Digby, Lady Ellenborough, and it is typical of his creative technique that in *Le Lys dans la vallée*, a novel contrasting platonic with sensual love, the presence of four women can be detected: "La Dilecta," Éveline, "La Contessa," and Lady Ellenborough. Other women who were not mistresses are traceable elsewhere: the admirable Zulma Carraud, for instance, as Renée de Maucombe in *Mémoires de deux jeunes mariées* (*The Two Young Brides*); his friend the novelist George Sand as the literary celebrity Camille Maupin in *Béatrix* and other works; Marie de Flavigny, Comtesse d'Agoult, mistress of the composer Franz Liszt, in the same novel. Of course, there is always fusion and transformation. It would be superfluous to complete the count of his loves. He was highly sexed, a fact that caused Madame Hanska many transports of jealousy.

Further highlights in Balzac's life include his unsuccessful ventures as editor of *La Chronique de Paris* (1836) and *La Revue Parisienne* (1840); culturally profitable journeys to Italy in 1836 and 1837 on a commission for "la Contessa" and her husband, which resulted in such interesting works as the musical fantasies *Gambara* (1837) and *Massimila Doni* (1839); an abortive expedition to Sardinia in 1838 on a mining quest that might have produced results had Balzac not been born, financially speaking, under an unlucky star; the building in 1838 of a rather fantastic house near Versailles called Les Jardies, which increased his debts and which he left in 1840, taking a house in Passy, another Paris suburb (now the Balzac Museum); from 1839 onward, renewed but frustrated bids for success in the theatre (with, for example, *Vautrin*, 1840); in 1839, efforts on behalf of authors' copyright as president of the Société des Gens de Lettres ("Society of Men of Letters"); vain attempts to save a former fellow journalist from the guillotine for murder; the signing of the contract for the *La Comédie humaine* in 1841; and failure to obtain election to the Académie Française.

In January 1842 Balzac learned of the death of Wenceslas Hanski. He now had good expectations of marrying Éveline, but there were many obstacles, not the least being his inextricable indebtedness. She in fact held back for many years, and the period of 1842–48 shows Balzac continuing and even intensifying his literary activity in the frantic hope of winning her, though he had to contend with increasing ill-health. He joined her again in St. Petersburg in the summer of 1843. After that, his life, apart from his writing and wrangles with publishers or editors, was a story of new meetings and holidays together in western Europe; of a longed-for paternity and miscarriage (1846); of anxiety, anguish, and misgivings: the tragedy of an ailing man longing to find stability. His literary productivity, however, did not visibly flag: *Les Parents pauvres* (*Poor Relations*—consisting of *La Cousine Bette* and *Le Cousin Pons*) is among his greatest works.

In the autumn of 1847 he went to Madame Hanska's château at Wierzchownia and remained there until February 1848. He returned again in October to stay, mortally sick, until the spring of 1850. Then at last Éveline relented. They were married in March and proceeded to Paris to live in a house that Balzac had bought and lovingly furnished. He lingered on miserably until his death on August 18, 1850.

Reputation. Until well into the 20th century, Balzac was chiefly regarded as the creator of realism, or naturalism, in the novel; as a man obsessed with the positive and sordid aspects of life; of limited vision, of too coarse a fibre to understand the aristocracy (though Marcel Proust did not accept this estimation), leaving the working classes out of account though acutely perceptive of every aspect of bourgeois life from the professional classes downward; as giving a vivid picture of artistic and bohemian circles and a grim view of the peasantry; obsessed also with the power of money (everywhere in his work, but above all in *César Birotteau*, and *La Maison Nucingen* [*The Firm of Nucingen*]); intensely sympathetic with, though critical of, the young men of his time who were desperately struggling for recognition and success (as in *Le Père Goriot* and *Illusions perdues*). Balzac's reputation and influence have been worldwide. On the whole, readers have more appreciated his human understanding than his harsher critical qualities. In England and the United States he has had great admirers, including Oscar Wilde, Henry James, W. Somerset Maugham.

His perception of bourgeois life

Balzac is openly acknowledged as one who established the technique of the orthodox classical novel, in which consequent and logically determined events are narrated by an all-seeing observer and characters are coherently presented.

The study of Balzac, however, has taken a new turn. Less attention is now given to him as a scientifically minded determinist, an ardent admirer of both the naturalists Georges Cuvier and Geoffroy Saint-Hilaire, and their differing theories of evolution, and more instead to Balzac the visionary (a quality discerned by his young contemporary the poet Charles Baudelaire): to the man claiming "second sight," the philosopher and the illuminate. Perhaps some critics have gone to extremes in this respect, but justice had to be done to such fascinating works as *La Peau de chagrin*, *Le Chef-d'oeuvre inconnu* (*The Unknown Masterpiece*), *La Recherche de l'absolu*, and *Ursule Mirouët*.

Certainly Balzac had exceptional powers of observation and a photographic memory, but he also had a sympathetic, intuitive capacity to get inside other people's skins. Even in thinking of him as an observer and determinist, it is difficult to decide how far he carried the principle of causation: whether he was in effect a convinced materialist or a mystically oriented idealist. His *Études philosophiques* are of help in resolving this problem, but in all the great "realistic" novels (*Le Père Goriot*, *Le Contrat de mariage* [*The Marriage Settlement*], *La Rabouilleuse* [*A Bachelor's Establishment*], *Les Parents pauvres*) he was bent on illustrating the relation between cause and effect, background and character. This preoccupation explains the main features of his novelistic technique; his long preparative descriptions of antecedents and environment—locality, architecture, houses, atmosphere, furniture, clothes, personal physiognomy—from which his characters emerge. His ambition was to "compete with the civil register," exactly picturing his contemporaries in their class distinctions and occupations. In this he certainly succeeded; but he went even further in his efforts to show that the human spirit has power over men and events—to become, as he has been called, "the Shakespeare of the novel."

Balzac's style has been both praised and disparaged. He is a master of the French language, though perhaps too exhaustive in description and often too emotionally hyperbolic when he tries to soar into the higher reaches of spiritual experience. His conscious wit and humour are questionable—not everyone can stomach the Rabelaisian coarseness of the *Contes drolatiques* (*Droll Stories;* not forming an integral part of *La Comédie humaine*). He is probably at his best in the dry, sardonic mood that he favours above all for winding up his narrations. Brought up on the Gothic novel, he never wholly succeeded in eliminating the lurid and melodramatic from his work. The basic question is to ask how deep did his understanding of human nature go. Most of his readers would answer that it reached rock bottom.

MAJOR WORKS

La Comédie humaine

ETUDES DE MOEURS: (*Scènes de la vie privée*): *La Maison du chat-qui-pelote* (1830); *La Bourse* (1832); *Modeste Mignon* (1844); *Un Début dans la vie* (1842); *Albert Savarus* (1842); *La Vendetta* (1830); *Une Double Famille*

(1830); *La Paix du ménage* (1830); *Madame Firmiani* (1832); *Étude de femme* (1830); *La Grenadière* (1832); *La Fausse Maîtresse* (1841); *Une Fille d'Ève* (1839); *Le Message* (1832); *La Femme abandonnée* (1832); *Honorine* (1843); *Beatrix* (1839); *Gobseck* (1830–35); *La Femme de trente ans* (1831–34); *Le Père Goriot* (1834–35); *Le Colonel Chabert* (1832); *La Messe de l'athée* (1836); *L'Interdiction* (1836); *Le Contrat de mariage* (1835); *Autre étude de femme* (1842).

(SCENES DE LA VIE DE PROVINCE): *Ursule Mirouët* (1841); *Eugénie Grandet* (1833); *Les Célibataires: Pierrette* (1840); *Le Curé de Tours* (1832); *La Rabouilleuse* (1841–42). *Les Parisiens en province: L'Illustre Gaudissart* (1833); *La Muse du département* (1843); *Les Rivalités–La Vielle Fille* (1837); *Le Cabinet des antiques* (1836–38, 1839). *Illusions perdues:* [*Les Deux Poètes*, 1837; *Un Grand Homme de province à Paris*, 1839; *Les Souffrances de l'inventeur*, 1843].

(SCENES DE LA VIE PARISIENNE): *Histoire des Treize-* [*Ferragus*, 1833; *La Duchesse de Langeais*, 1833–34; *La Fille aux yeux d'or*, 1834–35]; *Histoire de la grandeur et de la décadence de César Birotteau* (1834–37); *La Maison Nucingen* (1838). *Splendeurs et misères des courtisanes* (in four parts, 1839–47); *Les Secrets de la Princesse de Cadignan* (1839); *Facino Cane* (1836); *Sarrasine* (1830); *Pierre Grassou* (1840); *Les Parents pauvres* [*La Cousine Bette*, 1846; *Le Cousin Pons*, 1847]; *Un Homme d'affaires* (1845); *Un Prince de la Bohème* (1840); *Les Employés* (1837); *Gaudissart II* (1844); *Les Comédiens sans le savoir* (1846); *Les Petits Bourgeois* (1845–56). *L'Envers de l'histoire contemporaine: Madame de la Chanterie* (1842); *L'Initié* (1848).

(SCENES DE LA VIE POLITIQUE): *Un Épisode sous la Terreur* (1830); *Une Ténébreuse Affaire* (1841); *Le Député d'Arcis* (1847); *Z. Marcas* (1840).

(SCENES DE LA VIE MILITAIRE): *Les Chouans* (1829); *Une Passion dans le désert* (1830).

(SCENES DE LA VIE DE CAMPAGNE): *Les Paysans* (1844); *Le Médecin de campagne* (1833); *Le Curé de village* (1839); *Le Lys dans la vallée* (1835–36).

ETUDES PHILOSOPHIQUES: *La Peau de chagrin* (1831); *Jésus-Christ en Flandre* (1831); *Melmoth réconcilié* (1835); *Massimilla Doni* (1839); *Le Chef-d'oeuvre inconnu* (1831); *Gambara* (1837); *La Recherche de l'absolu* (1834); *L'Enfant Maudit* (1831–36); *Adieu* (1830); *Les Marana, 1832–33; Le Réquisitionnaire* (1831); *El Verdugo* (1830); *Un Drame au bord de la mer* (1835); *Maître Cornélius* (1831); *L'Auberge rouge* (1831); *Sur Catherine de Médicis* (1841); *L'Élixir de longue vie* (1830); *Les Proscrits* (1831); *Louis Lambert* (1832); *Séraphîta* (1834–35).

ETUDES ANALYTIQUES: *La Physiologie du mariage* (1829); *Petites misères de la vie conjugale* (1830, 1840, 1845).

The above arrangement of the *Comédie humaine* is that of the Conard edition (40 vol., 1912–40). The dates after the titles are the printed publication dates.

OTHER WORKS: *Contes drolatiques* (1832–37). (PLAYS): *Le Faiseur* (1844, produced 1851); *Vautrin* (1840); *Les Ressources de Quinola* (1842); *Paméla Giraud* (1843); *La Marâtre* (1848).

ENGLISH TRANSLATIONS: Balzac's works have been extensively translated. There have been two collected editions in English: Dent, 40 vol. 1895–98; Caxton, 53 vol. 1895–1900. The former is the better of the two, but neither is accurate and the language used is archaic. Many novels in these translations have been revised and issued separately, and an 18-vol. reprint series by Books for Libraries Press appeared in 1971. Several individual translations that aim at accuracy and modernity of language have been produced by Penguin Classics, Everyman's Library, and other publishers.

BIBLIOGRAPHY. Balzacian archives, mostly collected in the last half of the 19th century by Vicomte Charles V. Spoelberch de Lovenjoul, are kept at the Chantilly Museum. LOVENJOUL's *Histoire des Oeuvres de Honoré de Balzac* (1879, reprinted 1967), is still indispensable for Balzacian research. See also WILLIAM H. ROYCE, *A Balzac Bibliography* (1929, reprinted 1969); and CHARLES B. OSBURN (comp.), *The Present State of French Studies* (1971).

Editions of complete works: There are many excellent ones. For convenience: Pléiade, *La Comédie humaine*, ed. by M. BOUTERON, 11 vol. (1935–59); Editions du Seuil, *La Comédie humaine*, ed. by PIERRE CITRON, 7 vol. (1965–66); Bibliophiles de l'Originale, *Oeuvres complètes*, ed. by JEAN DUCOURNEAU (1967–), with facsimiles of the original edition; see especially PIERRE BARBERIS, *Aux sources de Balzac: Les romans de jeunesse* (1965). There are many special critical editions. The partly critical editions of the Classiques Garnier are especially useful.

Letters: ROGER PIERROT (ed.), *Correspondance*, 5 vol. (1960–69) and *Lettres à Madame Hanska*, 4 vol. (1967–71) a complete, compact collection.

Periodicals: Les Cahiers Balzaciens, ed. by M. BOUTERON, 8 vol. (1923–28); *Le Courrier Balzacien*, 10 vol. (1948–50); *Les Études Balzaciennes*, 10 vol. (1951–60); *L'Année Balzacienne* (1960–).

Biographies: (*Classical presentations*): LAURE DE SURVILLE, *Notice biographique* (1858); THEOPHILE GAUTIER, *Honoré de Balzac* (1859); EDMOND WERDET, *Portrait intime de Balzac* (1859) by Balzac's one-time editor; LEON GOZLAN, *Balzac en pantoufles* (1869); FERDINAND BRUNETIERE, *Honoré de Balzac, 1799–1850* (1906). (*Later biographies*): ANDRE BILLY, *Vie de Balzac*, 2 vol. (1944); H.J. HUNT, *Honoré de Balzac: A Biography* (1957, reprinted 1969, with a corrigendum sheet bringing it up to date); ANDRE MAUROIS, *Prométhée: ou, La vie de Balzac* (1965; Eng. trans., *Prometheus: The Life of Balzac*, 1965).

General presentations: SAMUEL ROGERS, *Balzac and the Novel* (1953), a very thoughtful work; SOMERSET MAUGHAM, *Ten Novels and Their Authors* (1954; U.S. title, *The Art of Fiction*, 1955); H.J. HUNT, *Balzac's Comédie Humaine* (1959), a complete historical and analytical study that shows the work's expansion and the 1964 ed. brings the Balzac bibliography up to date to that year; JULES BERTAUT (ed.), *Balzac* (1959), includes a highly interesting tribute by MICHEL BUTOR; PHILIPPE BERTAULT, *Balzac* (1962; Eng. trans., *Balzac and The Human Comedy*, 1963), a judicious general study by a great Balzacian scholar; HARRY LEVIN, *The Gates of Horn: A Study of Five French Realists* (1963), a valuable assessment; MAURICE BARDECHE, *Une Lecture de Balzac* (1964), a subtle introduction; F.W.J. HEMMINGS, *Balzac: An Interpretation of La Comédie Humaine* (1967), a perceptive analysis of Balzac in relation to his times; STEFAN ZWEIG, *Balzac* (1946; Eng. trans., 1946), a competent assessment.

Special studies: MARCEL BARRIERE, *L'Oeuvre de Honoré de Balzac* (1890), an early analysis of each work in relation to Balzac's general scheme; MAURICE BARDECHE, *Balzac, romancier*, rev. ed. (1940, reprinted 1967), an initial inquiry into Balzac's evolution up to 1834; ALBERT BEGUIN, *Balzac visionnaire* (1946), follows Curtius (below) in reacting against the view of Balzac as a mere "realist"; PHILIPPE BERTAULT, *Balzac et la religion* (1942), the definitive work on this subject; OLIVIER BONARD, *La Peinture dans la création balzacienne* (1969), describes Balzac's dependence on visual and artistic models for his initial inspirations; RENE BOUVIER and EDOUARD MAYNIAL *Les Comptes dramatiques de Balzac* (1938), the basic inquiry into Balzac's financial affairs; E.R. CURTIUS, *Balzac* (1923), an early perception of Balzac's philosophic importance; E.P. DARGAN and BERNARD WEINBERG, *The Evolution of Balzac's Comédie Humaine* (1942); W.L. CRAIN (ed.), *Le Secret des Ruggieri: A Critical Edition* (1970), is an admirable example of the Chicago "school" of research; J.H. DONNARD, *Balzac: Les Réalités économiques et sociales dans la Comédie Humaine* (1961), a corollary to Bouvier and Maynial; HENRI EVANS, *Louis Lambert et la philosophie de Balzac* (1951), the first serious appreciation of this novel; H.U. FOREST, *L'Esthétique du roman balzacien* (1950), an illuminating study; BERNARD GUYON, *La Pensée politique et sociale de Balzac*, 2nd ed. (1967), a fundamental work; his *La Création littéraire chez Balzac* (1951) initiated the study, now much pursued, of Balzac's method of composition as revealed in manuscripts, successive proofs, and editions; JEAN POMMIER, *L'Invention et l'écriture dans la Torpille d'Honoré de Balzac* (1957), has been very active in this field; SOPHIE DE KORWIN-PIOTROWSKA, *Balzac et le monde slave* (1933) and *Balzac en Pologne* (1933), basic works for Balzac's relations with Madame Hanska; PIERRE LAUBRIET, *L'Intelligence de l'art chez Balzac* (1961), an important thesis; M. LE YAOUANC, *Nosographie de l'humanité balzacienne* (1960). For Balzac's knowledge of pathology, see also J. BOREL, *Médecine et psychiatrie balzaciennes* (1971); FERNAND LOTTE, *Dictionnaire biographique des personnages fictifs de La Comédie Humaine* (1952, with supplementary vol., *Anonymes*, 1956), supersedes the pioneer *Répertoire de la Comédie Humaine de Honoré de Balzac* by ANATOLE CERFBERR and JULES CHRISTOPHE (1887; Eng. trans., *Repertory of the Comédie Humaine*, 1902); FELICIEN MARCEAU, *Balzac et son monde* (1955; Eng. trans., *Balzac and His World*, 1966), a complete study of Balzac's characters, weakened by some lack of historical sense; D.Z. MILATCHITCH, *Le Théâtre inédit d'Honoré de Balzac* (1930), the pioneer study. See also PIERRE DESCAVES, *Balzac dramatiste* (1960); PER NYKROG, *La Pensée de Balzac dans La Comédie Humaine* (1965), perhaps takes the study of Balzac's philosophy too far; DANIEL VOUGA, *Balzac malgré lui* (1957); ADRIEN PEYTEL, *Balzac, juriste romantique* (1950), studies Balzac's very considerable knowledge of the law; M.H. FAILLIE, *La Femme et le code civil dans La Comédie Humaine d'Honoré de Balzac* (1968); GEORGES POULET, *Études sur le temps humain*, vol. 2, *La Distance intérieure* (1956; Eng. trans., *Studies in Human Time*, vol. 2,

The Interior Distance, 1959), examines Balzac's sense of time values; A. PRIOULT, *Balzac avant La Comédie Humaine* (1936), important as a pioneer study of Balzac's "apprentice" works.

(H.J.H.)

Bangkok

Bangkok is the capital and chief port of Thailand. With a population of more than 3,000,000 it is the only cosmopolitan city in a country of small towns and villages; it is also the centre of Thailand's cultural and commercial activity.

One of the most colourful of Eastern cities, Bangkok is located on the delta plain of the Mae Nam Chao Phraya (River of Chao Phraya), about 25 miles (40 kilometres) from the Gulf of Thailand (Gulf of Siam). It was formerly divided into two municipalities—Krung Thep on the east bank and Thon Buri on the west—connected by three bridges. In 1971, the two were united as a single city-province with a single municipal government. Krung Thep is considered as Bangkok proper. It has a low architectural profile, punctuated by tall pagodas, radio and television masts, and occasional high-rise buildings. It is a bustling, crowded city, with temples, factories, shops, and homes standing juxtaposed along its roads and *khlongs* (canals).

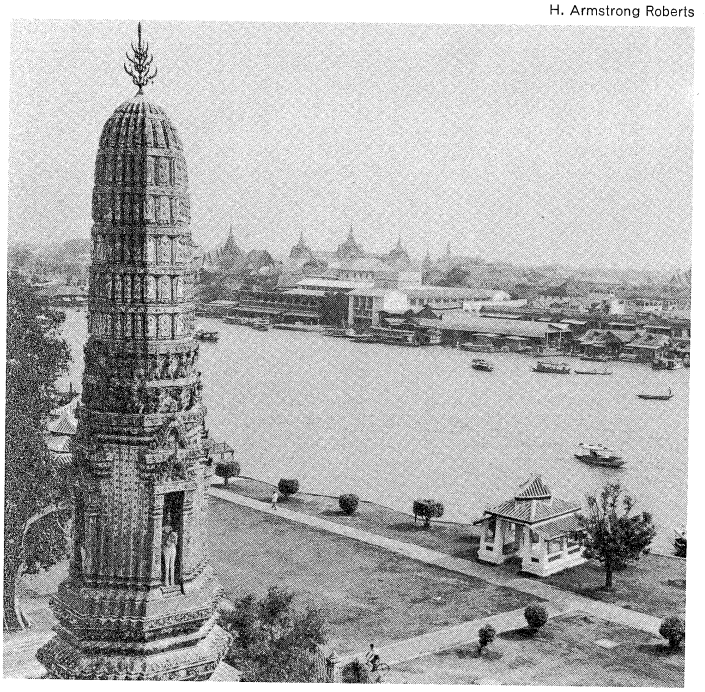
The name Bangkok, used commonly by foreigners and older provincials, is, according to one interpretation, derived from a name that goes back to the time before the city was built—the village or district (*bang*) of wild plums (*makok*). Most Thais call their capital Krung Thep (City of Angels), which is the first part of its mellifluous 27-word official name.

HISTORY

Bangkok became the capital of Siam (as Thailand was previously known) in 1782, when General Chakkri, on assuming the throne as Rama I, moved the citadel from the west to the east bank of the Chao Phraya. The move appears to have been dictated by strategic considerations, because the new site drew an advantage from the wide westward bend in the river that constituted a wide moat guarding the northern, western, and southern perimeters of the citadel. To the east stretched the vast, swampy delta plain, called the Sea of Mud, that could be traversed only with extreme difficulty. Rama I built a city that was intended as far as possible to be the equal in its grandeur of ancient Ayutthaya, 40 miles to the north, and capital of the kingdom for more than 400 years. By the end of the reign of Rama I the city was established, and the walled Grand Palace complex including Wat Phra Kaeo, which housed the most sacred Emerald Buddha, as well as Wat Po, and the city wall were completed. The wall, perhaps the most imposing structure, skirted the river and Khlong Ong Ang to the east; it was four and a half miles long (7 kilometres), 10 feet (three metres) thick, 13 feet (four metres) high, and had 63 gates and 15 forts. The area enclosed amounted to one and a half square miles (4 square kilometres), half of which was occupied by the Grand Palace.

Strategic location of the city

H. Armstrong Roberts



Bangkok on the Chao Phraya River. In the foreground (left) is a portion of Wat Arun (Temple of the Dawn).

More Buddhist monasteries were built during the reigns of Rama II (1809–24) and Rama III (1824–51). They served as schools, libraries, hospitals, and recreation areas, as well as religious centres. During these years Wat Arun, noted for its tall spire, Wat Yan Nawa, and Wat Bowon Niwet, were completed, Wat Po was enlarged, and Wat Sutat was begun. There were, however, few other substantial buildings and fewer paved streets; the river and the network of interconnected khlongs served as roadways.

Rama IV (1851–68) developed the city while continuing, at a reduced rate, the traditional building of *wats*. The Grand Palace was improved, a number of substantial dwellings constructed for members of the royal family, several new streets laid down, and a reduction made in the large number of floating houses anchored along the river front. A new route, Charoen Krung (New Road), leading southward was constructed and a new city moat, Khlong Phadung Krung Kasem, parallel to the city's first canal, was dug and fortified; a long canal led from it to the present port area (Khlong Toei), thus allowing small boats to bypass the big bend in the river immediately south of the city. A pony path, now Phra Ram Thi 4 Road, was laid atop the mud beside this waterway.

The transformation of the city under Rama V

The long reign of Rama V, King Chulalongkorn (1868–1910), saw the city transformed through a program of public works. This great triple-spired Chakkri Building in the Grand Palace was completed by 1880; later the Dusit Palace and an ancillary garden city were built beyond the Wall, being connected to the Grand Palace by European-inspired Ratchadamnoen Boulevard. A road- and bridge-building program was embarked on in earnest, for King Chulalongkorn, an early automobile enthusiast, foresaw the effect that the motor vehicle would have on city development. Most of the now obsolete city wall was pulled down to build the roads, but two forts, a large gate, and a section of the wall were preserved. A post and telegraph service was organized in 1885, an electric tram service instituted on New Road in 1892, and the first line of the State Railway, running from Bangkok to Ayutthaya, opened in 1900. Nor were aesthetic considerations forgotten, for other new buildings included the marble temple of Wat Benchamabopit (1900), elegant bridges in the French style, and the Italian-inspired Throne Hall (now the National Assembly Hall). King Chulalongkorn's reign was also one of population growth. By 1910 the inhabitants numbered about half a million.

Rama VI (1910–25) continued the program of public works. He established Chulalongkorn University in 1916, built a system of locks to control the level of waterways throughout the city, and gave the public their first and largest recreational area—Lumphini Park. During Rama VII's reign (1925–35), municipal areas were delimited as part of a general administrative reorganization aimed at decentralization. In 1937 Bangkok was formally divided into the municipalities of Krung Thep and Thon Buri; at the time of their establishment, the two municipalities, approximately equal in area, together covered about 37 square miles (97 square kilometres). The city's population had grown to over 650,000, four-fifths of whom lived in Krung Thep.

Since World War II, Bangkok—and particularly Krung Thep—has grown with unprecedented rapidity; the population doubled between 1960 and 1970 and the area of Krung Thep has been extended three times—in 1942, 1955, and 1965—to include more than 90 square miles

(238 square kilometres). Extended twice, in 1955 and 1966, Thon Buri now covers 20 square miles (52 square kilometres). With their union in 1971, the Bangkok-Thon Buri metropolis had a total area of 110 square miles (290 square kilometres). Since this growth had not been anticipated and since there is as yet no policy governing the use of or speculation in urban land, it is not surprising that the facilities of the city are now overtaxed, while problems associated with unregulated urban growth affecting transportation, communication, housing, water supply, drainage, and pollution are becoming acute.

THE CONTEMPORARY CITY

Climate. The climate is hot throughout the year, ranging from 77° F (25° C) in the "cold" season in December to 86° F (30° C) at the height of the hot season in April. The mean annual rainfall totals 60 inches (1,500 millimetres), four-fifths of which falls in brief torrential downpours during the late afternoons of the rainy season, which lasts from mid-May through September; the dry season lasts from December to February. Mean monthly humidity varies from a low of 70 percent in the "cold" season to well over 80 percent during the rains.

The city layout. Bangkok is not a planned city and is now undergoing rapid, if not chaotic, changes. The city is becoming much more closely packed in areas long built-up and, at the same time, is sprawling outward into the surrounding agricultural areas. Some districts, however, are evolving into functional units as the inner city becomes more clearly institutional and commercial, and the outer city more residential and industrial.

Principal districts of the city

The government district contains the central administrative offices, which are massed around the Grand Palace and also line Ratchadamnoen Boulevard. A number of large camps around and north of the National Assembly Hall compose the military area. The Chinese quarter of Sam Peng is the main commercial district. Both banks of the river just south of Sam Peng are the site of the city's warehouses, while industry is located at Sam Rong, south of the port. The rich live in Bang Kapi, the squatters in the Klong Toei area, and the poor and not so poor at the Din Daeng and Hua Mak welfare housing estates. Entertainment is concentrated on Pat Pong Road and along Phet Buri Road. The financial district straddles Silom Road.

Traditional areas. Two of the city's identifiable neighbourhoods—the governmental and commercial districts—occupy traditional sites. Government offices were first housed in the Grand Palace and, by the late 19th century, occupied surrounding palaces or palatial mansions. The bureaucracy then spread out into nearby colonial-style or Thai-style office buildings and homes along Ratchadamnoen Boulevard. Severe, multistoried buildings are now being erected to meet the ever-increasing demand for space, and the traditional government compounds are becoming overbuilt.

When the citadel was moved to the east bank of the river in the 18th century, Chinese merchants and tradesmen occupying the site were displaced and moved a short distance southward to the area now known as Sam Peng. Business was at first carried on in one-story wood and thatch houses of diverse shape and size. By the early 1900s a number of streets had been lined with two-story masonry shop-houses. Today, the ever-expanding district contains rows of shop-houses that are sometimes five or more stories high and that are so crowded together that often access is possible only on foot. The area may be likened to a huge department store in which the counters are represented by city blocks fronted by innumerable shops, and the aisles are crowded streets and alleyways.

Transportation. Bangkok's transportation system was originally based on water travel. The city's maze of khlongs connected with the river earned it the name of the Venice of the East. The advent of the automobile, however, brought drastic changes. The number of vehicles in the city has increased to nearly 260,000—including three-wheeled taxis, private cars, and buses that are colour coded according to the region of service—and a shortage of road space has developed. The problem was met first by filling in most of the smaller and a number of the larger canals; this proved to be more than an aesthetic loss because the waterway system had served to drain the deltaic waterlogged city site. Furthermore, though it afforded temporary relief, the measure did not solve the problem of lack of space; traffic today is so congested that movement is increasingly difficult.

Lines of communication radiate outward from the city. Roads run north to Laos and Chiang Mai, east to Cambodia, and south to Malaysia; railways run to the borders of Laos and Malaysia, to Chiang Mai in the north and to Ubon Ratchathani and Nong Khai in the east. Bangkok airport is utilized by 35 international airlines.

The port of Bangkok, located on the Chao Phraya, is connected to the sea by a channel dug through the sandbar at the river mouth some 17 twisting miles (27 kilometres) downstream. The port now handles nearly all of the nation's imports and seven-tenths of its exports. About 30 percent of all inbound cargo is sent elsewhere in the country, while the rest remains in Bangkok to be consumed or processed further.

Demography. The population's outstanding demographic characteristics—its youth and the low proportion of non-Thais—are due to a high rate of natural increase, and restrictive foreign immigration quotas adopted after World War II. Half the residents are under 20 years of age, while another fifth are not yet 30. About 90 percent of the 3,000,000 citizens are Thai, and of the remainder nine percent are Chinese, 0.02 percent are Indians, and 0.02 percent are Americans; the remainder constitute a mixed group of European and Asian nationals.

Despite their small size, the foreign communities tend to live in certain areas. The Chinese concentrate in and around the commercial area of Sam Peng, the Indians gather around mosques in the Wang Burapha section, and the Western community enjoys the affluence of the modern, eastern section of the city.

Of the foreign groups, the Chinese enter the most intimately into city life. They appear to assimilate readily, and intermarriage is the rule rather than the exception. Their offspring are Thai citizens, many families take Thai surnames, and many Chinese are naturalized.

Labour statistics reflect the city's commercial and industrial importance. Thirty percent of those employed are craftsmen and factory workers, 31 percent are sales workers, 19 percent are engaged in services, 10 percent are clerical workers, and 10 percent are administrators or professionals.

Population densities are highest in the Chinese district of Sam Peng where they reach about 390,000 persons per square mile (150,000 per square kilometre). They then diminish rapidly and regularly from this centre to about 25,000 persons per square mile (10,000 per square kilometre) in the more recently built-up outer portions of the city. Approximately half of the population live at densities of less than 50,000 per square mile (20,000 per square kilometre), while no more than one-fifth reside at densities of more than 130,000.

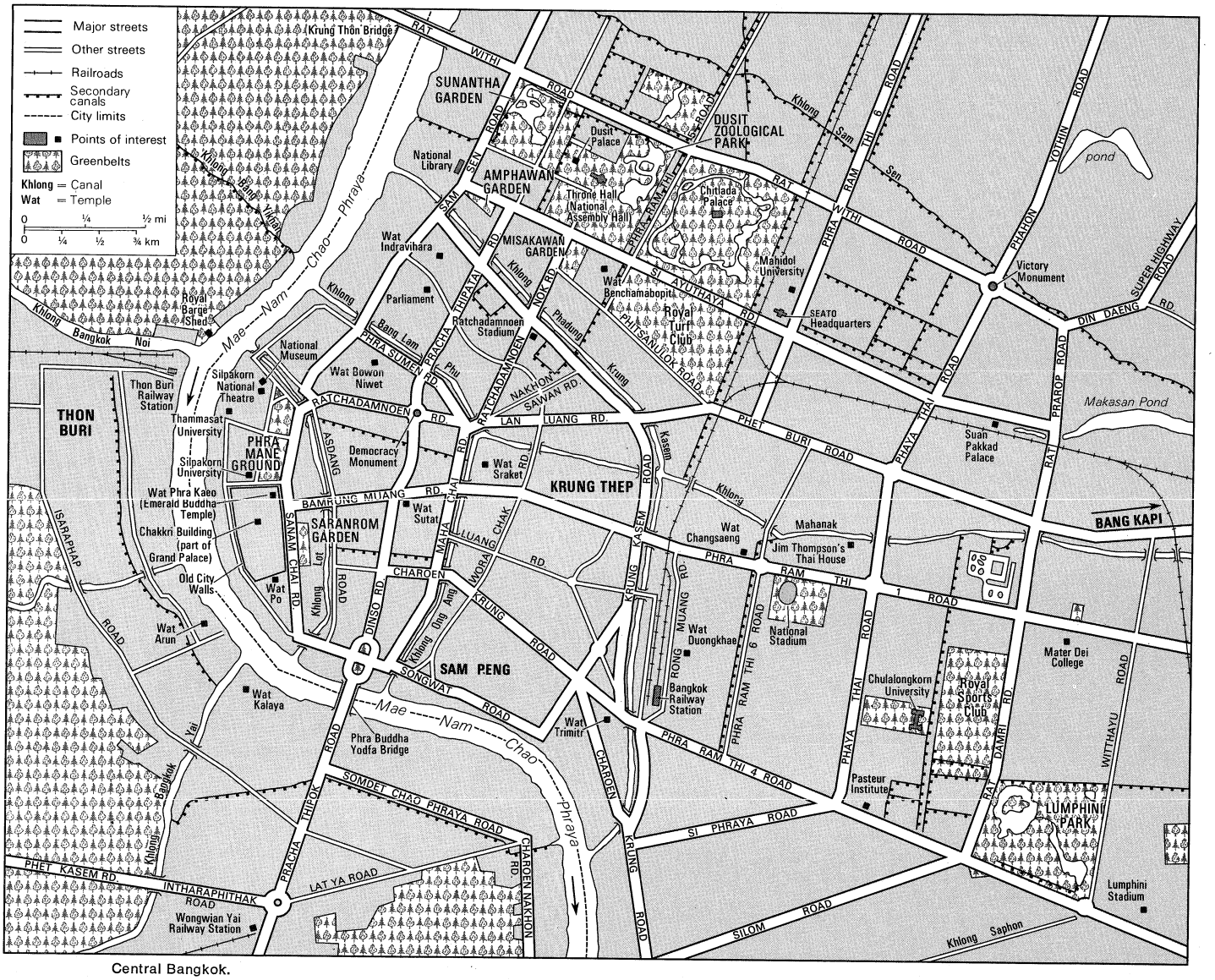
Housing. Homes for the most part consist of small, one- or two-story wooden structures standing close together. Most of these are overcrowded because there are far too few of them to house the ever-expanding population. Government housing programs are insufficient to meet the growing housing shortage; it is estimated that another 100,000 housing units are needed. Fewer than 9,000 units have been built in six public welfare housing projects, and many of these have replaced former homes. At Din Daeng, the largest project, high-rise concrete buildings of 50 to 80 apartments have been constructed.

Housing shortage

The government allows squatters temporarily to occupy unused public land. The number of squatters is small; most of them are concentrated around the port.

Private real estate developers are preoccupied with providing homes for middle-income groups, and many government agencies provide "estates" for their employees. Homes may be crowded onto small lots provided with elementary sanitation facilities. These developments spread out haphazardly on the periphery of the city.

Housing for the rich—mostly for the wealthy foreign



Central Bangkok.

community—usually takes the form of large, modern, two-story masonry structures equipped with separate servants' quarters and kitchens and set in spacious, landscaped compounds. So many of these homes have been built in Bang Kapi that the area has become a maze of dead ends and narrow, sharp-angled lanes. The shops offer Western goods, the modern office buildings house foreign firms and all manner of modern services are available.

Economy. *Industry.* There are many factories in the metropolitan area, the majority of them operating on a small scale. Larger plants are located in the vicinity of the port so as to be near the warehouses that store imported materials. Manufacturing is chiefly confined to food-processing, textiles, and the production of building materials. The city houses the Thai Chamber of Commerce and the Association of Thai Industries. Although trade unions were abolished in 1958, there are 14 specialized trade associations in Bangkok.

Commerce. Bangkok houses about one-third of the country's banking units, holding 75 percent of all deposits. There are 18 Thai banks and 14 branches of foreign banks, as well as the Industrial Finance Corporation of Thailand (IFCT) and the Office of the Board of Investment. The Bangkok Stock Exchange and several insurance companies are also located in the city.

Governmental and international institutions. Prior to their union in 1971, both of the municipalities—Krung Thep and Thon Buri—had a mayor and a municipal assembly whose members were elected by direct voting. The municipal councils were chosen by the assemblies from their members and had executive advisory functions. The government of the newly combined metropolis was to be administered by a governor and two deputies. The governor was to act as mayor of Bangkok–Thon Buri municipality and was to be appointed by the Interior Ministry. Governmental activities are greatly restricted, however, and developmental responsibilities rest with a large number of independent governmental agencies.

In addition to housing the headquarters of the United Nations Economic Commission for Asia and the Far East (ECAFE), Bangkok also houses various other United Nations agencies, including branch offices of the World Health Organization (WHO), the International Labour Organization (ILO), and the International Bank for Reconstruction and Development (World Bank). It is also the headquarters of the Southeast Asia Treaty Organization (SEATO).

Public utilities. Most of the water supply is drawn from the Chao Phraya and is then chemically treated at the municipal plant; nonchlorinated water is then added from deep wells. The supply is uncertain, and some areas are without water during periods of heavy demand.

Sanitation facilities consist of open sewage storm drains and the canals; large buildings are often equipped with septic tanks. Bangkok consumes about 85 percent of the country's electric power. It houses the National Energy Authority and its pilot nuclear plant.

Health and safety. Bangkok benefits by containing most of the country's health facilities. In the late 1960s there were more than 40 government-run general hospitals, 80 private hospitals, and several private clinics in operation. Special services are offered for patients with tuberculosis and venereal disease, and there are government homes for the indigent, handicapped, and aged. The Pasteur Institute supplies vaccines and antivenins.

The metropolitan police are a separate unit within the national police department. About 6,000 strong, it is divided into three sections—North Bangkok, South Bangkok, and Thon Buri—that are patrolled 24 hours a day. The traffic police division also provides mounted escorts and guards for the King. There is a police fire brigade; a detective-training school is located in the city.

Education. Because of its high proportion of school-age citizens, Bangkok's educational facilities are overburdened. In 1968 there were over 660,000 students in Bangkok and Thon Buri. There are too few schools, and the standard of instruction varies. Many of the government-built preprimary and primary schools are located on *wat* grounds. Private primary and secondary schools run by foreign religious missions—including Bangkok Christian College and Mater Dei College—train the children of the elite. There are about 100 private Chinese primary schools and night schools.

Universities

The city houses five universities—Chulalongkorn University (founded 1916), Kasetsart University (1943), Mahidol University (formerly the University of Medical Sciences), Silpakorn University (1943), and Thammasat University (1933).

Cultural life. The most important cultural feature is the *wat*, or Buddhist monastery. There are more than 300 such temples, representing classic examples of Thai architecture. Most *wats* are enclosed by walls. However, many *wats* have leased a portion of their grounds for residential or commercial use. Important monasteries are Wat Po (Temple of the Reclining Buddha), Wat Phra Kaeo (Temple of the Emerald Buddha), and Wat Trimitr (Temple of the Golden Buddha).

The National Museum houses prehistoric, Stone Age, and Bronze Age art relics, as well as royal objects dating back to the 6th century AD. The city also has 12 libraries, including the National Library and the Thai National Documentation Centre. Jim Thompson's Thai House, named for an American entrepreneur and devotee of Thai culture, is composed of five traditional Thai mansions; it contains the country's largest collection of 17th-century Thai religious paintings. There are also collections of Dhavaravadti and Khmer sculpture and of Thai and Chinese pottery and porcelain.

The media. All of the country's daily newspapers and most of its weeklies and monthlies are published in Bangkok. The press is tightly controlled by the government, and the number of newspapers on the stands changes rapidly. In 1971 there were 22 dailies, 20 weeklies, and several biweekly, monthly, and bimonthly publications. Newspapers are printed in Thai, English, and Chinese. The more important papers are the *Thai Rath*, which has the largest circulation; and the *Sri Nakorn*, one of the largest of the Chinese papers.

Radio and television are controlled by the public relations department of the Office of the Prime Minister and by the Royal Thai Army. More than one-half of the nation's 160-odd radio stations are located in or near Bangkok. The four television stations, two of which transmit in colour, broadcast about 100 hours a week. Most programs are in Thai, but some special programs are in English and Chinese.

Recreation. Motion pictures are extremely popular, with about 70 percent of the population attending the films at least once a week. There is a thriving Thai cinema industry but films are also imported from the U.S., Hong Kong, and the U.K. and are subject to governmental censorship. There are also many nightclubs. Fairs, festivals, and kite-flying contests are held at the three large public parks; the Ratchadamnoen and Lumphini stadiums present exhibitions of Thai boxing. Silapakorn National Theatre presents dancing, drama, and music.

BIBLIOGRAPHY. LITCHFIELD, WHITING, BOWNE AND ASSOCIATES; ADAMS, HOWARD AND GREELEY, *Greater Bangkok Plan-2533 (1990)* (1960), plan of the city containing invaluable information; ERIK SEIDENFADEN, *Guide to Bangkok with Notes on Siam* (1927), a beautifully illustrated guide with a wealth of historical data; LARRY STERNSTEIN, *Planning the Developing Primate City:* Bangkok 2000 (1971), a critique and translation of three Thai plans.

(La.S.)

Bangladesh

Bangladesh is an independent Asian state located in the delta of the Ganges and Brahmaputra rivers in the northeastern part of the Indian subcontinent. Until 1971 it was East Pakistan, one of five provinces of Pakistan, which was separated from the other four provinces by 1,100 miles of Indian territory. Bangladesh has an area of 55,126 square miles (142,776 square kilometres) and is one of the most densely populated areas in the world. In about 4 percent of the area the density exceeds 2,000 persons per square mile. The total population in the early 1970s was more than 73,000,000. Bangladesh is bounded by the Indian states of West Bengal to the west and north, Assam to the north, Meghalaya to the north and northeast, and Tripura to the east, and by the Indian union territory of Mizoram to the east, Burma to the southeast, and the Bay of Bengal to the south. The capital is Dacca.

Bangladesh is fringed on the south by the huge expanse of marshy deltaic forest known as the Sundarbans, the abode of the famous royal Bengal tiger. The Bay of Bengal is known for its cyclonic storms, which whip up its waters, sending them crashing onto the coastal areas and the offshore islands, occasionally causing flooding.

Relations with West Bengal

Throughout history the territory's relations with what is now the neighbouring Indian state of West Bengal have been marked by ambiguity. Although in certain periods the two areas were administratively united, their differences have remained: the riverine eastern part (the province) with its Muslim majority and the relatively dry western part (West Bengal) with its predominantly Hindu population. (For associated physical features, see BAY OF BENGAL; BRAHMAPUTRA RIVER; GANGES RIVER. For historical aspects, see INDIAN SUBCONTINENT, HISTORY OF THE.)

PHYSICAL GEOGRAPHY

Physiography. *The rivers.* The most significant feature of the landscape is provided by the rivers, which have molded not only its physiography but also the way of life of the people. The rivers may be divided into five systems—(1) The Ganges, or Padma, as the united streams of the Ganges and Brahmaputra are known, and their deltaic streams; (2) the Meghna and the Surma river system; (3) the Brahmaputra and its adjoining channels; (4) the North Bengal rivers; and (5) the rivers of the Chittagong Hill Tracts and the adjoining plains.

Five major river systems

The Ganges is the pivot of the deltaic river system of Bengal. The river and its tributaries enclose a large area, covering the districts of Kushtia, Jessore, Khulna, Faridpur, Patuākhāli, and Bakerganj. The Ganges Delta itself covers about 20,000 square miles. The Padma enters Bangladesh at the western extremity of the Rājshāhi district and forms, for about 90 miles, the international boundary between Bangladesh and West Bengal (India).

The Meghna, another mighty river, is formed by the union of the Sylhet-Surma and Kusiyāra rivers. These two rivers are branches of the Barāk River, which rises in the Nagar-Manipur watershed in India. The main branch of the Barāk, the Surma, is joined near Ajmiriganj in the Sylhet district by the Kālni and farther down by the Kusiyāra branch. The Dhaleswari, a distributary of the Jamura River, joins the Meghna a few miles above the junction of the Ganges and the Meghna. As it meanders south, the Meghna grows larger after receiving the waters of a number of rivers, including the Burhi Ganga and the Sitallakhya.

The Brahmaputra and its adjoining channels cover a large area from the eastern parts of the districts of North

Bengal to the Meghna River in the southeast. The Brahmaputra (Jamuna) receives waters from a number of rivers, especially on its right bank. The river, with its notoriously shifting channels, not only prevents permanent settlement along its banks but also inhibits communication between the northern area of Bangladesh and the eastern part, where Dacca, the capital, is situated.

The Tīsta is the most important water carrier of the northern districts of Bangladesh. Rising in the Himalayas near Darjeeling (India), it flows southward. After the floods of 1787, however, the Tīsta changed its course, moving southeastward to join the Brahmaputra. That shift caused the rivers of North Bengal to be cut off from the upland waters and led to the deterioration of the natural drainage. Similarly, a number of small and medium-sized rivers in the southwest are silting up, adversely affecting the economic life of that region.

Four main rivers constitute the river system of the Chittagong Hills and the adjoining plains—the Feni, the Karnaphuli, the Sangu, and the Mātāmuhari. Flowing generally west and southwest across the coastal plain, they empty into the Bay of Bengal.

Three main types of soils

Soils. There are three main categories of soils: the old alluvial soils, the new alluvial soils, and the hilly soils, which have a base of sandstone and shale. The fertile new alluvial soils, found mainly in flooded areas, are usually pale brown, sandy, micaceous, and chalky clays and loams. They are deficient in phosphoric acid, nitrogen, and humus but not in potash and lime. The old alluvial soils in the Bārind and Madhupur jungles are dark-brown clays and loams. They are sticky during the rainy season and hard during the dry. The hilly soils support dense forest growth.

Natural regions. Physical geographers, utilizing topographic criteria, have divided Bangladesh into as many as 20 regions; the following 11 are generally regarded as the most important.

The Bārind Tract. The Bārind Tract comprises the districts of Rājshāhi division, between the Ganges and Brahmaputra. The soil of this region is hard, reddish clay, and the region is comparatively elevated.

The Bhar Basin. The depression southeast of the Bārind Tract is called the Bhar Basin. It includes parts of Rājshāhi and Pābna districts, with its centre in the vast marshy area called the Chalan Bīl (bīl, "lake").

The Brahmaputra floodplains. The floodplains of the Brahmaputra (Jamuna) stretch from Bhurungamāri in Rangpur district in the north to Bera in Pābna district in the south. The eastern limit of this zone extends from Dewanganj to Jamālpur in Mymensingh district, while in the west it extends across the Indian border. The area is dominated by the Brahmaputra, which frequently overflows its banks in devastating floods.

The Madhupur Tract. The Madhupur Tract in the east consists of an elevated plateau, with hillocks varying in height from 30 to 60 feet. The valleys, mostly flat, are cultivated. The Madhupur Jungle contains sal trees, whose hardwood is second to teak in value.

The Northeastern Lowland. The southern and southwestern parts of Sylhet district and the northern part of Mymensingh district compose the Northeastern Lowland. It is characterized by a large number of lakes.

The Sylhet Hills. The Sylhet Hills in the east consist of a number of small hills and hillocks ranging from 100 feet to more than 1,100 feet in height. Among the northern Sylhet Hills, Kesara Pahar (500 feet, or 150 metres), Lubhachara (over 300 feet, or 90 metres), and Chatal Tila (over 400 feet, or 120 metres) stand out. A number of hills jut into southern Sylhet from India.

The Meghna Flood Basin. The Brahmaputra River in its old course built up the Meghna Flood Basin, which includes the low and fertile Meghna-Lakkhya Doāb, enriched by the Titās distributary, as well as the *diaras* and *chars* (land areas formed and changed by the deposition of silt and sand in riverbeds) of the Meghna, especially between Bhairab Bāzār and Daudkāndi.

The Central Delta Basin. The Central Delta Basin includes the extensive lakes in the central part of the Bengal Delta in Farīdpur district. The basin's total area is about 900 square miles (2,300 square kilometres).

The Immature Delta. The belt of land in southern Bangladesh bordering the Bay of Bengal constitutes the Immature Delta. The belt—a lowland of some 3,000 square miles (7,800 square kilometres)—contains, in addition to the vast forest known as the Sundarbans, the reclaimed and cultivated lands to the north of it. That area in southern Khulna and southwestern Bakerganj is crisscrossed by a network of streams that flow around roughly oblong islands.

The Active Delta. The Active Delta includes the Dhaleswari-Padma Doāb (*i.e.*, the land between those rivers) and the estuarine islands of varying sizes that are found from the Pusur River in Khulna district in the west to the island of Sandwīp in Chittagong district in the east.

The Chittagong region. Lying to the south of the

MAP INDEX

Political subdivisions

Bākerganj.....22·40n 90·30e
Bogra.....24·51n 89·22e
Chittagong.....22·20n 91·50e
Chittagong Hill Tracts.....22·30n 92·20e
Comilla.....23·28n 91·10e
Dacca.....23·43n 90·25e
Dinājpur.....25·38n 88·38e
Farīdpur.....23·36n 89·50e
Jessore.....23·10n 89·13e
Khulna.....22·48n 89·33e
Kushtia.....23·55n 89·07e
Mymensingh.....24·45n 90·24e
Noākhāli.....22·51n 91·06e
Pābna.....24·00n 89·15e
Patuākhāli.....22·21n 90·21e
Rājshāhi.....24·22n 88·36e
Rāngāmāti.....22·38n 92·12e
Rangpur.....25·45n 89·15e
Sylhet.....24·54n 91·52e
Tangail.....24·15n 89·55e

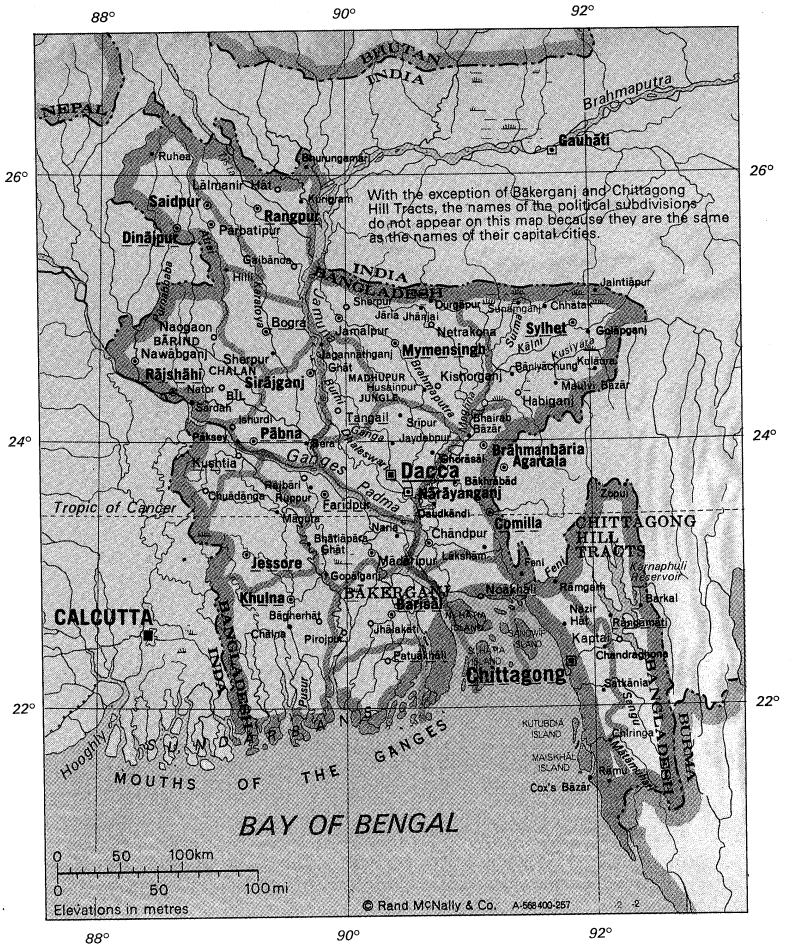
The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Bāgherhāt.....22·40n 89·48e
Bākhrābād.....23·43n 90·53e
Bāniyāchung....24·31n 91·22e
Barisāl.....22·42n 90·22e
Barkal.....22·44n 92·23e
Bera.....23·59n 89·40e
Bhairab Bāzār...24·04n 90·58e
Bhātiāpāra Ghāt.23·13n 89·42e
Bhurungamāri...26·07n 89·41e
Bogra.....24·51n 89·22e
Brahmanbāria...23·59n 91·07e
Chālna.....22·38n 89·30e
Chāndpur.....23·13n 90·39e
Chandraghona...22·28n 92·09e
Chhātak.....25·02n 91·40e
Chiringa.....21·45n 92·05e
Chittagong.....22·20n 91·50e
Chuādānga.....23·38n 88·51e
Comilla.....23·28n 91·10e
Cox's Bāzār.....21·26n 91·59e
Dacca.....23·43n 90·25e
Daudkāndi.....23·32n 90·43e
Dinājpur.....25·38n 88·38e
Durgāpur (Susang).....25·08n 90·41e
Farīdpur.....23·36n 89·50e
Feni.....23·01n 91·20e
Gaibānda.....25·19n 89·33e
Ghorāsāl.....23·56n 90·38e
Golāpganj.....24·52n 92·01e
Gopālganj.....23·01n 89·50e
Habiganj.....24·23n 91·25e
Hilli.....25·17n 89·01e
Husainpur.....24·25n 90·40e
Ishurdi.....24·09n 89·03e
Jagannāthganj Ghāt.....24·45n 89·49e
Jaintiāpur.....25·08n 92·07e
Jamālpur.....24·55n 89·56e
Jāria Jhānjail....25·02n 90·39e
Jaydebpur.....24·00n 90·26e
Jessore.....23·10n 89·13e
Jhālakāti.....22·39n 90·12e
Kaptai.....22·21n 92·17e
Khulna.....22·48n 89·33e
Kishorganj.....24·26n 90·46e
Kulāura.....24·32n 92·02e
Kurīgrām.....25·49n 89·39e
Kushtia.....23·55n 89·07e
Lākshām.....23·14n 91·08e
Lālmanir Hāt.....25·54n 89·27e
Mādārīpur.....23·10n 90·12e
Māgura.....23·29n 89·25e
Maulvi Bāzār.....24·29n 91·42e
Mymensingh.....24·45n 90·24e
Naogaon.....24·47n 88·56e
Nārāyanganj.....23·37n 90·30e
Naria.....23·18n 90·25e
Nator.....24·25n 88·59e
Nawābganj.....24·36n 88·17e
Nāzir Hāt.....22·38n 91·47e
Netrakona.....24·53n 90·43e
Noākhāli.....22·51n 91·06e
Pābna.....24·00n 89·15e
Pāksey.....24·05n 89·03e
Pārbatipur.....25·39n 88·55e
Patuākhāli.....22·21n 90·21e
Pirojpur.....22·34n 89·59e
Rājbāri.....23·46n 89·39e
Rājshāhi.....24·22n 88·36e
Rāmgarh.....22·59n 91·43e
Rāmu.....21·25n 92·07e
Rāngāmāti.....22·38n 92·12e
Rangpur.....25·45n 89·15e
Ruhea.....26·10n 88·25e
Ruppur.....23·41n 89·40e
Saidpur.....25·47n 88·54e
Sardah.....24·18n 88·44e
Satkānia.....22·04n 92·03e
Sherpur.....24·41n 89·25e
Sherpur.....25·01n 90·01e
Sirājganj.....24·27n 89·43e
Srīpur.....24·12n 90·29e
Sunāmganj.....25·04n 91·24e
Susang, see Durgāpur
Sylhet.....24·54n 91·52e
Tangail.....24·15n 89·55e
Zopui.....23·39n 92·14e

Physical features and points of interest

Atrai, *river*.....24·29n 89·03e
Bārind, *physical region*.....25·00n 88·40e
Bengal, Bay of...21·00n 90·00e
Brahmaputra, *river*.....24·02n 90·59e
Burhi Ganga, *river*.....23·37n 90·26e
Chalan Bīl, *wetland*.....24·27n 89·13e
Dhaleswari, *river*.....23·32n 90·34e
Feni, *river*.....22·46n 91·26e
Ganges, Mouths of the, *river mouth*.....22·00n 90·30e
Ganges (Padma), *river*.....23·22n 90·32e
Jamuna, *river*....23·51n 89·45e
Kālni, *river*.....24·21n 91·13e
Karatoya, *river*...24·07n 89·36e
Karnaphuli Reservoir.....22·30n 92·20e
Kusiyāra, *river*...24·36n 91·44e
Kutubdia Island..21·50n 91·52e
Madhupur Jungle, *forest*...24·43n 90·04e
Maiskhāl Island..21·36n 91·56e
Mātāmuhari, *river*.....21·39n 92·00e
Meghna, *river*....22·50n 90·50e
North Hātia Island.....22·40n 91·00e
Padma, see Ganges
Pūrnabhaba, *river*.....24·50n 88·18e
Pusur, *river*.....21·45n 89·36e
Sandwīp Island..22·30n 91·25e
Sangu, *river*.....22·08n 91·51e
South Hātia Island.....22·19n 91·07e
Sundarbans, *physical region*..22·00n 89·30e
Surma, *river*.....24·34n 91·14e
Tīsta, *river*.....25·23n 89·43e



BANGLADESH

Feni River, the Chittagong region is full of hills, hillocks, valleys, and forests and is quite different in aspect from other parts of Bangladesh. The coastal plain is partly sandy and partly composed of saline clay; it extends from the Feni River to Cox's Bāzār and varies in width from one to 10 miles. In the south, the plain is narrowed by the Jaldi and Khurushkul hills. The region has a number of offshore islands and one coral island, St. Martin's, off the coast of Burma.

Climate. Bangladesh has a typical monsoon climate characterized by rain-bearing winds, moderately warm temperatures, and high humidity. In general, maximum temperatures in the summer months, from April to September, range between 91° and 96° F (33° and 36° C). April is the warmest month in most parts. January is the coldest month in the winter season, which lasts from about November to March.

The conditions of lowest atmospheric pressure occur all over Bangladesh in June and July, the storm season. Winds are mostly from the north and northeast in winter, blowing gently at a rate of one to two miles per hour in northern and central areas and two to four miles per hour near the coast. During the period of the nor'westers (March to May), wind speeds may rise to 30 or 40 miles per hour.

Rainfall

Bangladesh receives heavy rainfall; except for some parts in the west, it generally exceeds 60 inches (1,500 millimetres) annually. Large areas of the south, southeast, north, and northeast receive from 80 to 100 inches (2,500 millimetres), and the northern and northwestern areas of Sylhet district receive from 150 to 200 inches. The maximum rainfall occurs during the monsoon period, from June to September or early October.

In the early summer (April and May) and late in the monsoon season (September to November), storms of very high intensity often occur; they may create winds with speeds of more than 100 miles (160 kilometres) per hour, piling up the waters of the Bay of Bengal to crests as high as 20 feet (six metres) that crash with tremendous force onto the coastal areas and the offshore islands, inundating them and causing heavy losses of life and property. One such cyclone and tidal wave that battered the coastal areas in November 1970 killed at least 250,000 people and rendered millions destitute.

Vegetation and animal life. *Vegetation.* About 16 percent of the land surface is covered with forests. Most of them are situated in the districts of Chittagong, Chittagong Hill Tracts, Sylhet, and Khulna, but there are also forests in the districts of Dacca and Mymensingh. Bangladesh in general possesses a luxuriant vegetation, with villages appearing to be virtually buried in groves of mango, jackfruit, bamboo, betel nut, and date palm.

Bangladesh has four different areas of vegetation. The northeastern zone, consisting of parts of Sylhet and Chittagong districts, has many low hills covered with jungles of bamboo and rattan (a species of climbing palm). The commonest plant is a large kind of bamboo that is the basis of the country's paper industry. The central zone, covering parts of Dacca and Mymensingh districts, contains a large number of lakes and swampy vegetation. The soil of part of this zone is laterite, which produces the Madhupur jungles. The area lying to the northwest of the Brahmaputra and to the southwest of the Padma forms a flat plain, the vegetation of which consists mostly of cultivated plants and orchards. Babul (*Acacia arabica*) is the most conspicuous plant. The southern zone, comprising the districts of Khulna and Barisāl, contains the Sundarbans, with their distinctive mangrove vegetation. Many commercially valuable trees, such as the *sundri,* after which the Sundarbans are named (*Heritiera fomes* or *minor*); *gewa,* or *gengwa* (*Excoecaria agallocha*), a softwood tree used for making newsprint; and *goran* (*Ceriops roxburghiana*), a kind of mangrove, grow in this vast forest.

Among the common flowers are marigolds, Bengal roses, water lilies, gandharaj (a tuberose), bokul (*Mimusops elengi*) and kamini (orange jasmine).

Animal life. Bangladesh is said to have about 200 species of mammals, 750 of birds, 150 of reptiles, and about 200 species of marine and freshwater fishes. Elephants are found in the Chittagong Hill Tracts living in herds ranging from less than a dozen to about 100. They are the so-called Indian elephants that can be captured and trained. Buffalo, both wild and domesticated, are also found. The domesticated, or water, buffalo are used for plowing and pulling carts. Of the different kinds of deer, the barking deer, the barosinga, or 12-horned deer, and the sambar deer, with its maned neck, are well-known. The barosinga, which reach a height of about four feet at the shoulders, mostly inhabit the Sundarbans. The sambar, which lives in the eastern jungles of the province, attains a height of four and a half feet and a length of six to seven feet. By comparison with these, the spotted deer, the barking deer, and the hog deer are smaller.

The elephant herds of the Chittagong Hill Tracts

Of the carnivores, the Bengal tiger is the best known. The famous denizen of the Sundarbans measures nine to ten feet from head to tail and feeds mostly on small animals. The leopard, though smaller than the tiger, is more active and is capable of climbing trees. The clouded leopard, dark gray and with spots that are oval or oblong in form, is smaller than the leopard. The leopard cat, about the size of the domestic cat but with longer legs, is ferocious.

Of the three types of bear usually seen—the sloth bear, the Himalayan black bear, and the Malayan bear—the sloth bear is the most numerous. Usually black and about two and a half feet high at the shoulder, the sloth bear is nocturnal and feeds on fruits, insects, and honey from beehives. The Malayan bear is smaller in size. The jackal is a common animal; it feeds on carrion of all kinds and is adept at stealing lambs and poultry from homesteads. Its eerie howling at night is a familiar sound in Bangladesh. The mongoose, grayish brown and about two and a half feet long, including the tail, feeds on rats, mice, lizards, and insects and also kills snakes. The Bengal, or rhesus, monkey is the commonest of primates in the province.

The common house crow is found everywhere, and its shrill cries are detested by the people, who regard them as a bad omen. The bulbul is a beautiful, small, smokey-brown bird. The colour of the magpie robin is a mixture of black and white. A wide variety of warblers are found; some are migrants that appear only in winter. Several kinds of flycatchers also occur. There are also myna birds of several kinds. Other species include the common game birds, cuckoos, hawks, owls, kingfishers, hornbills, and woodpeckers, as well as vultures, which act as scavengers. Among the eagles, the crested serpent eagle and the ring-tailed fishing eagle are the most common. There are also hoopoes, herons, storks, ducks, and wild geese.

POPULATION

Ethnic composition and distribution. Ethnically, Bangladesh is a melting pot of races. The following groups are the most important:

The proto-Australoids, sometimes called Veddas, who were longheaded, flat nosed, and dark brown in complexion, were one of the earliest groups to enter the area. According to some ethnologists, they were followed by longheaded Mediterranean caucasoids (whites), also known as Aryans. Short-headed or broad-headed Armenoids (of Indo-European stock) are believed to have entered as well. It is presumed that the process known as Aryanization began in the first millennium before Christ.

With the coming of the Muslims in the 8th century AD, new elements were introduced. Persons of Arab, Persian, and Turkish origin moved in large numbers to the subcontinent and gradually at different times entered what is now Bangladesh. The contention that Bengali Muslims are all descended from lower caste Hindus who were converted to Islām is incorrect; a substantial proportion are the descendants of the Muslims who reached the subcontinent from elsewhere.

The peoples known as the Santāls and the Khāsis possibly belong to the proto-Australoid group—the former living in the Bārind area in the districts of Rājshāhi and Dināj pur and the latter in the Khāsi Hills near the border with Assam. The Mongoloid element is represented by the Gāro, Hajang, Kachari, and Tipera groups. The Gāros live in the Susang Hills region; the Hajang are a branch of the Gāros. The Kacharis live mainly in the district of Kachar in Assam (India), but some of them are found in the Sylhet district of Bangladesh. The Tiperas mostly live in the Chittagong Hill Tracts.

Apart from these tribes, the rest of the people are Bengalis—an ethnic as well as a linguistic group. The Bengalis, however, are not homogeneous in origin. Differences in head shapes and in skin colour are found among them. In general, the people of the coastal areas with whom the Muslim merchants of the Middle East were in close touch show physical features that seem to be the result of the admixture of local peoples with peoples of Turkish and Semitic origin.

The Bengalis

Linguistic pattern. Bengali, the language spoken in Bangladesh, belongs to the Indo-Aryan group of languages. Like Pāli and various other forms of Prākrit in ancient India, Bengali originated beyond the influence of the Brahmin society of the Aryans. The Pāla rulers of Bengal, who were Buddhists and whose religious language was Pāli, did not inhibit the emergence of a colloquial tongue; it was known as the Gaudiya Prākrit, the language from which Bengali is derived.

Bengali is spoken by about 98 percent of the people. The remainder, largely immigrants who came to Bangladesh, then East Pakistan, after 1947, generally speak Urdu. English is an important means of communication with West Pakistanis, who speak four other languages.

While literary Bengali is the same all over Bangladesh, there are slight dialectal differences from region to region. The dialects of the districts of Sylhet, Noākhāli, and Chittagong are the most markedly different. Bengali contains a large number of loanwords from Portuguese, English, Arabic, Persian, and Hindi. Words derived either directly or indirectly from Sanskrit, however, predominate in the literary idiom.

Religious affiliations. An overwhelming majority—more than 80 percent—of the population profess the religion of Islām. The advent of a handful of Muslims to Bengal at the beginning of the 13th century and the rapid expansion of their rule permanently changed the character and culture of the area. When the Muslims first arrived, the Hindus were in an overwhelming majority; there were also some Buddhists and a few animists. The Hindus remained in the majority throughout the Turko-Afghan and Mughal periods. Even as late as 1872 there were more than 18,000,000 Hindus in Bengal, compared to about 16,000,000 Muslims. From the 1890s onward, however, the balance began to tilt in favour of the Muslims, and they have since remained in the majority.

There are several reasons for the increase in the Muslim population. Perhaps the most significant was the activity of ascetic divines and Ṣūfī (mystics), who won converts among the lower castes of the Hindus. Second, there was an influx of Muslims from northern India as well as from outside India. Third, the Muslims had a higher birthrate than that of the Hindus, since, for religious reasons, Hindu widows were not allowed to remarry.

The majority of the Muslims belong to the Sunnī sect. There is, however, a small sprinkling of Shīʿite Muslims, mostly descendants of immigrants from Persia.

Hindus are divided into two broad groups—high-caste and low-caste. The high caste Hindus constitute almost 9 percent of the population and the low castes almost 10 percent. The Buddhists, numbering about 400,000, form less than 1 percent of the population. Of the tribes in the Chittagong Hill Tracts, the Chakma, Chak, Magh, and the Mru are mostly Buddhists. The Kumi and some of the Mru are animists. While most of the Bon, Kuki, and Lushais are Christians, the Tiperas are high-caste Hindus, and the majority of the Piangs are low-caste Hindus.

Bangladesh, Area and Population

	area		population	
	sq mi	sq km	1961 census	1971 estimate
Divisions				
Chittagong				
Districts				
Chittagong	2,705	7,006	2,983,000	...
Chittagong Hill Tracts	5,093	13,191	385,000	...
Comilla	2,594	6,718	4,389,000	...
Noakhali	1,855	4,804	2,383,000	...
Sylhet	4,785	12,393	3,490,000	...
Dacca				
Districts				
Dacca	2,882	7,464	5,096,000	...
Faridpur	2,694	6,977	3,179,000	...
Mymensingh	5,060	13,105	5,532,000	...
Tangail	1,301	3,370	1,487,000	...
Khulna				
Districts				
Bakerganj (Barisal)	2,609	6,757	3,068,000	...
Jessore	2,547	6,597	2,190,000	...
Khulna	4,652	12,049	2,449,000	...
Kushtia	1,371	3,551	1,166,000	...
Patuākhāli	1,631	4,224	1,194,000	...
Rājshāhi				
Districts				
Bogra	1,502	3,890	1,574,000	...
Dinajpur	2,609	6,757	1,710,000	...
Dabna	1,877	4,861	1,959,000	...
Rājshāhi	3,654	9,464	2,811,000	...
Rangpur	3,704	9,593	3,796,000	...
Total Bangladesh	55,126*	142,776*	50,840,000*	73,068,000

*Figures do not add to total given because of rounding.
Source: Official government figures.

The pattern of rural and urban settlement. The economy is predominantly agrarian. At the time of the first census, in 1872, the population of the area now forming Bangladesh totalled about 22,000,000; it rose steadily to 35,000,000 in 1931, 50,000,000 in 1961, and more than 73,000,000 in 1971.

Urban settlement. Bangladesh is one of the least urbanized areas in South Asia. There are only three cities, although there are more than 80 towns of varying sizes. In 1951 the urban population numbered 1,800,000 —about 4 percent of the total population. By the early 1970s about 3,000,000 lived in cities—still only 4 percent of the total. Dacca, the capital, is the biggest city and in 1972 was estimated to have a population of about 960,000. Ten miles from Dacca is the town of Nārāyanganj, the centre of the jute industry. Dacca and Nārāyanganj are sometimes conjointly referred to as Greater Dacca City, with a combined population of nearly 1,400,000.

Major cities

Chittagong is the second most important city. Once a sleepy little river port, since 1947 it has been the most important port in the territory. In 1972 Chittagong's population was estimated at 488,000. A number of industrial areas, such as Kalurghat, Sholāshahar, and Faujdār Hat are developing around Chittagong.

Khulna, in the southwest, is becoming a commercial and industrial centre; the opening of the port of Chālna nearby and the growth of the Daulatpur industrial area have increased its population to more than 200,000 inhabitants.

The houses in the towns and cities are mainly of one or two stories. It is common for a new town to grow around an older, historical one, as at Dacca. Industrial development has prompted migration to the cities from the rural areas.

Rural settlement. The rural area throughout the province is so thickly populated that it is difficult to distinguish any well-defined pattern of settlement. There are, however, some noticeable features. The inundation of most of the fields during the rainy season makes it necessary to build homes on higher ground. Continuous strings of settlements along roads are very common in the districts of Kushtia, Jessore, and Faridpur, in parts of Khulna, Bakerganj, and Patuākhāli, and in the floodplains of the Mahānanda, Tīsta, Jamuna, Ganges, and Meghna rivers. Similar settlements are also to be found in the hilly regions of southern Sylhet and in the Chittagong region. Settlements are more scattered, however, in parts of Khulna, Bakerganj, and Patuākhāli districts, in the floodplains of the Brahmaputra, in eastern and southern Sylhet, in Comilla district, and in parts of Chittagong district. In the Haor Basin of Sylhet and in the Chittagong Hill Tracts, settlements occur in a nucleated, or clustered, pattern.

The villages are composed of thatched bamboo huts. There is an adequate water supply, and the buildings within the settlements are scattered.

THE ECONOMY

Agriculture. Bangladesh is overwhelmingly agricultural, with more than 80 percent of the population engaged in farming. Of Bangladesh's total land area of 36,000,000 acres (15,000,000 hectares), about 23,000,000 acres are cultivated. The soil is fertile and has been so for centuries. Travellers from the 7th century onward consistently testified to the richness of Bengal's soil and the abundance of its harvests. Jute and rice are the most important agricultural products. Before 1971, the territory produced about half the world's supply of jute. The total yield in an average year was more than 5,000,000 bales (of 400 pounds each); the total area devoted to jute cultivation was over 2,000,000 acres. The average yield is 1,000 pounds of jute per acre. Other important agricultural products are pulses (leguminous plants, such as peas, beans, and lentils), potatoes, oilseeds of various kinds, sugarcane, tobacco, and fruit.

Jute and rice production

Agriculture has in the past been wholly dependent upon the vagaries of the monsoon. A poor monsoon always meant poor harvests or none and the threat of famine. Among the remedial measures adopted in recent years have been the construction of a number of irrigation projects designed to control floods and to conserve rainwater for use in the dry months. The most important are the Karnaphuli Multipurpose Project, the Tīsta Barrage Project, and the Ganges–Kobadak Project. The first is in the Chittagong district in the southeast, the second is in the north, and the third will serve the districts of Kushtia, Jessore, and Khulna.

Fisheries. Bangladesh, a riverine area, has great possibilities in pisciculture (the breeding and raising of fish). Its rivers and seacoast offer opportunity for the operation of the usual type of fisheries. There are approximately 2,200,000 acres of inland fishery waters and 20,000,000 acres of marine fishery waters, mostly in the estuaries of the Bay of Bengal. It has been estimated that the annual yield from inland fisheries is around 240,000 tons and from marine fishing over 58,000 tons. Among the varieties of fish caught are the marine rupchanda, or pomfret, and the freshwater hilsa, a relative of the shad.

Industry. Excessive—until recently almost exclusive—dependence upon agriculture in an area in which during the monsoon period no work in the fields is possible has led to seasonal unemployment among peasants, as well as to a low standard of living. To counteract this imbalance, a policy of industrialization was adopted after 1947 and, until 1971, was pursued through the successive five-year plans. The main obstacle to its fulfillment has been the comparative lack of mineral resources.

Power resources. Oil in marketable quantities has not been struck anywhere in Bangladesh; vigorous prospecting has, however, been in progress. The consumption of petroleum products in the territory rose from 70,000 tons in 1948 to more than 1,000,000 tons in 1968. Bangladesh is also dependent upon imports for coal. Two coalfields have been discovered, one at Sylhet and the other at Bogra, but commercial exploitation has not yet begun.

Natural gas holds more promise. It has been found at Harīpur, Chhātak, Kailash-Tila, Rashīdpur, and Shahi Bāzār (all in the district of Sylhet), as well as at Brāhmanbāria and Bakherabad, in Comilla district. The Harīpur field is estimated to have a reserve of about 180,000,000,000 cubic feet, with gas of fairly good quality. The Harīpur field and the gas fields of Rashīdpur and Kailash-Tila supply the Fenchuganj Natural Gas Fertilizer Factory, which is also in Sylhet. The Titās field in Brāhman-

bāria began supplying gas to industrial consumers in Dacca and its suburbs in 1968.

The output of electricity

In 1947 the territory had an installed capacity of 21,000 kilowatts of electricity. This rose to 173,000 in 1960 and to about 535,000 in 1970. Thermal power stations have been built at Ghorāsāl, in Dacca district, and at Chittagong, producing, 110,000 kilowatts and 60,000 kilowatts, respectively; a few other similar projects have also been constructed. The annual per capita consumption of electricity was about 13 kilowatt-hours in 1970. A nuclear-power station is under consideration at Ruppur, in the district of Pābna.

Industrial development. Industrial policy after 1947 was to give priority to industries based on indigenous raw materials such as jute, cotton, hides, and skins. Up to 1971, the principle of free enterprise in the private sector was accepted, subject to certain conditions including the national ownership of public utilities. The policy also aimed to develop as quickly as possible consumer-goods industries with a view to avoiding dependence on imports. Attempts were also being made to establish heavy industries, such as steel.

To promote this policy, various institutions were established. Some of the important ones were the East Pakistan Industrial Development Corporation (EPIDC), the East Pakistan Small and Cottage Industries Corporation, the Industrial Development Bank of Pakistan, the Equity Participation Fund, and the Pakistan Industrial Credit and Investment Corporation.

The actual task of establishing industries rested on the EPIDC, which sponsored new industrial units producing jute textiles, paper and board, sugar, cotton textiles, nonmetallic mineral products, basic metals, chemicals and fertilizers, and fuels and minerals, as well as engineering and shipbuilding plants.

Manufacturing and other industries. Since the export of raw jute was not very remunerative, efforts were made to establish mills to produce and export jute products and thus earn foreign exchange. About 45 percent of the jute produced was processed in the territory; the balance was exported raw.

The abundance of bamboo in the Chittagong Hill Tracts and the various softwood trees growing in the Sundarbans provide excellent raw material for papermaking. There are paper mills at Chandraghona (Chittagong) and at Pāksey, in Pabna district, as well as a paper and board mill at Khulna.

Bangladesh has several fertilizer factories and one cement factory, which is located at Chhātak, in the district of Sylhet, and which, although it has a capacity of 250,000 tons, is quite unable to meet the growing demand for cement. A shipyard has been opened at Khulna for repairing and reconstructing ships. Two dry docks—one at Chittagong and one at Khulna—were under construction in 1970.

Cottage industries

By far the most important cottage industry centres on the production of yarn and textile fabrics. More than 300,000 people produce coarse and medium-quality fabrics. Another cottage industry produces cigarettes known as *bidi*s and employs more than 20,000 workers. Carpets, ceramics, and cane furniture are also products of cottage industries. (S.S.H.)

ADMINISTRATION AND SOCIAL CONDITIONS

Administration. The administration of Bangladesh began in a *de facto* manner on December 16, 1971, after the surrender of the Pakistan army in Bangladesh. During much of 1972 the new administration had to cope with the confusion engendered by the 1971 civil war, which had resulted in the country's independence. Thus reliable statistics, as opposed to rapid assessments, were not available. The administrative structure, however, remained substantially the same in form as it had been before the civil war, although certain policies and many of the personnel were not the same.

The constituent assembly. In late 1972 the draft of the Bangladesh Constitution was completed for submission to the constituent assembly, a body comprised at that time of 403 members. The Assembly approved the draft, and it became effective on December 16, 1972, establishing a parliamentary form of government and dissolving the Assembly. Sheik Mujibur Rahman, leader of the Awami League political party, which had won a landslide victory in the 1970 elections, formed a Bangladesh government. His cabinet was composed primarily of persons who before independence had been members of the government-in-exile, and he administered the affairs of state through the office of the presidency by means of presidential ordinances and orders. The Cabinet under the new constitution was to consist of members of the Parliament to be elected in March 1973. In 1972 the president was the titular head of the People's Republic of Bangladesh, and he acted upon the advice of the prime minister.

Post-war problems. The new administration faced a formidable combination of social, economic, and sometimes political problems arising out of the 10-month civil war. The restoration of law and order, for example, posed a major problem; weapons had to be recalled from youths who had joined the liberation forces; citizens accused of "war crimes" had to be arrested and detained in an orderly fashion; members of communities of non-Bengali origin (Biharis) were frequently open to charges of "collaboration"; and the army and police forces required major reorganization. The administrative tasks arising from these circumstances were made more difficult by the shocked and emotional state in which the people emerged from a civil war that had produced many inhuman acts, including systematic terrorization and massacres of civilians.

The refugee problem

Approximately 8,000,000 to 10,000,000 refugees, most of whom were Hindus, had fled across the border to India; substantial numbers of Muslims, including students, Awami League supporters, and others, were, however, also among them. The number of those who were displaced from their homes but remained within the territory that is now Bangladesh could not be reliably assessed; figures of between 10,000,000 and 20,000,000 were frequently assumed. The refugees returned in early 1972, but their physical and financial resources were at a low ebb, and many of them returned to derelict farms or businesses, and to ruined houses.

The difficulties of administering adequate supplies of relief and rehabilitation materials, and of providing necessary services, were compounded by the destruction of road and rail communications (particularly bridges), the obstruction of major ports and channels by sunken vessels, and by increases in the price of essential items on the open market at a time when employment opportunities were scarce, and when smuggling and hoarding were financially profitable. The normal fixed-price ration shop distribution system was unable to meet the acute demand for some time, although the United Nations Relief Operation in Dacca gave technical assistance, particularly with transport facilities. The quantities of foodstuffs that had to be imported was variously estimated at between two and three or more million tons up to the end of 1972. (The lower figure also represented the territory's normal shortfall in food production.) The United Nation's Children's Fund (UNICEF) sponsored a nationwide child feeding program to combat malnutrition. Wherever possible the Bangladesh government pursued a policy of "Test Relief," using imported, donated, or locally acquired foodstuffs as payment in kind to needy persons for manual work on public projects. Although the incidence of malnutrition inevitably increased in the worst-affected communities, widespread famine did not prevail.

The civil service. The civil service was modelled upon the pre-1947 Indian Civil Service, whose members were at the apex of the administration. Civil servants were appointed by the Civil Service Commission, and selection was made by competitive examinations. The administrative structure had two broad categories—the Secretariat, and the "field officers." The former was the administrative policy-making body, while the latter was organized into a pyramidical hierarchy with a well-defined chain of command. During the civil strife, many depart-

ments functioned in only a token manner, and were often completely dependent upon military exigencies. At the conclusion of the civil war, the civil service continued as a national institution of substantially the same character as it had been previously. The four area divisions, previously subdivided into 19 districts and 62 subdivisions, in 1972 were reorganized as 62 new district units, thus promoting the decentralization of administrative authority. At a lower level, however, the former administrative structure of "circles" (*thanas*), each having a population of between 100,000 and 200,000, was continued. Each unit was headed by a "circle officer" responsible for revenue collection and for certain developmental functions. In each of these units there were also police officers, serving side by side with the civil servants responsible for administration.

The autonomous corporations. Under the Pakistani administration, new types of autonomous corporations were developed over the years to deal with industrial development, water and sewerage management, the development of forest industries, and road transportation. The Bangladesh government in 1972 nationalized these corporations, and then established 12 new corporations to manage the nationalized enterprises. At the same time, the government revealed its intention of having these corporations managed by personnel with professional and commercial experience, rather than relying on civil servants, as had formerly been the practice.

Justice. Bangladesh has continued with substantially the same judicial system as had been in operation when the territory was a province of Pakistan, and which owed its origins to the system in operation under the British Raj before 1947. The Supreme Court of Pakistan had previously existed at the apex of the system as the final court of appeal for three types of jurisdiction: original, appellate, and advisory. High Court judges were appointed by the President of Pakistan, and could not be removed except on charges of misconduct. Although the judiciary was comparatively free, the separation of judicial and executive powers was not fully achieved, and officials from one power had at times been appointed to the other. After the civil war, the Bangladesh Supreme Court became the highest court in the land. The 1972 constitution divided the Supreme Court into Appellate and High Court divisions and envisaged a complete separation of the judiciary and executive branches.

In rural areas, district courts were situated in town centres, where district judges tried both civil and criminal cases. The district magistrates were of different classes, corresponding to the seriousness of the cases that they were allowed to hear, and to the limits of punishment that they could impose.

In general, the concept of justice suffered greatly during the civil war; the Pakistan army considered that the continued unity of the country as an Islamic state was the paramount consideration. The military were assisted by ad hoc "peace committees" which were established in most localities, purportedly to help restore normalcy and actively to cooperate with military administrators and commanders in the field operating against the Mukti Bahini (nationalist guerrillas) and their sympathizers. Many of those who cooperated with the military were later arrested, and shared the blame for the tragic excesses of killing and destruction that took place. The trials of "war criminals" and "collaborators" was a major issue in the earlier period of Bangladesh politics. A tribunal of the type employed at the Allied War Crimes trials at Nürnberg after World War II was mooted, especially for those Pakistani officers who were captured.

The "peace committees"

Local government. Although the tradition of local self-government through *panchayats* (councils of village elders) throughout the subcontinent goes far back into history, it was only in 1880 that, under British administration, the system was organized along modern lines. After 1947, under the initial Pakistan administration, the functions of local self-government that were obligatory covered such public services as primary education, local road construction and maintenance, sanitation, drainage, water supply, lighting, slaughterhouses, markets, and local health services, as well as the power to hear petty criminal and civil cases. These functions were administered under a Department of Local Self-Government through district boards or municipalities, and with "union councils" as the lowest unit in the administrative hierarchy. In 1958, after the introduction of martial law, a system known as "Basic Democracy" was substituted; this partly reflected a gradualist approach to democracy, and also promoted community development and change through greater citizen participation. Each "Basic Democrat" was elected by a population of between 1,000 and 1,500 (roughly one village), and ten such representatives, together with an additional five who were nominated to serve, made up one "union council." These councils had municipal, judicial, and community organization functions. At the next level in the hierarchy were the "*thana* councils," which were primarily concerned with developmental activities within their respective areas. Half of the *thana* council members were chosen by indirect elections conducted in the union councils, while the other half were *ex-officio* government servants. While the system of unions and *thanas* that Bangladesh has inherited has remained, the political concept of "Basic Democracy," and the councils as such, have been abolished.

"Basic Democracy"

Social conditions. *Education.* The foundation of the educational system was laid down during the period of British rule; the system has three tiers—the primary, secondary, higher educational levels. Primary education, which was not compulsory, was for children up to about ten years old. Fewer than half of the children of primary school age attended; although tuition was free, books were not. (In 1972 it was the intention of the Bangladesh government to issue free books up to a certain level.) Secondary education was subdivided into three levels—junior secondary, high school, and higher secondary (intermediate college)—with public examinations being held at the conclusion of each level of schooling. Both primary and secondary schools varied greatly in quality in different areas; those in cities and towns were generally better staffed and financed than those in rural areas, where the quality of teaching left much to be desired.

Higher education was in two stages—college and university. There were some 250 colleges, most of them affiliated with one of three universities—the University of Dacca, the University of Rājshāhi, and the University of Chittagong. Apart from these universities, there was an agricultural university at Mymensingh, a rural academy at Comilla, and a university of engineering and technology at Dacca. There were also more than 20 technical schools, which trained high school graduates.

From 1963 onward, intensive efforts were made to promote literacy. In 1972 the literacy rate among the general population was approximately 20 percent.

The educational system was greatly disrupted by the civil war; schools and colleges ceased to function, there was a massacre of the staff and students at Dacca University, and Bengali intellectuals were generally considered a target group. In its first budget (1972–73), the Bangladesh government allocated 20 percent of budget revenue to education. Bengali, and occasionally English, are the languages of instruction; Urdu has lost the prominence it formerly enjoyed under Pakistani administration. The policies of secularism and Bengali nationalism were expected, in the early 1970s, to be reflected in changes of emphasis and curricula.

Health. The aims of the health services under the Pakistani administration were: the provision of essential health services for all, the intensification of the fight against such diseases as malaria and tuberculosis, and the improvement of sanitation and nutritional standards. There were about 8,000 doctors of all categories in the territory—a ratio of about one doctor for every 8,500 people. There was also a shortage of trained nurses. Rural health centres were established, but there was a reluctance on the part of some doctors to go to remote rural areas. Malaria constitutes the most serious threat to health. Before the civil war, tuberculosis affected about 4 percent of the population; subsequent privations, however, are known to have raised the incidence of this

Prevalent diseases

disease. There have also been periodic outbreaks of cholera, which can assume plague proportions after natural disasters, such as the cyclone and tidal wave that occurred in 1970. Both smallpox and cholera were spread in Bangladesh in 1972 by refugees returning from exile, where they had been obliged to live in squalid camp conditions. A cholera research laboratory in Dacca, however, has evolved an effective approach, which is both suitable and inexpensive, to treatment of this disease. In the early 1970s it was estimated that there were more than 65,000 cases of leprosy in the territory; measures for leprosy control, including the establishment of clinics, were undertaken.

Welfare. Under the Pakistani administration, most social services were provided by private agencies. Public projects included community development projects, schools for handicapped children, youth centres, orphanages, training institutes for social workers, and other units. In 1972 the government of Bangladesh welcomed voluntary agency assistance in its task of relief and rehabilitation after the successive disasters of the 1970 cyclone, the 1971 civil war, and the floods that occurred in 1972; more than 50 international agencies had responded to this appeal by 1972. (A.C.Cl.)

CULTURAL LIFE

Music. There are four main types of music in Bangladesh—classical, light-classical, devotional, and popular. Classical music has many forms, of which *dhrupad* —Hindustani devotional songs—and *khayal*—a blending of the Perso-Arab and Indian musical systems—are the best known. The *thumri* and *tappa* forms belong to the light-classical variety. Devotional music also is represented by *qawwāli* and *kīrtan*. These forms are part of the common musical heritage of the subcontinent. It is, however, in the field of popular music that the territory can best claim originality. The forms known as *bhatiali, bhawaiya, jari, sari, marfati,* and *baul* have no exact equivalents outside the country. While they may appear to lack the sophistication and artistry of the classical forms, they are characterized by a spontaneity and vigour wholly missing in classical music. They combine, in some respects, the vigour of jazz, the pathos of Negro spirituals, and the unpredictability of pop.

Music in Bengal made great advances during the period of the independent rulers. Sultan Ghiyāṣ ud-Dīn Aʿzam Shāh (who ruled in the late 14th and early 15th centuries) was a great patron of music, as was ʿAlāʾ-ud-Dīn Ḥusayn Shāh who ruled from 1493 to 1518. During the Mughal period, regular and constant cultural contact was maintained between northern India, the seat of classical music, and Bengal. Noted musicians accompanied the Mughal viceroys to Bengal, and classical music flourished in the region through the centuries; the *rāga-pradhan* Bengali music of modern times is an important form of classical music.

Between the rigid and formal classical music and free modern songs are found the songs called *rabindra sangeet.* They represent an experiment in mixed beats. Another significant form of modern Bengali music is the *nazrul geeti.*

Dance. Apart from such classical dances as *kathākali* and *bhārata-nāṭya*—forms that are popular all over the subcontinent—the territory has evolved highly original indigenous dances. The best known are the *dhali, baul, manipuri,* and snake dances. Each form expresses a particular aspect of tribal or communal life and is danced on particular occasions. In popular music and dancing alike, improvisation has been traditional. With the increasing commercialization of the arts, however, improvisation has been on the wane, and stereotypes have become the order of the day.

Painting. Painting in the territory is a recently introduced art form. The main figure behind the art movement was Zainul Abedin, whose sketches of the Bengal famine of 1943 first attracted attention. He was able, after 1947, to gather around him a school of artists who have experimented with various forms, both orthodox and original.

The media. *The press.* There are several daily newspapers published in Dacca and Chittagong in Bengali and English. There are also numerous Bengali and English weeklies, biweeklies, and periodicals.

Radio and television. Broadcasts are made from Dacca, Rājshāhi, Chittagong, Khulna, Sylhet, and Rangpur. Programs are presented in more than 20 languages, including English, Urdu, and Bengali. Limited television service is broadcast from Dacca.

BIBLIOGRAPHY. Comprehensive studies of the history, geography, economy, and culture of the territory are not readily available. NAFIS AHMAD, *An Economic Geography of East Pakistan,* 2nd ed. (1968), is one of the few broad surveys of the country's economic resources in relation to its geography. K. AHMED, *Agriculture in East Pakistan* (1965); and MOHAMMAD AFSARUDDIN, *Rural Life in East Pakistan* (1964), provide useful information. Demographic and agricultural statistics are available from government publications such as *East Pakistan on the March* (1963) and *Abstracts of Agricultural Statistics of East Pakistan* (annual). The history of industrialization is traced in *Some Aspects of Rural Capital Formation in East Pakistan: A Systematic Regional Geography and Its Deade of Industrial Progress* (1968), a government of Pakistan publication; and *Planning in Pakistan: A Review by a Panel of Economists* (1968). For further study, see HAROUN ER-RASHID, *East Pakistan: A Systematic Regional Geography and Its Development Planning Aspects* (1965); WARREN C. ROBINSON, *Studies in the Demography of Pakistan* (1967); and SUBHASH C. KASHYAP *et al.* (eds.), *Bangla Desh* (1971). M. RASHIDUZZAMAN in his *Pakistan: A Study of Government and Politics* (1967), analyzes current political problems. KAMRUDDIN AHMAD, *The Social History of East Pakistan* (1967), presents a historical survey of the growth and development of the area. *Folk Tales of East Pakistan* by JASIMUDDIN (1953), is a translation of some of the best known popular stories of Bangladesh. S. SAJJAD HUSAIN, *Contemporary Writing in East Pakistan* (1958), provides detailed information on modern writing in Bengali.

(S.S.H.)

Bankruptcy, Laws Concerning

Bankruptcy legislation is designed to provide an orderly and equitable liquidation of the estate of an insolvent debtor. This has always been the principal purpose of the institution of bankruptcy. Because in earlier times it was frequently coupled with the imposition of penalties and loss of civil rights upon fraudulent debtors, the designation bankrupt came to be associated with dishonesty. Eventually, however, bankruptcy legislation was extended to a broader scope to permit the rehabilitation of embarrassed estates and to provide judicial proceedings for the adjustment of debts so as to avoid bankruptcy. Modern bankruptcy laws, therefore, include various types of arrangement and reorganization proceedings to prevent the liquidation of estates that have run into financial difficulties.

In addition, bankruptcy laws of England, the United States, and some British Commonwealth countries came to include provisions for the discharge of the unpaid portion of debts incurred prior to bankruptcy in order to give honest but unfortunate debtors a new start in life. The bankruptcy laws of the European continent and of most of Latin America, by contrast, do not have such provisions. The staggering number of bankruptcies in the United States may be explained by those provisions for the discharge of hopeless debts, coupled with the possibility of voluntary proceedings by nonmerchants. Thus, the number of proceedings under the Bankruptcy Act of the United States commenced during the fiscal year of 1969 totalled 184,470. This figure included 154,054 voluntary petitions in straight bankruptcy, of which 140,530 were consumer cases filed for the purpose of obtaining a discharge, while the remaining 13,524 cases were voluntary business bankruptcy proceedings.

Since bankruptcy laws aim at the liquidation or rehabilitation of insolvent estates, bankruptcy proceedings involve all nonexempt assets of the debtor; all creditors entitled to share in the proceeds of liquidation of the estate are called upon to participate. Hence bankruptcy proceedings are general or universal collection procedures, as distinguished from individual collection remedies

available to particular creditors for the enforcement of their claims.

HISTORY OF BANKRUPTCY LAW

Early developments

Modern bankruptcy law has been formed from a number of distinct historical strands. In ancient Roman law, an unpaid judgment creditor could have his debtor's estate sequestered (*missio in bona*) and sold for the benefit of all creditors (*venditio bonorum*). Proceedings of this type caused loss of civil rights. To alleviate this hardship a debtor was given the privilege of relinquishing voluntarily his assets to his creditors by petitioning a magistrate (*cessio bonorum*).

During the Middle Ages both institutions underwent a revival and development. The Italian medieval cities enacted statutes dealing with the assets of debtors, especially merchants, who had absconded or fraudulently caused insolvency. Such "bankrupts" (*rumpentes et falliti*) were subjected to severe penalties, and their estates were liquidated. In addition, medieval Spanish law restored the judicial *cessio bonorum*. The famous Siete Partidas, a codification published by authority of Don Alfonso X the Wise, king of Castile and León, during the second half of the 13th century, contained detailed provisions relating to insolvent debtors, applicable to merchants and nonmerchants alike, enabling them to secure a voluntary liquidation of their assets under judicial supervision. An unpaid creditor could insist on either payment or assignment by the debtor of his estate to all creditors.

Laws dealing with the property of absconding and fraudulent debtors, modelled after the statutes of the medieval Italian cities, spread throughout western Europe. Provisions of this type were adopted in the commercial centres of France, Brabant, and Flanders during the 15th and 16th centuries.

The customs of Antwerp, printed in 1582, contained comprehensive rules on the treatment of bankrupts and their estates. The emperor Charles V, as count of Flanders, inserted stringent provisions for the repression of bankruptcies in his Decree for the Administration of Justice and Good Order of 1531. There can be no doubt that the first English "acte againste suche persones as doo make Bankrupte" passed in 1542/43 was inspired by the northern European models, as the very title reproduces the Flemish expression. It governed proceedings instituted against absconding or concealed debtors. It was superseded by a more detailed act of 1571 that applied only to merchants and other traders. Voluntary proceedings were not provided in England until 1844 and not in the United States until 1841.

In France, national rules on insolvency and bankruptcy were inserted into the celebrated Ordonnance du Commerce of 1673. It regulated both voluntary assignments for the benefit of creditors made by merchants (Title X) and the proceedings and effects flowing from bankruptcy (Title XI). It was interpreted to restrict bankruptcy proceedings to merchants only, and the laws of many other countries followed the French lead. Thus in Spain the limitation of bankruptcy to merchants was adopted by the famous Ordinances of Bilbao, which were sanctioned in 1737 and subsequently applied in Latin America, especially Argentina.

The restriction of bankruptcy legislation to persons engaged in commerce created a need for liquidation proceedings applicable to other debtors. As mentioned before, the Siete Partidas contained provisions for voluntary liquidation proceedings applicable to all classes of debtors. On that basis a Spanish jurist of the 17th century, Salgado de Somoza, elaborated detailed rules for the initiation and conduct of voluntary judicial liquidation proceedings, which were styles concourse of creditors. His tract, entitled *Labyrinthus Creditorum*, influenced the course of Spanish law and also had great impact on the common law of Germany. As a result, Spanish law developed two classes of liquidation proceedings, one for merchants and one for nonmerchants. A similar approach was adopted in Portugal, Argentina, and Brazil. Other countries, including Germany, Austria, the United States, and England, brought nonmerchants under their bankruptcy laws. France and Italy, however, as well as some Latin American countries, still do not provide true insolvency procedures for ordinary debtors.

Consequences of bankruptcy

The dire consequences of bankruptcy for the debtor, such as the loss and liquidation of his assets, criminal penalties, and loss of civil rights, resulted in the need for procedures avoiding such sanctions. A remedy was found in the right of a deserving debtor to reach an agreement for an extension or reduction of his debts with a majority of his creditors that was binding on dissenters. The cradle of this institution was again the statutes of the medieval cities. Provisions to that effect were also contained in the Siete Partidas mentioned before. In England, similar procedures were developed by the Privy Council through bills of conformity, but this practice ended with the abolition of the council's civil jurisdiction in 1641. In France the Ordonnance du Commerce of 1673 recognized majority compositions as a legitimate means of handling the estates of insolvents without liquidation. The Commercial Code of 1807, however, and following it the laws of other countries, restricted them to a method of terminating rather than preventing bankruptcy proceedings. Preventive compositions were resumed as legitimate means of dealing with embarrassed or insolvent estates only during the second part of the 19th century; they are now recognized in most countries as important devices for economic rehabilitation.

At one time all bankrupts were considered defrauders and criminals. They were subjected to severe social and professional sanctions, including even a degrading form of dress. In recent times, however, great efforts have been made to remove the disgrace attached to bankruptcy. Even the terms bankrupt and bankruptcy (or their equivalents in other languages) are used less and less frequently in the statutory language. Modern French legislation, for example, totally suppresses the traditional term *faillite* as the name of liquidation proceedings and restricts it to special procedures entailing the loss of civil rights by insolvents guilty of commercial misconduct.

LIQUIDATION OF INSOLVENT ESTATES

Most nations with private enterprise economies possess legislation providing for the liquidation of hopelessly insolvent estates. The socialist countries, with the exception of Yugoslavia, handle financial difficulties of their economic organizations in a different manner, aiming at rehabilitation rather than elimination. Liquidation proceedings are often referred to as "straight" bankruptcy, to distinguish them from arrangements and other rehabilitation proceedings.

The existing bankruptcy laws of the major commercial nations are of diverse vintages. The following list gives the dates of some currently operating bankruptcy laws, disregarding subsequent amendments but including the dates of the codes of civil procedure in nations such as Portugal and Spain in which the governing procedural rules are contained in these codes: Germany 1877, Spain 1881, Switzerland 1889, Austria 1914, England 1914, India 1920, Sweden 1921, Japan 1922 and 1952, Chile 1931, Peru 1932, Argentina 1933, United States 1938, Italy 1942, Mexico 1942, Brazil 1945, Canada 1949, Portugal 1961, Australia 1966, France 1967, New Zealand 1967, Colombia 1969. Some of these laws, especially those of the United States, Canada, and Mexico, were undergoing revision at the beginning of the 1970s.

Although liquidation laws differ greatly in policies and details, they all deal with six well-defined major problems: the persons subject to judicial liquidation of their estates; the persons entitled to initiate judicial liquidation proceedings and the prerequisites of an order for liquidation; the assets encompassed in liquidation proceedings and the effect of such proceedings upon prior dispositions by the bankrupt; the creditors entitled to share in distribution and the priorities among them; the role of the court, the administrator, and the creditors in the conduct of the proceedings; and the effect of liquidation on unpaid or partially unpaid debts.

Persons subject to estate liquidation. Bankruptcy or insolvency laws vary considerably in their applicability to

Differences in coverage of laws

particular classes of persons. The German act and, following its example, the Austrian and Japanese acts extend bankruptcy proceedings to all natural and juristic persons, whether or not they are engaged in commerce and without differentiating between petitions by the bankrupt himself or by creditors. In the United States and Canada the bankruptcy acts likewise have a very broad sweep. In the United States, individuals, whether merchants or nonmerchants, as well as private corporations, with the exception of certain financial institutions, are subject to the Bankruptcy Act. Involuntary petitions, however, cannot be filed against low-income wage earners, farmers, and nonprofit corporations. Canada likewise applies its act to individuals and corporations. It excludes, however, not only certain financial institutions but, in addition, nonbusiness corporations in general. In England the Bankruptcy Act covers only individuals, whether merchants or not. Registered companies are liquidated under the winding-up provisions of the Company Law. A great number of the provisions of the Bankruptcy Act, however, are made applicable in such proceedings. The same system governs in Australia, New Zealand, and India. A number of nations, following the model of the French law of 1838, extend their bankruptcy laws only to persons qualifying as merchants or engaging in trade but do not differentiate between individuals and corporations. To that class belong the bankruptcy laws of Italy (with the exception of small enterprises), Spain, Portugal, Switzerland, and a number of Latin American countries, including Argentina, Brazil, Colombia, Mexico, and Venezuela. Chile and Peru, however, chose to follow the German pattern and subjected to their bankruptcy laws all individuals and corporations, whether merchants or not. A number of the countries that restrict bankruptcy to merchants have, however, inserted provisions for insolvency proceedings governing nonmerchants in their Codes of Civil Procedure. France extends its new insolvency law to merchants and all juristic persons even if not merchants.

Persons entitled to initiate liquidation. Modern bankruptcy laws provide for the initiation of liquidation proceedings upon petition by either the bankrupt himself or his creditors. There are differences as to the number of creditors who must join in a creditor's petition. A great number of laws are satisfied with a petition by a single creditor regardless of the amount of his claim, the total number of creditors, or the amount of the outstanding indebtedness, so long as the debtor is unable to meet his current payments or has committed a so-called act of bankruptcy. Petition by a single creditor suffices according to the law of Germany and, following it, of Japan; the laws of Austria, France, Italy, Portugal, Spain and Switzerland; and the laws of such Latin American countries as Argentina, Brazil, Chile, and Mexico. In Austria at least one other creditor must exist, although he need not join in the petition; but the sufficiency of one unpaid creditor is provided expressly in the Chilean and Mexican acts.

A somewhat different regime exists in the common-law countries. In England, Canada, Australia, and New Zealand, as well as India, a single creditor may be a petitioner if the unsecured part of his claim equals or exceeds a specified amount. Otherwise, other creditors must join until the aggregate amount of their claim equals the requisite sum. In the United States in 1971, the total indebtedness of the bankrupt had to be at least $1,000, and a petition by three or more creditors with unsecured claims aggregating at least $500 was required if the total number of creditors was 12 or more; otherwise one or two creditors with claims totalling at least that amount might be petitioners.

In some countries the initiation of liquidation may also be decreed by the court ex officio or upon petition by public officials. Ex officio action by the court is provided, for example, in Italy and France. Italy, Portugal, and Mexico authorize public officials to file petitions for the liquidation of bankrupt estates. In a number of countries, following in that respect the traditional French approach, a bankrupt debtor is under a duty to file a petition, and his initiation is not left to his own judgment as in the common law countries.

Prerequisites of involuntary liquidation

Legal systems differ widely as to the substantive conditions that must be met in order to justify a decree for compulsory adjudication. In England, the United States, Canada, Australia, New Zealand, and India, a creditor's petition requires that the debtor have committed an "act of bankruptcy" or an "act of insolvency" within a specified period prior to the petition. The acts of bankruptcy are specifically defined in the applicable bankruptcy laws, their definitions and numbers varying from country to country. They include public manifestations of insolvency as well as certain conduct that endangers the collection of debts by ordinary means or entails preferential treatment of certain creditors. The laws of Germany and Italy, as well as the legislation of other countries influenced by them such as Austria and Japan, do not include a special catalog of acts of bankruptcy but condition bankruptcy upon inability of the debtor to meet his obligations or upon his state of insolvency. Cessation of payment is deemed to be an important manifestation of this condition. In France and Argentina a cessation of payments is made the grounds for the decree of liquidation, and not only an appropriate manifestation of insolvency. A number of countries have mixed systems. Liquidation is decreed if the creditor proves that the debtor has either proceeded to a cessation of payments or has committed other specified acts of bankruptcy. Systems of this type exist in Portugal, Spain, Switzerland, Brazil, and Chile. Mexico lists a number of acts that result in a rebuttable presumption of a cessation of payments. In Chile cessation in the payment of one commercial obligation justifies adjudication.

Postadjudication and preadjudication dispositions

Assets subject to distribution. One of the most important aspects of bankruptcy legislation is the determination of the assets to be seized and sold for the purpose of distributing the proceeds among the creditors as so-called dividends. Various legal systems have vastly different approaches. The disparities relate mainly to the status of assets acquired by the bankrupt subsequent to his adjudication or conveyed away by him prior to that date.

In Germany all nonexempt assets belonging to the bankrupt at the date of the adjudication form the bankrupt estate. Assets that are no longer owned by the bankrupt at the time of the adjudication are not included in the bankrupt estate unless their alienation is voidable under special rules permitting the avoidance of fraudulent or preferential transactions. The Bankruptcy Law of the United States follows a similar approach, except that the "date of cleavage" is not the date of the adjudication but is the date of the filing of the petition. Postpetition acquisitions are part of the bankrupt estate under the U.S. law only if they constitute narrowly defined "windfalls," such as inheritances or bequests vesting in the bankrupt within six months from the filing date. On the other hand, many other bankruptcy laws include within the estate subject to distribution all nonexempt assets acquired after the adjudication and during the period of the proceedings. The English Bankruptcy Act declares expressly,

> The property of the bankrupt divisible amongst his creditors . . . shall comprise . . . all such property as may belong to or be vested in the bankrupt at the commencement of the bankruptcy, or may be acquired by or devolve on him before his discharge.

Hence the bankrupt estate is formed by assets owned by the bankrupt when he commited an act of bankruptcy within the three months preceding the filing of the petition, or vesting in him after that time until his discharge. Similar rules apply in other Commonwealth countries, such as Australia, New Zealand, and Canada. In Canada, however, the so-called relation-back doctrine has been restricted and the date of cleavage is not that of the commission of the relevant act of bankruptcy but that of the filing of the involuntary or voluntary petition.

In many civil-law countries the bankrupt estate likewise includes property acquired during the course of the proceedings. Rules to this effect govern in Austria, Switzerland, France, Italy, Portugal, Spain, Argentina, and Bra-

zil. Chile, however, restricts the inclusion in the bankrupt estate of after-acquired assets to those acquired by gift, bequest, or inheritance and leaves assets purchased by the bankrupt to his own disposition.

The bankruptcy acts of England, the United States, and the Commonwealth countries vest the title to the assets forming the bankrupt estate in a trustee or assignee in bankruptcy. The bankruptcy laws of most of the other countries, however, leave the title in the bankrupt and transfer only the power of administration and disposition to a third party; as a result, the relation-back doctrine of the common-law countries, which dates the vesting of the title of the assignee to a date of cleavage prior to the adjudication, is inapposite in the countries whose bankruptcy laws do not affect the title. In some of these countries, however, parallel doctrines have been developed. The model for that approach has been the French commercial code of 1807. It required the decree of adjudication to fix the date of the cessation of payments and provided that transactions by the bankrupt after that date and before the adjudication (called the critical or suspect period) were ineffective against the estate. In 1838 France relinquished the general relation-back theory and replaced it with a catalog of different transactions, mostly various kinds of preferential transfers, that are rendered ineffective against the estate if made during the critical period. In 1967 that period was limited to a maximum of 18 months prior to the adjudication. In the early 1970s Spain seemed to be the only civil-law country that still provided for a general avoidance of acts of the bankrupt from the cessation of payments until the adjudication.

Preferential transfers

One of the cardinal principles governing the liquidation of insolvent estates is the equal treatment of creditors—the classical *par conditio creditorum*. Debtors on the eve of bankruptcy, either on their own volition or under pressure, may accord preferential treatment to certain creditors. The bankruptcy laws of all countries, therefore, contain special rules aiming at the reintegration of the insolvent estate through the avoidance of such preferences given after insolvency or cessation of payments or even earlier. These provisions are either in addition to or in lieu of a general relation-back of the date of the loss of title or power of disposition produced by the adjudication. Again, the laws of the different nations vary greatly with respect to the elements that must be present to make a transfer voidable as a preference. In the United States a transfer of property to a creditor is voidable by the trustee as a preference if it was made on account of a pre-existing debt and perfected within four months preceding the petition at a time when the bankrupt was insolvent and if, in addition, the recipient knew or should have known of the insolvency. In England a preference given by an insolvent to a creditor is subject to avoidance if it was made within six months prior to the filing of the bankruptcy petition and if it was made with a view to giving the recipient a preference. Canadian law is similar to the English in its approach, except that the critical period ordinarily covers only three months prior to the petition but is extended to 12 months if the recipient creditor is related to the bankrupt; moreover, if the transfer had a preferential effect, it is presumed that it was made with a view to giving a preference. In Australia the critical period commences six months prior to the filing of the petition. Preferential transfers by an insolvent to a creditor during this time are voidable unless the creditor had no reason to suspect that the debtor was insolvent at the time of the transfer and that the transfer gave him a preference over other creditors. In New Zealand the critical period extends to two years preceding the adjudication. Any transfer made during this time by an insolvent debtor to his creditor with a view to giving him a preference can be avoided by the trustee. In addition, all preferential transfers made by an insolvent during the month preceding his adjudication are voidable regardless of intent.

In France and a number of other civil-law countries, certain classes of preferential transfers are ineffective regardless of the intent of the debtor or the creditor's knowledge of the financial embarrassment of the debtor, provided that the transfer was made after cessation of payments or after the date when the debtor became insolvent. This critical date is usually fixed in the order of adjudication; usually, it may not antedate the adjudication or initiation of the bankruptcy proceedings by more than a fixed period. Systems of this type have been adopted by Argentina, Brazil, Chile, and Mexico. The length of the permissible period varies from country to country, being a maximum of two years in Chile but only 60 days in Brazil. Transactions considered preferential and ineffective are payments of debts made before their maturity; payments of matured debts by means other than cash, commercial paper, or securities; and grants of security interests for antecedent debts.

In Germany and in countries influenced by the German approach, preferential transfers are only voidable if the recipient knew the transfer to be preferential; in certain cases the trustee may rely on a presumption of such knowledge. Thus, in Germany, transfers granting a creditor a security or satisfaction to which he was not then entitled are voidable if made after the tenth day prior to the cessation of payments or the filing of the petition, unless the recipient is able to show that he knew neither of the cessation of payments or the petition nor of a preferential intent. If the payment or the grant of a security interest was made in performance of an existing obligation, it is voidable if its date is subsequent to the cessation of payments or the filing of a petition. In that case, however, the burden of proof of the recipient's knowledge of these facts is on the trustee. Voidability on the ground of the creditor's knowledge of the cessation of payment cannot be claimed if the transfer was perfected at a date preceding the filing of a petition by more than six months. The same system governs in Japan except that the ten-day period in cases of prepayment or newly granted security is extended to 30 days and the special limitation period of six months is lengthened to one year. The Portuguese bankruptcy law likewise contains a catalog of voidable preferences differentiating the length of the critical period preceding adjudication according to various classes of preferential transactions; voidability is aided by a rebuttable presumption of a preferential intent. Comparable provisions govern in Austria and Italy. Switzerland authorizes the avoidance of payments of debts prior to their maturity or of grants of security interests for pre-existing debts if such preferences were given when the bankrupt was insolvent, unless the creditor can show that he did not know of that circumstance. In addition, in Switzerland all transfers made with preferential intent known to the creditor are voidable within five years.

Distribution of proceeds

Claims to the estate. One of the principal objectives of liquidation is the distribution of the proceeds of the estate among the creditors. The designation of the classes of claims entitled to share in such distribution and the regulation of the procedures for their establishment form an important ingredient of modern bankruptcy legislation. Entitlement to "dividends" by reason of a "provable claim" must be distinguished from the right by means of "reclamation proceedings" and from the right to regain specific property to recover the proceeds from certain assets by a lien claimant. Except for the rules governing fraudulent or preferential transfers, bankruptcy does not affect existing consensual or statutory security interests although their enforcement may come within the jurisdiction of the bankruptcy court. In a number of countries, however, for example in Italy, Argentina, and Brazil, creditors holding security interests in assets belonging to the estate must file proof of the claims so secured in order to assert their rights to the proceeds from the collateral.

Provable claims are held by creditors of the bankrupt. Creditors who have become such as a result of acts in, or expenditures for, the administration of the estate are not creditors of the bankrupt and are entitled to payment ahead of such creditors. Provable claims result from transactions or events that took place prior to a specific phase of the bankruptcy proceedings. The controlling date in the United States and Canada is usually the time

of the filing of the petition. In England the dividing date is that of the issuance of the receiving order that precedes the formal adjudication, while in Australia it is that of the sequestration order or of a voluntary petition that renders the debtor a bankrupt. In New Zealand the crucial date is that of the adjudication, a rule also governing in continental European and South American countries and in countries influenced by continental European laws, such as Austria, France, Germany, Italy, Portugal, Spain, Argentina, Brazil, Chile, and Japan.

Under most bankruptcy laws provable claims include all classes of pre-bankruptcy obligations, whether matured or unmatured, liquidated or unliquidated, unconditional or contingent. A provable unliquidated liability may be based not only on contract but also on the infliction of personal injury or damage to property. Systems providing for such a broad spectrum of provability exist in most of the civil-law countries and also in Canada and New Zealand. In England and Australia, however, claims for unliquidated damages arising otherwise than by reason of contract are not provable. Hence personal injury claims, not reduced to judgment or settled by compromise, are not provable. In the United States a similar exclusion exists, except that claims for the negligent infliction of personal injuries are provable if such an action was pending at the time of the petition. In Brazil the view is held that a creditor with an unliquidated claim cannot prove it in bankruptcy proceedings but may only demand the setting apart of a reserve.

Although in principle bankruptcy laws aim at equality among creditors, priorities traditionally are given to certain creditors. Generally speaking, such priorities exist for taxes or other debts owed to the state and certain public entities and for wage claims and other private liabilities deserving preferential treatment. In some civil-law countries the system of priorities is full of complexities. In Spain, for example, proceeds from personal property are distributed according to priority rules that differ substantially from those governing proceeds from real property.

Creditors not entitled to priorities are designated as general creditors. Secured creditors whose claims are not covered by their security are entitled to dividends for the unsecured portion of their debt.

Release of debts

Discharge. An English statute of 1705 provided for a discharge (release) of all debts owed by the bankrupt and due at the time of bankruptcy provided the bankrupt had faithfully complied with his statutory duties. Since that time, relief of the honest but unfortunate bankrupt from his provable debts has become one of the main objectives of the bankruptcy laws of the English-speaking countries. The right to a discharge, however, is not unqualified. A bankrupt may forfeit his right to a discharge by certain reprehensible conduct or by neglect of special duties. Moreover, certain debts are excepted from the operation of a discharge, as, for example, liabilities for support or debts resulting from fraudulent dealings.

In England, Canada, Australia, and New Zealand, the court has the power to grant suspended or conditional discharges. In the first three countries the court may not grant an immediate and absolute discharge if the bankrupt has been guilty of certain acts listed in a statutory catalogue. In New Zealand the discretion of the court in deciding the character of the discharge is not restricted. Conversely, in the United States, conditional or suspended discharges are not permitted, and the grounds barring a discharge are absolute. Japan introduced discharge provisions modelled after the law of the United States in 1952. Most civil-law countries do not provide for discharges in liquidation proceedings and leave the discharge of debts to subsequent agreement among the parties. An exception is the bankruptcy law of Brazil. In that country payment of dividends in excess of 40 percent of the provable debts entails the release of the balance; the unpaid balance is extinguished after five years from the closure of the proceedings, unless the bankrupt was convicted of a bankruptcy crime.

Conduct of bankruptcy proceedings. Since bankruptcy aims at a judicially supervised liquidation of the insolvent estate, the proceedings must be initiated in a court, which in some measure controls the proceedings. The bankruptcy laws of the various countries, however, vary considerably with respect to (1) the jurisdiction of the bankruptcy court and (2) the relative roles assigned to the judicial officers, the creditors, and others participating in the liquidation.

The "force of attraction" theory

Jurisdiction of the bankruptcy court. During the formative period of bankruptcy law in Europe, the courts developed a theory of the "force of attraction," in the sense that all litigation involving the creditors or assets was to be concentrated in the bankruptcy court. Modern bankruptcy laws are still under the influence of this idea, but the jurisdiction of the bankruptcy court differs greatly from country to country. The statutes of England, the United States, Australia, Canada, and New Zealand contain very broad grants of jurisdiction to the bankruptcy courts; extensive jurisdictional powers of the bankruptcy courts exist likewise under the acts of a number of civil-law countries, as for example in Italy and Argentina. Conversely, the bankruptcy law of Germany and to a lesser degree the bankruptcy legislation of countries influenced by German law, such as Austria, have left the settlement of individual controversies to the ordinary courts—*i.e.*, in such matters as the rights of the estate to certain assets or the determination of contested provable debts.

In many countries a vast array of functions of the bankruptcy court are conferred upon special judicial officers who may be either actual members of the judiciary (as in France) or judicial officers without full judicial status. In England, Registrars in Bankruptcy have jurisdiction over the initiation of bankruptcy proceedings and a long catalog of matters not requiring hearings in open court. Similar rules apply in Canada and New Zealand. In the United States the Referee in Bankruptcy may exercise practically all judicial functions. In Australia, for constitutional reasons, the law authorizes only the delegation of functions of an administrative nature. In Germany, conversely, a wide range of judicial bankruptcy functions are delegable to judicial officers outside the regular judiciary, called *Rechtspfleger*.

Growing powers of the courts

The roles of the court, the creditors, and others. Creditors have traditionally played an important part in the conduct of bankruptcy proceedings. In most countries, however, they are no longer the dominant parties; the courts, assisted by official administrators, have emerged as key figures. In England, creditors gained an active role with the legislation of 1706; they have been vested with important powers except for an interval from 1831 to 1869, when a system of "officialism" governed. It was followed by a system of creditor autonomy. In 1883, however, shared responsibility was instituted. Creditors retained the power of selecting the trustee, who was subjected to supervision by a creditors' committee as well as by the Board of Trade. English law also provides for official receivers who act as trustees until one is appointed either by the creditors or, in default of such action, by the Board of Trade. In Canada the position of judicial and administrative officers in liquidation proceedings is even stronger. There the trustee is appointed initially either in the court's receiving order, issued upon an involuntary petition, or by an official receiver in the case of a voluntary assignment. The creditors, however, may replace such a trustee by one of their choice. Official receivers in Canada are special officers of the court, appointed by the governor general, who are charged with a general supervision of bankruptcies in their districts as well as the examination of bankrupts in individual cases. Creditors exercise supervisory powers over the actions of the trustees. Similarly, in Australia, official receivers serve as initial trustees but may be replaced by registered trustees selected by the creditors. Under the New Zealand Insolvency Act of 1967, the administration of bankrupts' estates is in the hands of official assignees appointed under the State Services Act. The creditors may appoint experts or committees to assist in the administration. In the United States creditors have preserved a greater power over the conduct of the proceedings, al-

though the bankruptcy judge is in control of them. No official trustee or administrative control is provided. In France and Italy the creditors are almost totally deprived of any voice in the proceedings; creditors' meetings in liquidations have been abolished in France and are convoked in Italy only for the establishment of provable debts. In both countries creditors are represented through a creditors' committee selected by the judge. Its establishment is mandatory in Italy but only optional in France. In Chile and Switzerland the actual liquidation is entrusted to administrative officials. In Chile, bankrupt estates are liquidated by official trustees, constituting the staff of a central governmental agency called Sindicatura General de Quiebras. In Switzerland, bankruptcy offices are established for each of the bankruptcy districts created by the individual cantons. The bankruptcy office is in charge of the progress of the proceedings and acts as trustee, unless the creditors select one of their choice.

REHABILITATION OF INSOLVENT ESTATES

Liquidation is a wasteful and socially undesirable way of disposing of insolvent estates. For that reason most bankruptcy laws establish procedures aiming at the rehabilitation of the estate by means of an arrangement with a qualified majority of the creditors. Such an arrangement, if it meets statutory requirements and has the approval of the court, becomes binding upon dissenters. An arrangement may provide merely for a deferment of payments (extension agreement) or for a scaling down of the indebtedness (composition). During the 19th century the bankruptcy laws of many countries recognized such arrangements only as a method of terminating pending bankruptcies. In more recent statutes, however, arrangement proceedings have been established for the avoidance of liquidation proceedings. In England, after an experiment with preventive compositions outside bankruptcy, composition proceedings were permitted by the Bankruptcy acts of 1883 and 1914 only after a petition in bankruptcy had been filed either before or after the adjudication. In France, likewise, arrangement proceedings constitute a mandatory phase after the filing of a petition and before adjudication. In the United States, preventive compositions prior to the initiation of liquidation proceedings were introduced in 1933 and greatly expanded by the Chandler Act of 1938. Canada reintroduced preventive "proposals" in 1949, and similar provisions have been adopted in New Zealand. Systems of preventive arrangements exist also, for example, in Austria, Germany, Italy, Portugal, Argentina, Brazil, Chile, Mexico, Japan, and Spain. Australia authorizes preventive compositions among other preventive arrangements. Usually compositions affect only the rights of the unsecured creditors. In the case of corporate enterprises, however, effective rehabilitation frequently requires that the whole debt structure be reconstituted; in some countries, therefore, including the United States, Canada, and Japan, special provisions for corporate reorganization affecting also secured creditors have been adopted.

Avoidance of liquidation proceedings

The United States has also enacted special provisions permitting wage earners to work out plans for a reduction of their indebtedness and payment of their creditors from their earnings. Canada authorizes nonbusiness insolvents to arrange for the orderly payment of their debts by means of consolidation orders. Proceedings of this kind, however, do not afford protection against an involuntary bankruptcy petition.

INTERNATIONAL ASPECTS

Perhaps the majority of countries extend their bankruptcy proceedings to all foreign assets of a resident debtor without differentiating between personal and real property. This is done, for example, in England, the United States, and Switzerland. Japan, on the other hand, applies its law solely to property located in that country. Only a few nations, among them the United States and Austria, provide for adjudication of nonresidents merely because they have assets in the national territory. Germany follows a similar rule with respect to nonresidents who maintain a business branch or operate a farm in that country but restricts the proceedings to local assets. A number of countries expressly refuse to recognize foreign adjudications and leave the domestic assets to individual executions; these include Argentina, Germany, Japan, and Portugal. Other countries permit foreign trustees to reach property in their territory, especially if the foreign adjudication has been recognized by special procedures to that effect.

Normally countries will not differentiate between foreign and domestic creditors in proceedings involving residents, at least if reciprocity exists. Provisions to that effect exist in the laws of Germany and Japan and by implication in Italy. A number of countries, however, especially in Latin America, give priority to local creditors if there are concurring bankruptcies.

Even greater complexities exist with respect to the extraterritorial effects of releases flowing from compositions or discharges in bankruptcy.

Because of the above difficulties, some groups of countries have regulated the subject among themselves by multipartite treaty, in particular the five Scandivanian countries (by a treaty of November 7, 1933) and 15 Latin American countries (by a treaty of February 20, 1928; the so-called Bustamante Code of Private International Law). The Common Market countries had by 1971 prepared a draft convention governing liquidation, composition, and extension proceedings in their respective territories, accompanied by the draft of a uniform law governing fraudulent and preferential transfers and certain other matters.

BIBLIOGRAPHY. While the general aims and features of the bankruptcy law of most countries are similar, the substantive and procedural rules vary greatly. The following works are commentaries on, or proposals for, reform of the laws of particular countries: *Argentina:* FRANCISCO GARCIA MARTINEZ, *El concordato y la quiebra*, 4th ed., 3 vol. (1962–63). *Australia:* ARNDELL LEWIS and DENNIS ROSE, *Lewis' Australian Bankruptcy Law*, 5th ed. (1967). *Austria:* HANS SABADITSCH (ed.), *Die Konkurs-, Ausgleichs- und Anfechtungsordnung*, 5th ed. (1970). *Brazil:* WALTER T. ALVARES, *Direito Falimentar* (1966); J.C. SAMPAIO DE LACERDA, *Manual de Direito Falimentar*, 4th ed. (1967). *Canada:* LEWIS DUNCAN and JOHN D. HONSBERGER, *Bankruptcy in Canada*, 2nd ed. (1961); *Report of the Study Committee on Bankruptcy and Insolvency Legislation* (1970). *Chile:* ARTURO DAVIS, *Código de comercio: origenes concordancias, jurisprudencia*, vol. 4, *Ley de quiebras* (1966). *England:* R.V. WILLIAMS, *William's Law and Practice in Bankruptcy*, 18th ed. (1968). *France,* GILBERT BORD, *Règlement Judiciaire et liquidation des bien* (1969). *Germany:* ERNST JAEGER, *Konkursordnung, mit Einführungsgesetzen*, 2 vol., 8th ed. (1958–70). *India:* RAMESHWAR DIAL, *The Provincial Insolvency Act*, 3rd ed. (1969). *Italy:* FRANCESCO FERRARA, *II Fallimento*, 2nd ed. (1966); RENZO PROVINCIALI, *Manuale di Diritto Fallimentare*, 5th ed. (1969). *Japan:* M. HASEBE, "Problems in Implementing the Corporate Reorganization Law," *Law in Japan*, 21:164 (1968). *New Zealand*, E.H. FLITTON, "The Insolvency Act, 1967," *New Zealand Journal* 394 (1968). *Portugal:* JOAO FARINHA, *Código de Processo Civil Anotado*, 2 vol. (1965); PEDRO DE SOUSA MACEDO, *Manual de Direito das Falências*, 2 vol. (1964–68). *Spain:* L. PRIETO-CASTRO FERRANDIZ, *Derecho Procesal Civil*, 2nd ed. (1968–69); RODRIGO URIA, *Derecho Mercantil*, 6th ed. (1968); JOSE A. RAMIREZ, *Derecho Concursal Español, La Quiebra*, 3 vol. (1959). *Switzerland:* C. JAEGER and M. DAENIKER, *Schuldbetreibung und Konkurs*, 8th ed. (1967). *United States:* WILLIAM COLLIER, *Bankruptcy*, 14th ed. (1971).

The best studies in English of the various international aspects are by K.H. NADELMANN, "Foreign and Domestic Creditors in Bankruptcy Proceedings, Remnants of Discrimination?" *University of Pennsylvania Law Review*, 91:601 (1943), "International Bankruptcy Law: Its Present Status," *University of Toronto Law Journal*, 5:324 (1944), and "Assumption of Bankruptcy Jurisdiction over Non-Residents," *Tulane Law Review*, 41:75 (1966).

(S.A.Ri.)

Banks and Banking

The principal types of banking in the modern industrial world are commercial banking and central banking. A commercial banker is a dealer in money and in substitutes for money, such as checks or bills of exchange. He also provides a variety of financial services. The basis of

his business is borrowing from individuals, firms, and occasionally governments—*i.e.*, receiving "deposits" from them. With these resources and also with his own capital, the banker makes loans or extends credit and also invests in securities. The banker makes his profit by borrowing at one rate of interest and lending at a higher rate and by charging commissions for services rendered.

A bank must always have cash balances on hand in order to pay its depositors upon demand or when the amounts credited to them become due. It must also keep a proportion of its assets in forms that can readily be converted into cash. Only in this way can confidence in the banking system be maintained. Provided it honours its promises (*e.g.*, to convert notes into gold or provide cash in exchange for deposit balances), a bank can create credit for use by its customers by issuing additional notes or by making new loans, which in their turn become new deposits. The amount of credit it extends may considerably exceed the sums available to it in cash. But a bank will only be able to do this as long as the public believes the bank can and will honour its obligations, which are then accepted at face value and circulate as money. So long as they remain outstanding, these promises or obligations constitute claims against that bank and can be transferred by means of checks or other negotiable instruments from one party to another. These are the essentials of deposit banking as practiced throughout the world today, with the partial exception of Soviet-type institutions.

Another type of banking is carried on by central banks. Central banks are bankers to governments and "lenders of last resort" to commercial banks and/or other financial institutions. They are often responsible for formulating and implementing their country's monetary and credit policies, usually in cooperation with the government. In some cases—*e.g.*, the U.S. Federal Reserve System—they have been established specifically to lead or regulate the banking system; in other cases—*e.g.*, the Bank of England—they have come to perform these functions through a process of evolution.

Some institutions often called banks, such as finance companies, savings banks, investment banks, trust companies, and home-loan banks, do not perform the banking functions described above and are best classified as financial intermediaries. Their economic function is that of channeling savings from private individuals into the hands of those who will use them, in the form of loans for building purposes or for the purchase of capital assets. These financial intermediaries cannot, however, create money (*i.e.*, credit) as the commercial banks do; they can lend no more than savers place with them.

The article is divided into the following sections:

I. The development of banking systems
II. The structure of modern banking systems
  Unit banking: the U.S.
  Branch banking: Great Britain
  Hybrid systems
  Banking in planned economies
III. The business of banking
  Functions of commercial banks
  Industrial finance
IV. The principles of central banking
  Responsibilities of central banks
  Techniques of credit control

I. The development of banking systems

Money changers and merchant bankers

Banking is of ancient origin, though little is known about it prior to the 13th century. Many of the early "banks" dealt primarily in coin and bullion, much of their business being money changing and the supplying of foreign and domestic coin of the correct weight and fineness. Another important early group of banking institutions was the merchant bankers, who dealt both in goods and in bills of exchange, providing for the remittance of money and payment of accounts at a distance but without shipping actual coin. Their business arose from the fact that many of these merchants traded internationally and held assets at different points along the trading routes. For a certain consideration, a merchant stood prepared to accept instructions to pay money to a named party through one of his agents elsewhere; the amount of the bill of exchange would be debited by his agent to the account of the merchant banker, who would also hope to make an additional profit from exchanging one currency against another. Because there was a possibility of loss, any profit or gain was not subject to the medieval ban on usury. There were, moreover, techniques for concealing a loan by making foreign exchange available at a distance but deferring payment for it to a later date so that the interest charge could be camouflaged as a fluctuation in the rate of exchange.

Another form of early banking activity was the acceptance of deposits. These might derive from the deposit of money or valuables for safekeeping or for purposes of transfer to another party; or, more straightforwardly, they might represent the deposit of money in current account. A balance in current account could also represent the proceeds of a loan that had been granted by the banker, perhaps based on an oral agreement between the parties (recorded in the banker's journal) whereby the customer would be allowed to overdraw his account.

English bankers in particular had by the 17th century begun to develop a deposit banking business, and the techniques they evolved were to prove influential elsewhere. The London goldsmiths kept money and valuables in safe custody for their customers. In addition, they dealt in bullion and foreign exchange, acquiring and sorting coin for profit. As a means of attracting coin for sorting, they were prepared to pay a rate of interest, and it was largely in this way that they began to supplant as deposit bankers their great rivals the "money scriveners." The latter were notaries who had come to specialize in bringing together borrowers and lenders; they also accepted deposits.

It was found that, when money was deposited by a number of people with a goldsmith or a scrivener, a fund of deposits came to be maintained at a fairly steady level; over a period of time, deposits and withdrawals tended to balance. In any event, customers preferred to leave their surplus money with the goldsmith, keeping only enough for their everyday needs. The result was a fund of idle cash that could be lent out at interest to other parties.

About the same time, a practice grew up whereby a customer could arrange for the transfer of part of his credit balance to another party by addressing an order to the banker. This was the origin of the modern check. It was only a short step from making a loan in specie or coin to allowing customers to borrow by check: a loan account would be debited with the full amount borrowed and the amount credited to a current account against which checks could be drawn, or the customer would be allowed to overdraw his account up to a specified limit. In the first case, interest was charged on the full amount of the debit, and in the second the customer paid interest only on the amount actually borrowed. A check was a claim against the bank, which had a corresponding claim against its customer.

Origin of the check

Another way in which a bank could create claims against itself was by issuing bank notes. The amount actually issued depended on the banker's judgment of the possible demand for specie, and this depended in large part on public confidence in the bank itself. In London, goldsmith bankers were probably developing the use of the bank note about the same time as that of the check. (The first bank notes issued in Europe were by the Bank of Stockholm in 1661.) Some commercial banks are still permitted to issue their own notes, but in most countries this has become a prerogative of the central bank.

Bank money or credit

In Britain the check soon proved to be such a convenient means of payment that the public began to use checks for the larger part of their monetary transactions, reserving coin (and, later, notes) for small payments. As a result, banks began to grant their borrowers the right to draw checks much in excess of the amounts of cash actually held, in this way "creating money"—*i.e.*, claims that were generally accepted as means of payment. Such money came to be known as "bank money" or "credit."

Excluding bank notes, this money consisted of no more than figures in bank ledgers; it was acceptable because of the public's confidence in the ability of the bank to honour its liabilities when called upon to do so.

When a check is drawn and passes into the hands of another party in payment for goods or services, it is usually paid into another bank account. Assuming that the overdraft technique is employed, if the check has been drawn by a borrower, the mere act of drawing and passing the check will create a loan as soon as the check is paid by the borrower's banker. Since every loan so made tends to return to the banking system as a deposit, deposits will tend to increase for the system as a whole to about the same extent as loans. On the other hand, if the money lent has been debited to a loan account and the amount of the loan has been credited to the customer's current account, a deposit will have been created immediately.

One of the most important factors in the development of banking in England was the early legal recognition of the negotiability of credit instruments or bills of exchange. The check was expressly defined as a bill of exchange and thus recognized as a negotiable instrument. In continental Europe, on the other hand, limitations on the negotiability of an order of payment prevented the extension of deposit banking based on the check. Continental countries developed their own system, known as Giro payments, whereby transfers were effected on the basis of written instructions to debit the account of the payer and to credit that of the payee.

II. The structure of modern banking systems

The banking systems of the world have many similarities, but they also differ, sometimes in quite material respects. The principal differences are in the details of organization and technique. The differences are gradually becoming less pronounced because of the growing efficiency of international communication and the tendency in each country to emulate practices that have been successful elsewhere.

Unit banking and branch banking

Banking systems may be classified in terms of their structure as unit banking, branch banking, or hybrids of the two. For example, unit banking prevails in large areas of the United States, which in 1971 had nearly 14,300 banks but only about 24,600 bank branches (in relation to a population of 205,000,000). In other countries it is more usual to find a small number of large commercial banks, each operating a highly developed network of branches. In England and Wales, for example, there were in 1971 only six clearing banks (four of which are very much larger than the other two), which carried on nearly all the domestic banking business through more than 14,300 branch offices (in relation to a population of 49,000,000). Examples of hybrid systems include those of France, West Germany, and India, where banks that are national in scope are supplemented by regional or local banks. Some of these hybrid systems are slowly changing their character, the banks becoming fewer in number and individually larger, with a larger number of branches.

UNIT BANKING: THE U.S.

Bank organization in the United States during the years after World War II was still passing through a phase of structural development that many other countries had completed some decades earlier. Development in the U.S. has been subject to constraints not found elsewhere. The federal Constitution permits both the national and state governments to regulate banking. Some states prohibit branch banking, largely because of the political influence of small local bankers, thus encouraging the establishment and retention of a large number of unit banks. (In Australia, which also has a federal constitution, the national government has more power to regulate banking generally; consequently, there have been no legal impediments to the spread of branch banking, and the system has for a number of years been well developed.)

Even in its early years the United States had an unusually large number of banks. As the frontiers of settlement were pushed rapidly westward, banks sprang up across the country. One reason for this was the demand for capital in the expanding frontier economy. There was also an obvious need for a large number of banks to serve the diverse and rapidly expanding demands of a growing and constantly migrating population. It must be remembered, too, that at this time communications between the frontiers of settlement and the established centres of commerce and finance were still inadequately developed.

As long as communications remained imperfect, the existence of large numbers of competing institutions is not difficult to explain. The subsequent failure of bank mergers or amalgamations to produce a concentration of financial resources in the hands of large banking units can be attributed in part to the character of the federal Constitution as noted above. Among the people, moreover, there was a widespread distrust of monopoly and a deep-rooted fear that a "money trust" might develop. This went hand in hand with a political philosophy that emphasized the virtues of individualism and free competition; restrictions on branching, merging, and on the formation of holding companies were a feature of both the state and the federal banking laws. Where permitted, however, bank branches are numerous in the U.S. (especially in California, where branching is statewide, and in New York); in states in which branching is prohibited, one often finds local bank monopolies in small towns. The banking system of the United States would not work without a network of correspondent bank relationships, which are more highly developed there than in any other country.

The U.S. concern for competition

BRANCH BANKING: GREAT BRITAIN

If the United States banks can be taken as representative of a unit banking system, the British system is the prototype of branch banking. Its development was linked to the growth of transportation and communications, for otherwise banks cannot clear checks drawn on other banks and effect remittances speedily and efficiently. The Scots favoured branch banking from the very beginning (the Bank of Scotland was founded in 1695), but at first they were not very successful—largely because of poor communications and the difficulty of supplying branches with adequate amounts of coin. Not until after the Napoleonic Wars, when the road system of Scotland had been greatly improved, did branch banking begin to develop vigorously in Scotland. As the Industrial Revolution progressed and as the size of businesses increased, the structure of English banking underwent a corresponding change. Greater resources were required for lending, and banks also needed more extensive interconnections in order to provide an increasing range of services. Where banks remained small, they were frequently unable to take the strain of the larger demand; they tended to become overextended and often failed.

The growth in size of banks was also greatly encouraged by legislation that encouraged joint-stock ownership, beginning in 1826. Joint-stock ownership, which reduced the risk to any particular individual, must be distinguished from limited liability, which did not become widely accepted until the failure of the City of Glasgow Bank in 1878 demonstrated the need for a legal device that would protect the stockholder. The early joint-stock banks tended to remain somewhat localized in their business interests, and it was only gradually (along with the spread of limited liability and disclosure of accounts) that amalgamations began to convert the banking system in England and Wales into its highly concentrated modern form. The main movement was completed before World War I, though there was to be a further degree of concentration in the years after World War II. By these means, the British banks were able to attract deposits from all parts of the country and also to spread the banking risk over a wide range of industries and areas.

HYBRID SYSTEMS

A third group of banking systems differs from the unit-banking system of the United States and also from the

branch-banking systems of countries that have followed the British model (such as Australia, Canada, New Zealand, and South Africa). This group is characterized by the existence of a small number of banks with branches throughout the country, holding a significant part of total deposits, along with a relatively large number of smaller banks that are regional or local in emphasis. Such systems exist in France, West Germany, Italy, and India. Japan has a small number of large city banks with branch networks but a larger number of local banks.

France. Banking institutions in France were classified after World War II into three main groups: deposit banks, *banques d'affaires* (or investment banks), and institutions that were either specialized or operated mainly outside France. New banking legislation in 1966 greatly reduced the importance of the distinction between deposit banks and *banques d'affaires;* the main distinction is now between commercial banks on the one hand and a range of other banking and financial institutions on the other. Along with this there has been (1) a further concentration of banking resources, as a result of several large mergers and also of greater financial integration through share-exchange agreements and interlocking directorates, and (2) the conversion of a number of *banques d'affaires* into deposit banks, which have hived off their investment interests into separate investment or holding companies.

Integration in France

The main reason for this development was the relative superfluity of deposit bank funds, along with a scarcity of funds at the disposal of the *banques d'affaires*. There had always been a bigger demand for capital than the *banques d'affaires* themselves were able to provide, but, so long as a high proportion of medium-term commercial notes were rediscountable at semipublic institutions or at the Bank of France, the *banques d'affaires* could turn over their funds fairly frequently by rediscounting; this was now no longer possible on the same scale. In consequence, the *banques d'affaires* had to seek deposits more actively. Meanwhile, the deposit banks had begun to attract an increasing proportion of their funds from private individuals (who were making much less use of notes), and these banks sought to hold such business by expanding their retail banking facilities (*e.g.*, personal loans and mortgage lending). But they were also permitted to carry more time deposits, and these funds are now used to finance their medium-term lending; hence, they have less need to rediscount.

In 1970 there were about 330 banks registered in France, a decrease of more than 200 since 1945. The thinning-out process affected primarily the deposit banks, which in 1970 consisted of the three nationalized banks (with branches throughout the provinces and accounting for well over one-half of total deposits), four relatively large non-nationalized banks (one of them strictly a Parisian bank that heads a group of regional banks), about 20 regional banks (very few of which are now independent of Paris, one of them being the sixth-largest deposit bank in the country), 75 Parisian deposit banks, and 70 local banks operating in departmental centres. Most of the reduction in numbers was among the local banks, and the rest was among the Paris deposit banks. The number of regional banks remained virtually unchanged.

All of the regional banks and some local banks have branches. The balanced character of the regional economies often provides these banks with a good portfolio of risks; they serve not only a prosperous agriculture but a number of local industries. Some of the local banks are also very sound institutions, despite their small size.

The survival of a hybrid system in France, despite the long-run trend toward centralization, reflects certain characteristics of French society. These included, until recently, a strong emphasis on small business, together with a preference for the individual and personal service that only the small bank could provide. Particularism in some parts of France manifests itself in support for local institutions, and the local banker also often has an advantage because of his special knowledge of local industries and people, which enables him to accept risks that the big banks decline. Until recent years, moreover, the large French banks preferred to open branches in new areas rather than to absorb local and regional banks.

West Germany. An even more direct conflict between the forces favouring concentration and those working against it may be seen in the Federal Republic of Germany. Before 1848, when the first joint-stock bank was established, there was little industrial development. Banking grew in the latter part of the 19th century along with industry; a number of the banks established after 1848 were from the beginning specifically linked with the promotion of industrial enterprises. For this reason, the banks were inclined to rely mainly on their own capital resources and did not at first try to attract deposits from the public. Not until 1874 did the Deutsche Bank A.G. begin to seek deposits through offices specially opened for the purpose. This was done to provide cheap finance for traders, the deposits being invested in mercantile bills that were regarded as both safe and liquid.

This was an important period in the evolution of German banking, when the principles of deposit banking on the British model were combined with long-term financing of the kind done by the French *banques d'affaires.* In pursuit of deposits, the banks built up a widespread network of branch offices, which were also used to establish and maintain industrial contacts throughout the country. The unification of Germany in 1871 removed the political obstacles to a more integrated banking system, and the selection of Berlin as the capital made that city the country's financial centre. Four of the largest banks were already established there; the new Reichsbank was set up in 1876. In addition, the larger and more enterprising of the provincial banks were now attracted to the capital. The Berlin stock exchange rapidly displaced Frankfurt am Main as the country's leading securities market.

The Berlin banks

The Berlin banks extended their influence throughout the provinces. They did this initially by developing correspondent relationships and subsequently by acquiring a financial interest in the provincial banks and being represented on their boards. Sometimes a provincial bank might acquire an interest in a Berlin bank, but almost invariably it was the provincial bank that lost some of its independence. Each of the big Berlin banks came to be associated with a group of provincial banks more or less under its control. At the same time, all of the banks, Berlin and provincial alike, expanded their business by opening branches, usually by absorbing a private banking business or by buying up the assets of a provincial bank that had been forced into liquidation.

During World War I the degree of centralization increased; by 1918, the big Berlin banks held over 65 percent of total deposits. In the early 1920s there were amalgamations, and branch systems became much larger. Bank failures and the financial crisis of 1931 resulted in further consolidation until the German banking system was dominated by three giants. Under the Nazis, it became virtually an instrument of the state.

But there were countervailing forces. Probably the most important of these was the establishment of publicly owned banking institutions, such as the communal savings banks, which became of increasing importance after World War II. In 1970 the three largest banks held only about one-fifth of total demand deposits (repayable on sight or demand) and about 10 percent of savings bank deposits. The other big commercial banking group included the regional banks, holding less than 15 percent of sight deposits. The commercial banks also accepted time deposits (repayable after a period of time). But sight and time deposits together were much less than savings deposits; indeed, the savings banks (some of which were bigger than certain of the regional banks) constituted the chief competition of the commercial banks. They attracted an even larger volume of sight deposits than did the commercial banks—over 35 percent of the total—and almost 60 percent of the much larger sum of savings deposits. On the lending side, also, the competition was quite significant; in the long-term lending field (in which the big banks were small), the savings banks extended about one-quarter of the total amount loaned, and their

central institutions, the *Girozentralen*, extended a further one-sixth. Even the specialist long-term lending of the private and public mortgage banks amounted to something less than 30 percent of the total.

The savings banks offer a wide range of services, especially to lower income groups and smaller businesses. The large commercial banks have concerned themselves more with big business and with wealthy individuals. The savings banks now compete in wholesale banking as well. A number of them, together with their *Girozentralen*, are to all intents and purposes "universal banks," like the Big Three and the larger regional banks. The Big Three remain unchallenged only in stock exchange and foreign banking business.

There were also about 170 private bankers in 1970, though only about half a dozen were of any size. The bigger private banks are important in the fields of investment and wholesale banking, while the smaller ones flourish in the leading stock exchange cities, such as Düsseldorf and Frankfurt am Main. Many of these private bankers, however, are not bankers in the true sense; they subsist mainly on stock transactions, investment services, portfolio management, and insurance and mortgage brokerage. There are also consumer finance institutions, mortgage and other specialist banks, and a large number of cooperatives.

The German banking system is a reflection of the decentralized nature of the country, with its numerous provincial cities. Economic and cultural life is dispersed. Some strong banks thrive upon regional sentiment. The belief in economic competition also remains very strong and is encouraged. At the same time, the influence of the Big Three is more important than the foregoing figures indicate. Regional and private banks are often within the sphere of influence of the Big Three. In some cases the latter have a financial interest in these banks, and in some cases they own them. The Big Three also have shares in certain of the private mortgage banks. There are also "cooperation agreements," and a number of mergers have taken place. In these several ways, much more integration exists than appears on the surface. While banking in West Germany remains a hybrid system, a trend toward greater concentration is evident.

Italy. The large Italian banks began, as in a number of other European countries, by operating as "mixed banks." After the crisis of the early 1930s, however, short-term banking operations were separated from the various types of medium- and long-term banking by the Banking Law of 1936, which reserved each kind of business specifically for a separate type of organization. Italian arrangements thus contrast strongly with the West German concept of a "universal" bank, the tendency in Italy being to specialize. The specialization has been somewhat blurred in recent years as credit terms have lengthened and as the ordinary banks have expanded their security portfolios.

Wholesale and retail banking in Italy

The big banks have tended to function as wholesalers of credit, operating on a national scale, and to do business mostly with the large and medium-sized firms. They have had branches mainly in the capital towns of the provinces and in busy centres of industry and commerce. The medium-sized and smaller banks have traditionally catered to medium-sized and smaller customers. As indicated, there have been changes in recent years. The big banks, stimulated partly by the growth in middle-range personal income, have begun to develop a retail banking business, especially in the field of personal loans; while certain of the medium-sized banks, as well as some of the savings banks, have grown sufficiently in size and importance to attract business even from very large concerns. In addition, the big banks now operate in such fields as agricultural credit, requiring a nationwide network of branches, and have extended their branching arrangements accordingly, sometimes by taking over smaller banks.

Although there is some concentration of banking resources, much of Italian banking is still decentralized, a pattern stemming from the time when Italy consisted of a number of separate states and kingdoms. Regionalism is still evident. Only in northern and central Italy is banking strongly integrated, though the economic development of the south will doubtless lead to further integration. As in other countries, some large banks have taken control of smaller banks without proceeding to a merger, so that the degree of actual concentration is greater than it appears.

In 1970 there were about 1,200 banks in Italy. Five of them were major banks, and another five might be described as large. There was also one major savings bank. Many of the others, including ordinary credit banks, cooperative people's banks, savings banks, and rural and artisan banks, were quite small. About 345 banks accounted for 97 percent of total deposits.

The major banks concentrate almost entirely on commercial banking, foreign banking, and the securities business. They also have subsidiaries that provide industrial credit and real estate financing. But the other banks offer a greater variety of services and therefore attract a greater share of the business.

For a long time the Italian government sought to stem the movement toward concentration and the absorption of small banks. By the late 1960s, however, the official attitude toward bank mergers had changed; the government recognized a need for rationalization and automation, as well as a need to spread risks more widely, and began to encourage the merging of smaller banks into larger units.

India. Until the 1950s, banking in India was carried on by a large number of small banks. India is still primarily an agricultural country, with an economic and social structure based largely on the villages. The integration of banking has been impeded by poor communications, by illiteracy, and by the barriers of language and caste. Banking and credit have remained largely in the hands of the so-called indigenous banker and the village moneylender. Although their influence has been greatly reduced in recent years, they still remain important in many an up-country area. The indigenous banker, who is also a merchant, offers genuine banking services: accepting deposits and remitting funds; making loans quickly and with a minimum of formality; and by means of the *hundi* (a credit instrument in the form of a promissory note) financing a still significant, if declining, portion of India's internal trade and commerce.

Problems in the integration of banking in India

Efforts were made to eliminate the moneylender by developing a network of rural credit cooperatives. When progress proved to be slow, a more successful alternative was found in requiring banks to open "pioneer" branches in rural areas. The first branches were those of the semi-public Imperial Bank of India and its nationalized successor, the State Bank of India (and its subsidiaries). Many smaller banks began to disappear, sometimes by merger and sometimes as a result of failure. The number of "reporting" banks fell from 517 in 1952 to 66 in 1970. Of these, 22 had been nationalized, including the State Bank, its seven subsidiaries, and the 14 large joint-stock banks taken over after 1969. The foregoing figures do not, of course, include the indigenous bankers and moneylenders.

The main path of banking development in India is now the expansion of bank branches into the under-banked areas. The authorities have sought to expand the number of branches but to avoid their concentration in the larger towns and cities and, in particular, to provide the rural areas with adequate facilities. The ultimate objective is to enable the mobilization of deposits on a massive scale throughout the country and a stepping up of lending to weak sectors of the economy. This is a massive undertaking in a country of 550,000 villages. In 1970, despite years of encouragement, there were only about 10,000 commercial bank offices in India.

Japan. Banking business in Japan is largely concentrated in the hands of the big banks (some of which are specialized), though a number of small banks still survive. At the end of 1971 there were 85 banks operating in Japan. Of these, 14 were city banks (over half with head offices in Tokyo and three in Osaka), three were special long-term credit banks, 61 were local banks, and

City banks and local banks in Japan

seven were trust banks. There were also 71 mutual loans and savings banks, 484 credit associations, and 528 credit cooperatives. The city banks have widespread networks of branches, though some are more regional in their interests than others. Some of the local banks, which on occasion are similar in size to city banks, also have a number of branches, making them really more regional than local. At the end of 1971 there were some 8,000 bank offices in Japan; about 2,400 belonged to the city banks, about 4,300 to the local banks, 262 to the trust banks, and 36 to the long-term credit banks. Thus, despite the concentration of business in the hands of the city banks, which at the end of 1971 held about 60 percent of total deposits, the Japanese system may be classed with the hybrid system.

It should be noted that the distinction between city banks and local banks has no legal basis; it rests on the fact that the city banks are under the supervision of the Department of Banks, while the local banks are regulated mainly by the Local Boards of Finance; both the city and the local banks belong to the Federation of Bankers' Associations of Japan; the local banks also belong to the Local Banks Association. The city banks service mainly manufacturing industry and commerce, particularly the big firms, while the local banks operate in and around a single city or in the rural areas, collecting deposits, lending to local business and the smaller firms, but also investing in securities and in the call-money market.

The local banks have city bank correspondents, not only to hold surplus balances but also for assistance in investing their funds, especially in the call-money market. In addition, a city bank may introduce certain of its large customers to a local bank (*e.g.*, a big company having a local factory) with a view to a loan from the local bank to the customer concerned. City correspondents in Japan do not, however, provide the wide range of ancillary services that is common in the United States.

Since World War II there has been much stability in Japanese banking, but there are many in Japan who are anxious to see changes. This is particularly true of the city banks, which have suffered a relative decline in the importance of their business in competition with other institutions. This has been true in particular of the agricultural cooperatives, which attract the larger part of the Treasury's payments on account of government purchases of the rice crop; there has also been a relative increase in the importance of the life insurance companies and the trust funds, which have attracted sizable funds from the general public.

Because the need for radical overhaul had become so apparent, the problem was referred to the Committee on Financial System Research, set up to advise the Minister of Finance in 1966. As a result, two laws were passed in 1968, the first of which was concerned to revise arrangements relating to mutual loans and savings banks and to improve the financing of small business, and the second to permit mergers and conversions of status between different types of financial institutions. In 1967, the committee began to investigate the system of private financial institutions (other than those for small businesses and agriculture) with an eye to the possible diversification of commercial bank business, the promotion of efficiency in banking (and the possible introduction of a deposit insurance scheme), and arrangements for financing foreign trade within the context of a changing international monetary environment. In order to sustain stable economic growth, the committee recommended in 1970 that greater resort be made to interest rates (including interest rates on deposits) as a means of influencing business activity; that greater strength be imparted (*e.g.*, through mergers, but on the initiative of the institutions concerned) to financial institutions serving the national economy; that, although specialization was to be favoured, there should be some flexibility in the choice of business undertaken, thereby helping to stimulate "fair competition" (*e.g.*, in medium-term financing); and that a deposit insurance scheme should be introduced (relevant legislation was enacted in 1971).

A major banking merger took place in 1971. In the short run, extension of existing cooperation agreements is the more likely pattern (*e.g.*, for sharing Telex services among local banks, the joint use of electronic computers, resort to mutual credit cards, and so on). In due course these agreements will almost certainly lead to interlocking capital holdings, on which basis their more complete integration will tend to follow. In this way, banks in the provinces may well build themselves up into units large enough to compete directly even with merged city banks.

BANKING IN PLANNED ECONOMIES

The Soviet Union. The present-day Soviet banking system was established by the credit reforms of the early 1930s, which centralized practically all short-term credit in the hands of the Gosbank (State Bank, established in 1921). There was much restructuring of banking during succeeding years, mainly to ensure that the system became an effective instrument for carrying out the national economic plan. The Gosbank's control over payments flows was also tightened in order to maintain stability of prices. The activities of the Gosbank are by no means limited to purely financial operations. For example, Gosbank economists are also employed, on a voluntary basis, by organs of the Communist Party and the government to assist in regulating activity in a wide range of economic sectors and to conduct investigations into the business and financial activities of the enterprises that are their customers.

Functions of the Gosbank

The Gosbank was originally concerned with the provision of short-term credit; a number of other banks were created to finance capital investment in the socialized economy. Even in the 1930s there was a tendency to consolidate these banking units, and this continued into the postwar period. In 1959 there were further mergers and a reallocation of activity, as a result of which the Investment Bank (Stroibank) emerged as the means of canalizing state budgetary appropriations into capital investment. The savings bank system, with over 76,000 branches, became part of the Gosbank in 1963. The only other banking institution remaining in existence is the Bank for Foreign Trade (Vneshtorgbank), whose field of operations was considerably expanded in 1961. Originally concerned mainly with the provision of currency for tourists and diplomatic missions and with remittances from abroad, the Bank for Foreign Trade came to handle all foreign-exchange transactions, including those relating to trade.

Since the functions of the Investment Bank are essentially administrative (supervising the disbursement of budgetary grants), the Gosbank is the only domestic bank servicing the cash, credit, and payment needs of a population exceeding 240,000,000. The organization of the Gosbank is as follows: it has a policy-making head office and principal offices in the various Soviet republics. There are also regional offices and a network of about 4,000 local branches; the latter are the bank's main points of contact with a variety of economic enterprises, collective farms, and lower level government units. To serve the needs of the urban population, the Gosbank also maintains numerous collection offices (part of the network of communal banks abolished in 1959), which collect rent, taxes, and other compulsory payments and contributions. It maintains a small number of special cash service agencies in large industrial establishments and at construction projects. Seasonal agencies are operated at remote places where large purchases of farm products are made at certain times of the year.

Organizational structure of the Gosbank

In the industrial area, the Gosbank services hundreds of thousands of state enterprises that operate on the basis of cost accounting; each of these enterprises has its own working capital and prepares a balance sheet and a statement of income; it may borrow regularly or occasionally, depending on its needs. In addition, there are tens of thousands of collective farm accounts. But the Gosbank has hundreds of thousands of other customers comprising party, trade union, cultural and other organizations, and individuals. Most of the accounts held by individuals are loan accounts for financing homes.

The aggregate balances maintained by Gosbank customers are small in comparison with the cash balances held by business and government accounts in the United States.

The operation of a centralized payments system in a country the size of the Soviet Union gives rise to many problems. The system of payments, based mainly on documentary drafts, involves the processing of a very large number of items, an operation that in recent years has been to some extent automated. Changes in several sectors of the economy have greatly increased the complexity of the bank's operations.

Transactions of individuals. Although consumer loans are extended by stores rather than by banks, there has also been a rapid increase in the transactions of individuals. This has resulted from rising incomes, an increase in savings, and the provision of facilities for crediting wages to savings accounts and making periodic payments from them. Housing loans and tourism have also been growing in importance. As savings bank offices handle virtually all the accounts and transactions of individuals, the considerable increase in the volume of these transactions was probably the main reason for absorbing the savings bank system into the Gosbank in 1963.

Transactions of collective farms. An even greater problem has been created by the rise in transactions involving the collective farms. Before 1953 the collectives paid in kind for the services performed by the state-owned tractor and farm-machinery stations; they were also required to deliver to the state a large part of their output. Since 1953 these transactions have gradually shifted from payment in kind to payment in cash. As a first step, the tractor and machinery stations were closed down and their equipment sold to the collectives. The farms now had to pay in cash for all machinery and fuel, as well as for building materials, fertilizers, and other supplies. Subsequently, they were enabled to sell their output to the state for money.

Farm labour likewise has come to be remunerated in cash rather than in kind. In 1953 only one-third of the "compensation" for work contributed by members of the collectives was in cash; five years later the proportion had risen to more than half, and by 1963 it was nearly three-quarters. In 1965 the flow of money income to the farm population was increased further by the introduction of state pensions for collective farmers. In the summer of 1966, it was decided to make minimum monthly payments to the members of collective farms, and this resulted in an even greater use of money in the farm sector.

The growth of money flows and of bank lending in rural areas, as well as the need to service a growing clientele in the villages, greatly added to the complexity of Gosbank operations, which had been geared primarily to the needs of industry and government. The Gosbank attempted to resolve its problems by simplifying payments procedures, though its efforts to work out a system of offsetting mutual claims were not initially very successful. The settlements mechanism remained one of the major operating problems of the Gosbank.

Interest rates in the U.S.S.R.

In the Soviet economy, interest rates serve primarily to pay the operating expense of the Gosbank. After the credit reforms of 1930–32, a uniform system of interest rates was applied to all short-term credits, irrespective of the purpose of the loan or the financial condition of the borrower. Higher rates were charged as a penalty on overdue loans. In the 1960s there was a move to differentiate rates, it being accepted as a matter of principle that bank funds should be more expensive than an enterprise's own working capital and that borrowings made necessary by shortcomings of management (*e.g.*, excessive inventories or the erosion of working capital) should carry higher rates than loans to finance normal needs. Penalty rates for overdue loans and for late payments were also increased, and collection was more vigorously enforced. But the rate of interest was employed only to a very limited extent as a means of allocating resources among alternative uses.

Yugoslavia. The banking system that had emerged in the Socialist Federal Republic of Yugoslavia by 1971 represented an evolution away from the highly centralized institutional arrangements of the Soviet-type system. It also differed from the Western capitalist model. In Yugoslavia, management is ultimately in the hands of the depositors, these consisting mainly of enterprises or sociopolitical bodies having large deposits in the bank.

Before World War II Yugoslavia's banking system had consisted of a number of private and state banks. When the Communists came to power, the private banks disappeared, and the state banks were reorganized. The National Bank, a descendant of the Serbian National Bank founded in 1883, was retained. The former State Mortgage Bank became the State Investment Bank, and the Agricultural Bank founded in 1929 continued to operate. An Industrial Bank was also established to assist in implementing the ambitious program of industrialization, and there were six regional banks. In 1946, all the existing banks were merged into the National Bank and the State Investment Bank; the former was to engage in short-term transactions, while the other handled investments and foreign loans. The National Bank also issued currency, provided general banking and agency facilities for the government, and served as a clearing house for the entire economy. In 1948 the system was expanded again, with the formation of 89 communal banks and six regional state banks for lending to agricultural cooperatives. Though it was not deemed appropriate in a socialist system to charge interest for the use of capital, a commission of 1 percent was charged for bank services. Additional changes came in the early 1950s, when the communal banks were abolished and the others merged with the National Bank. A "reasonable" rate of interest was instituted so that credit would be used economically. In 1954 the National Bank experimented with credit auctions as a means of allocating funds among borrowers, but the less profitable enterprises tended to offer the highest rates; in 1955 the bank returned to the practice of evaluating every application for credit separately.

In another revolt against overcentralization, the communal banks were re-established, though they were required to keep substantial reserves with the National Bank. Two-thirds of each bank's managing board was to be nominated by workers' councils of the enterprises located in its area. Three specialized federal government banks were also set up—a Foreign Trade Bank, an Investment Bank, and an Agricultural Bank.

The banking reforms of 1965 in Yugoslavia

These continued to operate until the reforms of 1965, when the National Bank was given the power to regulate the commercial banks (formerly the federal government banks and the communal banks) by varying their reserve requirements, limiting the amounts they could borrow from it, and sometimes imposing a ceiling on commercial bank lending. The National Bank was also empowered to vary the special credits granted to the commercial banks, used to finance about half the short-term credits extended by the commercial banks to their customers. Three types of bank credit were contemplated: investment, commercial, and consumer. Investment banks were established to finance the capital requirements of enterprises; commercial banks to provide short-term credit; "mixed banks" that served as both investment and commercial banks; and savings banks. In principle, all banks could operate throughout the country. These changes were followed by a marked concentration of banking: in November 1964 there had been 217 banks (206 communal banks, eight republican investment banks, and the three specialized federal banks); by June 1967 there were only 111 banks (61 commercial banks, 39 "mixed banks," and 11 investment banks), and by the end of 1968, the number of banks had been further reduced to 74.

The biggest change, however, was in the organization and operation of the banks. Banks were now established by enterprises in partnership with sociopolitical communities (federal, republican, and local). Each bank had its own capital, or "credit fund." The founders invested their capital in the credit fund of the bank and became shareholders. No single shareholder was entitled to more than

10 percent of the total number of votes in a bank's assembly of shareholders (which also contained representatives of the bank's personnel). No enterprise or sociopolitical community could be refused the right to invest in a bank and to take part in its management. Shareholders were entitled to dividends, depending on profitability, but these tended to be invested in the bank as new capital. Management was placed in the hands of an executive committee appointed by the assembly of shareholders.

The decentralization of the banking system was offset to some extent by mergers, some of which produced several very large banks. A number of banks entered into agreements with foreign banks as a means of encouraging greater foreign participation in the Yugoslav economy. In the field of banking, as in economic affairs generally, the Yugoslavs were prepared to experiment on a broad scale.

III. The business of banking

The business of banking consists of borrowing and lending. As in other businesses, operations must be based on capital, but banks employ comparatively little of their own capital in relation to the total volume of their transactions. The purpose of capital and reserve accounts is primarily to provide an ultimate cover against losses on loans and investments. In the United States capital accounts also have a legal significance, since the laws limit the proportion of its capital a bank may lend to a single borrower.

FUNCTIONS OF COMMERCIAL BANKS

The balance sheet

The essential characteristics of the banking business may be described within the framework of a simplified balance sheet. A bank's main liabilities are its capital (including reserves) and deposits. The latter may be from domestic or foreign sources (corporations and firms, private individuals, other banks, and even governments). They may be repayable on demand (sight deposits or current accounts) or repayable only after the lapse of a period of time (time, term, or fixed deposits and, occasionally, savings deposits). A bank's assets include cash (which may be held in the form of credit balances with other banks, usually with a central bank but also, in varying degrees, with correspondent banks); liquid assets (money at call and short notice, day-to-day money, short-term government paper such as treasury bills and notes, and commercial bills of exchange, all of which can be converted readily into cash without risk of substantial loss); investments or securities (substantially medium-term and longer term government securities—sometimes including those of local authorities such as states, provinces, or municipalities—and, in certain countries, participations and shares in industrial concerns); loans and advances made to customers of all kinds, though primarily to trade and industry (in an increasing number of countries, these include term loans and also mortgage loans); and, finally, the bank's premises, furniture, and fittings (written down, as a rule, to quite nominal figures).

All bank balance sheets must include an item that relates to contingent liabilities (*e.g.*, bills of exchange "accepted" or endorsed by the bank), exactly balanced by an item on the other side of the balance sheet representing the customer's obligation to indemnify the bank (which may also be supported by a form of security taken by the bank over its customer's assets). Most banks of any size stand prepared to provide acceptance credits (also called bankers' acceptances); when a bank accepts a bill, it lends its name and reputation to the transaction in question and, in this way, ensures that the paper will be more readily discounted.

Deposits. The bulk of the resources employed by a modern bank consists of borrowed money (that is, deposits), which is lent out as profitably as is consistent with safety. Insofar as an increase in deposits provides a bank with additional cash (which is an asset), the increase in cash supplements its loanable resources and permits a more than proportionate increase in its loans.

The way in which deposits may be increased

An increase in deposits may arise in two ways. (1) When a bank makes a loan, it may transfer the sum to a current account, thus directly creating a new deposit; or it may arrange a line of credit for the borrower upon which he will be permitted to draw checks, which, when deposited by third parties, likewise create new deposits. (2) An enlargement of government expenditure financed by the central bank may occasion a growth in deposits, since claims on the government that are equivalent to cash will be paid into the commercial banks as deposits. In the first instance, with the increase in bank deposits goes a related increase in the potential liability to pay out cash; in the second case, the increase in deposits with the commercial banks is accompanied by a corresponding increase in commercial bank holdings of money claims that are equivalent to cash.

Taking one bank in isolation, an increase in its loans may result in a direct increase in deposits. This may occur either as a result of a transfer to a current account (as above) or a transfer to another customer of the same bank. Once again, there is an increase in the potential liability to pay out cash. On the other hand, if there has been an increase in loans by another bank (including an increase in central bank loans to the government), this may give rise to increased deposits with the first bank, matched by a corresponding claim to cash (or its equivalent). For these reasons a bank can generally expect that, if there is an increase in deposits, there will also be some net acquisition of cash or of claims for receipt of cash. It is in this way that an increase in deposits usually provides the basis for further bank lending.

Except in countries where banks are small and insecure, banks as a whole can usually depend on their current account debits being largely offset by credits to current account, though from time to time an individual bank may experience marked fluctuations in its deposit totals, and all banks in a country may be subject to seasonal variations. Even when deposits are repayable on demand, there is usually a degree of inertia in the deposit structure that prevents sharp fluctuations; if money is accepted contractually for a fixed term or if notice must be given before its repayment, this inertia will be greater. On the other hand, if a significant proportion of total deposits derives from foreign sources, there is likely to be an element of volatility arising from international conditions.

In banking, confidence on the part of the depositors is the true basis of stability. Confidence is steadier if there exists a central bank to act as a "lender of last resort." Another means of maintaining confidence employed in some countries is deposit insurance, which protects the small depositor against loss in the event of a bank failure. Such protection was the declared purpose of the "nationalization" of bank deposits in Argentina between 1946 and 1957; banks receiving deposits acted merely as agents of the government-owned and government-controlled central bank, all deposits being guaranteed by the state.

Maintaining liquidity

Reserves. Since the banker undertakes to provide his depositors with cash on demand or upon prior notice, he must hold a cash reserve and maintain a "safe" ratio of cash to deposits. The safe ratio is determined largely through experience. It may be established by convention (as it was for many years in England) or by statute (as in the United States and elsewhere). If a minimum cash ratio is required by law, a portion of a bank's assets is in effect frozen and not available to meet sudden demands for cash from the bank's customers. In order to provide more flexibility, required ratios are frequently based on the average of cash holdings over a specified period, such as a week or a month. In addition to holding part of his assets in cash, a banker will hold a proportion of the remainder in assets that can quickly be converted into cash without significant loss. No banker can safely ignore the necessity of maintaining adequate reserves of liquid assets; some prefer to limit the sum of loans and investments to a certain percentage of deposits (*e.g.*, 70 percent), not allowing their loan–deposit ratio to run for any length of time at too high a level.

Unless a bank holds cash covering 100 percent of its demand deposits it could not meet the claims of depositors if they were all to exercise in full and at the same time their rights to demand cash. If that were a common phe-

nomenon, deposit banking could not long survive. For the most part, the public is prepared to leave its surplus funds on deposit with the banks, confident that they will be repaid if needed. But there may be times when unexpected demands for cash exceed what might reasonably have been anticipated; therefore, a bank must not only hold part of its assets in cash but also must keep a proportion of the remainder in assets that can be quickly converted into cash without significant loss. Indeed, in theory, even its less liquid assets should be self-liquidating within a reasonable time.

A bank may mobilize its assets in several ways. It may demand repayment of loans, immediately or at short notice; it may sell securities; or it may borrow from the central bank, using paper representing investments or loans as security. Banks do not precipitately call in loans or sell marketable assets because this would disrupt the delicate debtor–creditor relationships and increase any loss of confidence, probably resulting in a run on the banks. Ready cash may be obtainable in this way only at a very high price. Banks must either maintain their cash reserves and other liquid assets at a high level or have access to a "lender of last resort," such as a central bank, able and willing to provide cash against the security of eligible assets. In some countries the commercial banks have at times been required to maintain a minimum liquid assets ratio. This has been common, for example, in some of the western European countries, in England, Australia, Canada, and in several of the new African countries. But central banks impose such requirements less as a means of maintaining appropriate levels of commercial bank liquidity than as a technique for influencing directly the lending potential of the banks (see below).

Kinds of assets of commercial banks

Among the assets of commercial banks, investments are less liquid than money-market assets such as call money and treasury bills. By maintaining an appropriate spread of maturities, however, it is possible to ensure that a proportion of a bank's investments is regularly approaching redemption, thereby producing a steady flow of liquidity and in that way constituting a secondary liquid assets reserve. Some banks, particularly in the United States and Canada, favour a "dumbbell" distribution of maturities, a significant proportion of the total portfolio being held in long-dated maturities with a high yield, a small proportion in the middle ranges, and another significant proportion in short-dated maturities. Following redemption, the banks usually reinvest all or most of the proceeds in longer term maturities that in due course become increasingly short-term. Investments and money-market assets merge into each other. The dividing line is arbitrary, but there is an essential difference: the liquidity of investments depends primarily on marketability (though sometimes it also depends on the readiness of the government or its agent to exchange its own securities for cash); the liquidity of money-market assets, on the other hand, depends partly on marketability but mainly on the willingness of the central bank to purchase them or accept them as collateral for a loan. This is why money-market assets are more liquid than investments.

INDUSTRIAL FINANCE

Long-term and medium-term lending. Banks that do a great deal of long-term lending to industry must ensure their liquidity by maintaining relatively large capital funds and a relatively high proportion of time deposits, as well as valuing their investments very conservatively. Such banks, notably the French *banques d'affaires* and the West German commercial banks, have developed special means of reducing their degree of risk. Every investment is preceded by a thorough technical and financial investigation. The initial advance may be an interim credit, later converted into a participation. Only when market conditions are favourable is the original investment converted into marketable securities, and an issue of shares to the public is arranged. One function of these banks is to nurse an investment along until the venture is well established. Even assuming its ultimate success, a bank may be obliged to hold such shares for long periods before being able to liquidate them. In addition, they often retain an interest in a firm as an ordinary investment as well as to ensure a degree of continuing control over it.

The long-term provision of industrial finance in Britain and the Commonwealth countries is usually handled by specialist institutions, with the commercial banks providing only part of the necessary capital. In Japan, the long-term financial needs of industry are met partly by special industrial banks (which also issue debentures as well as accepting deposits) and partly by the ordinary commercial banks. In West Germany the commercial banks customarily handle long-term finance.

Term loans

Since World War II, the commercial banks in the United States have developed the so-called term loan, especially for financing industrial capital requirements. The attempt to popularize the term loan began in the economic depression of the 1930s, when the banks tried to expand their business by offering finance for a period of years. Most term loans have an effective maturity of little more than five years, though some run for 10 years or more. They are usually arranged between the customer and a group of lending banks, sometimes in cooperation with other institutions such as insurance companies. Banks in other countries, including Britain and Australia, began during the 1960s to give term loans both to industry and to agriculture.

Short-term lending. Short-term loans are the core of the banking business even in countries where commercial banks make long-term loans to industry. Much short-term lending consists in the provision of working capital, but the banks also provide temporary finance for fixed capital development, aiding a customer until he can find long-term finance elsewhere.

Overdrafts

Much of this short-term lending is done by overdraft, particularly in Britain and a number of the Commonwealth countries, in The Netherlands and Norway, and—on the basis of somewhat similar arrangements—in Austria, Denmark, and Switzerland. The overdraft permits a depositor to overdraw his account up to an agreed limit. In theory, overdrafts are repayable on demand or after reasonable notice has been given, but often they are allowed to run indefinitely, subject to a periodic review. An advance is reduced or repaid whenever the account is credited with deposits and recreated when new cheques are drawn upon it, interest being paid only on the amount outstanding.

An alternative method of short-term lending is to debit a loan account with the amount borrowed, crediting the proceeds to a current account; interest is usually payable on the whole amount of the loan, which normally is for a fixed period of time. (In Britain arrangements are sometimes more flexible, and the term of the loan may be set by oral agreement.)

Discountable paper

In some countries, including the United States, France, West Germany, and Japan, short-term finance is often made available on the basis of discountable paper—commercial bills or promissory notes. Some of this paper is usually rediscountable at the central bank, thus becoming virtually a liquid asset, unlike a bank advance or loan.

Credit may be offered with or without formal security, depending on the reputation and financial strength of the borrower. In many countries a customer may use a number of banks, and these institutions usually freely exchange information about joint credit risks. In Britain and The Netherlands, however, most concerns tend to use a single banking institution for most of their needs.

Liability management

Traditionally bankers took the view that the liabilities of a bank (and, in particular, its deposits) were more or less stable and concerned themselves primarily with the investment of these funds. In recent years, however, especially in North America, there has been a change of emphasis. During the late 1950s and 1960s the banks found it more difficult to obtain deposits. Interest rates rose to high levels, and banks were obliged to compete with each other and with other institutions for funds. At the same time, there was little point in paying a high rate of interest for money unless it could be employed profitably. Bankers began to relate the cost of borrowed money directly to the return on investments. Previously the main limitation on a bank's expansion had been its

ability to find profitable new business, but now the determining factor became the availability of funds to lend out. The essence of liability management, as it came to be called, was deciding what kinds of new money to buy and what to pay for it. There was an incentive to borrow at longer term or to borrow money in forms that were less volatile than demand deposits had become: savings deposits, time deposits (including certificates of deposit, which could be sold by the lender to someone else if he wanted his funds before they matured), federal funds in the United States (entitlements to balances at a Federal Reserve Bank, the excess supplies of which were regularly traded throughout the United States), and even Eurodollars (entitlements to dollar balances usually held outside the United States, chiefly in Europe). In short, if loan demand was strong and the business likely to be profitable, the banks would seek to buy new money to meet these demands. To some extent banks tended to relate the periods for which money was borrowed to the periods for which it was to be lent. Since large banks generally found it easier to obtain money, they tended to become larger; the smaller banks, on the other hand, felt more directly than larger institutions the impact of central bank restrictions, which they could not so easily circumvent.

IV. The principles of central banking

The need for a central bank

The principles of central banking grew up in response to the recurrent British financial crises of the 19th century and were later adopted in other countries. Modern market economies are subject to frequent fluctuations in output and employment. Although the causes of these fluctuations are various, there is general agreement that the ability of banks to create new money may exacerbate them. Although an individual bank may be cautious enough in maintaining its own liquidity position, the expansion or contraction of the money supply to which it contributes may be excessive. This raises the need for a disinterested outside authority able to view economic and financial developments objectively and to exert some measure of control over the activities of the banks. A central bank should also be capable of acting to offset forces originating outside the economy, although this is much more difficult.

RESPONSIBILITIES OF CENTRAL BANKS

The first concern of a central bank is the maintenance of a soundly based commercial banking structure. While this concern has grown to comprehend the operations of all financial institutions, including the several groups of nonbank financial intermediaries, the commercial banks remain the core of the banking system. A central bank must also cooperate closely with the national government. Indeed, most governments and central banks have become intimately associated in the formulation of policy.

The central bank as "lender of the last resort"

Relationships with commercial banks. One source of economic instability is the supply of money. Even in relatively well controlled banking systems, banks have sometimes expanded credit to such an extent that inflationary pressures developed. Such an overexpansion in bank lending would be followed almost inevitably by a period of undue caution in the making of loans. Frequently the turning point was associated with a financial crisis, and bank failures were not uncommon. Even today, failures occur from time to time. Such crises in the past often threatened the existence of financial institutions that were essentially sound, and the authorities sometimes intervened to prevent complete collapse.

The willingness of a central bank to offer support to the commercial banks and other financial institutions in time of crisis was greatly encouraged by the gradual disappearance of weaker institutions and a general improvement in bank management. The dangers of excessive lending came to be more fully appreciated, and the banks also became more experienced in the evaluation of risks. In some cases, the central bank itself has gone out of its way to educate commercial banks in the canons of sound finance. In the United States, the Federal Reserve System examines the books of the commercial banks and carries on a range of frankly educational activities. In developing countries such as India and Pakistan, central banks have also set up departments to maintain a regular scrutiny of commercial bank operations.

The most obvious danger to the banks is a sudden and overwhelming run on their cash resources in consequence of their liability to depositors to pay on demand. In the ordinary course of business, the demand for cash is fairly constant or subject to seasonal fluctuations that can be foreseen. It has become the responsibility of the central bank to protect banks that have been honestly and competently managed from the consequences of a sudden and unexpected demand for cash. In other words, the central bank came to act as the "lender of last resort." To do this effectively, it was necessary that the central bank be permitted either to buy the assets of commercial banks or to make advances against them. It was also necessary that the central bank have the power to issue money acceptable to bank depositors. But if a central bank was to play this role with respect to commercial banks, it was only reasonable that it or some related authority be allowed to exercise a degree of control over the way in which the banks conducted their business.

Role of central banks in the banking system

Most central banks now take a continuing day-to-day part in the operations of the banking system. The Bank of England, for example, has been increasingly in the market to assure that the banks have a steady supply of cash, even during periods of credit restriction. It also lends regularly to the discount houses, supplementing their resources whenever the commercial banks feel the need to call back money they have on loan to them. In the United States the Federal Reserve System has operated in a similar way by buying and selling securities on the open market and by lending to dealers in government securities on the basis of repurchase agreements. The Federal Reserve may also discount paper submitted by the commercial banks through the Federal Reserve banks. The various techniques of credit control used in the United States, Great Britain, and other countries are discussed in greater detail below.

The evolution of those working relations among banks implies a community of outlook that in some countries is relatively recent. The whole concept of a central bank as responsible for the stability of the banking system presupposes mutual confidence and cooperation. For this reason, contact between the central bank and the commercial banks must be close and continuous. The latter must be encouraged to feel that the central bank will give careful consideration to their views on matters of common concern. Once the central bank has formulated its policy after a full consideration of the facts and of the views expressed, however, the commercial banks must be prepared to accept its leadership. Otherwise, the whole basis of central banking would be undermined.

The central bank and the national economy. *Relationships with other countries.* Since no modern economy is self-contained, central banks must give considerable attention to trading and financial relationships with other countries. If goods are bought abroad, there is a demand for foreign currency to pay for them. Alternatively, if goods are sold abroad, foreign currency is acquired that the seller ordinarily wishes to convert into his own currency. These two sets of transactions usually pass through the banking system, but there is no necessary reason why, over the short period, they should balance. Sometimes there is a surplus of purchases and sometimes a surplus of sales. Short-period disequilibrium is not likely to matter very much, but it is rather important that there be a tendency to balance over a longer period, since it is difficult for a country to continue indefinitely as a permanent borrower or to continue building up a command over goods and services that it does not exercise.

Maintaining the balance of foreign payments

Short-period disequilibrium can be met very simply by diminishing or building up balances of foreign exchange. If a country has no balances to diminish, it may borrow, but normally it at least carries working balances. If the commercial banks find it unprofitable to hold such balances, the central bank is available to carry them; indeed,

it may insist on concentrating the bulk of the country's foreign-exchange resources in its hands or in those of an associated agency.

Long-period equilibrium is more difficult to achieve. It may be approached in three different ways: price movements, exchange revaluation (appreciation or depreciation of the currency), or exchange controls.

Price levels may be influenced by expansion or contraction in the supply of bank credit. If the monetary authorities wish to stimulate imports, for example, they can induce a relative rise in home prices by encouraging an expansion of credit. If additional exports are necessary in order to achieve a more balanced position, the authorities can attempt to force down costs at home by operating to restrict credit.

The objective may be achieved more directly by revaluing a country's exchange rate. Depending on the circumstances, the rate may be appreciated or depreciated, or it may be allowed to "float." Appreciation means that the home currency becomes more valuable in terms of the currencies of other countries and that exports consequently become more expensive for foreigners to buy. Depreciation involves a cheapening of the home currency, thus lowering the prices of export goods in the world's markets. In both cases, however, the effects are likely to be only temporary, and for this reason the authorities often prefer stability in exchange rates even at the cost of some fluctuation in internal prices.

Quite often governments have resorted to exchange controls (sometimes combined with import licensing) to allocate foreign exchange more or less directly in payment for specific imports. At times, a considerable apparatus has been assembled for this purpose and, despite "leakages" of various kinds, the system has proved reasonably efficient in achieving balance on external payments account. Its chief disadvantage is that it interferes with normal market processes, thereby encouraging rigidities in the economy, reinforcing vested interests, and restricting the growth of world trade.

Whatever method chosen, the process of adjustment is generally supervised by some central authority—the central bank or some institution closely associated with it—which can assemble the information necessary to ensure that the proper responses are made to changing conditions.

Moderating booms and slumps

Economic fluctuations. As noted above, monetary influences may be an important contributory factor in economic fluctuations. An expansion in bank credit makes possible, if it does not cause, the relative overexpansion of investment activity characteristic of a boom. Insofar as monetary policy can assist in mitigating the worst excesses of the boom, it is the responsibility of the central bank to regulate the amount of lending by banks and perhaps by other financial institutions as well. The central bank may even wish to influence in some degree the direction of lending as well as the amount.

An even greater responsibility of the central bank is that of taking measures to prevent or overcome a slump. Recessions, when they occur, are often in the nature of adjustments to eliminate the effects of previous overexpansion. Such adjustments are necessary to restore economic health, but at times they have tended to go too far; depressive factors have been reinforced by a general lack of confidence, and, once this has happened, it has proved extremely difficult to stimulate recovery. In these circumstances, prevention is likely to be far easier than cure. It has therefore become a recognized function of the central bank to take steps to preclude, if possible, any such general deterioration in economic activity.

For the central bank to be effective in regulating the volume and distribution of credit so that economic fluctuations may be damped, if not eliminated, it must at least be able to regulate commercial bank liquidity (the supply of cash and "near cash") because this is the basis of bank lending. There is now little dispute about the broad objectives, though the techniques of control are various and depend to some extent on environmental factors. It would be incorrect to suppose, however, that the actions of the central bank can, unaided, achieve a high degree of stability. It can by wise guidance contribute to that end, but monetary action is in no sense a panacea; at all times, the degree to which it is likely to be effective depends on the provision of an appropriate fiscal environment (see FISCAL AND MONETARY POLICY).

Developing the banking system

Banking services. Another responsibility of the central bank is to ensure that banking services are adequately supplied to all members of the community that need them. Some areas of a country may be "under-banked" (*e.g.*, the rural areas of India and the northern and more remote parts of Norway), and central banks have attempted, directly or indirectly, to meet such needs. In France, this need underlay the early extension of branches of the Bank of France to the *départements*. In India, as noted above, the authorities encouraged the opening of "pioneer" branches by the former Imperial Bank of India and its successor, the State Bank of India, as well as the extension of bank branches to rural and semirural areas. In Pakistan, officials of the State Bank of Pakistan played an active part in the foundation of the semipublic National Bank of Pakistan.

A different sort of problem arises when the business methods of existing banks are unsatisfactory. In such circumstances, a system of bank inspection and audit organized by the central banking authorities (as in India and Pakistan) or of bank "examinations" (as in the United States) may be the appropriate answer. Alternatively, the supervision of bank operations may be handed over to a separate authority, such as France's Banking Control Commission or South Africa's Registrar of Banks.

In developing countries, central banks may undertake to encourage the establishment and growth of specialist institutions such as savings institutions and agricultural credit or industrial finance corporations. These serve to improve the mechanism for tapping existing liquid resources and to supplement the flow of funds for investment in specific fields.

The central bank as a public institution

Responsibilities to the government. Central banks have over the years acquired a number of well-defined responsibilities to their respective national governments. Some, notably the Bank of England, developed into central banks after being, in origin, bankers to the government. More recently it has become a matter of course for a new central bank to accept responsibility for the financial affairs of its government. The reasons are self-evident. Government transactions have become of increasing importance in influencing the workings of the economy, and the institution that holds the government's account is in a strategic position to cushion the commercial banks against the impact of large movements of cash originating in this way. As banker to the government, furthermore, the central bank has an obvious responsibility to provide routine banking services, such as arranging loan flotations and supervising their service, renewal, and redemption. The central bank also usually issues the currency.

Equally important are its responsibilities as an adviser on the probable monetary consequences of any proposed action. In this role the central bank should scrutinize the government's proposals with a certain amount of objectivity and state its point of view with vigour. One may cite a now-famous dictum of Montagu Norman as governor of the Bank of England:

> I think it is of the utmost importance that the policy of the Bank and the policy of the Government should at all times be in harmony—in as complete harmony as possible. I look upon the Bank as having the unique right to offer advice and to press such advice even to the point of "nagging"; but always of course subject to the supreme authority of the Government.

Many central banks are now nationalized, reflecting the increasingly general recognition of the central bank as a servant, if not a creature, of the government. This is also, in a way, a final recognition of the central bank as a responsible public institution, whose function it is to serve the community as a whole, untrammelled by narrow dictates of profit and loss. Most central banks, nevertheless, make very handsome profits.

TECHNIQUES OF CREDIT CONTROL

Central banks have developed a variety of techniques for influencing, regulating, and controlling the activities of commercial banks. These may be divided into (1) the so-called classical, or indirect, techniques and (2) various "direct" controls. The classical techniques make use of open-market purchases or sales by the central bank of certain types of assets, invariably associated with changes in interest rates. Direct or quantitative credit controls are used to influence the cash and liquidity bases of commercial bank lending by freezing or unfreezing their liquid resources; sometimes ceilings are imposed on bank loans.

Purchases and sales of securities by the central bank

Open-market operations. The way in which open-market operations influence the cash reserves and, through them, the general liquidity of the commercial banks is essentially simple. If the central bank buys securities in the open market, the cash it offers in exchange adds to the reserves of the banks; if the central bank sells securities in the open market, the cash necessary to pay for them is either withdrawn from the banks' reserves or obtained by diminishing holdings of other assets (with the possibility of capital losses in consequence of these sales). It does not matter whether this buying and selling takes place between the central bank and the commercial banks directly or between the central bank and other financial sectors, including the public at large, since these are the customers of the commercial banks.

Open-market operations are invariably associated with related changes in one or more "strategic" rates of interest, the most influential of these rates being the minimum rate at which the central bank does business (the bank rate, or the discount rate), since other rates tend to move in sympathy with it. The central bank seeks to achieve an appropriate and consistent structure of interest rates. If a particular rate structure is desired (*e.g.*, prior to a new issue of government securities or in order to change the emphasis of institutional investment between, say, long-term and short-term securities), it may be necessary to precondition the market by means of open-market operations. To achieve its purposes the central bank must possess (if it is selling) or be willing to absorb (if it is buying) the appropriate types of securities.

In London, the specialist banks known as discount houses effectively put to work the revolving fund of cash that circulates through the British banking system. If temporarily there is an inadequate supply of cash, the Bank of England either lends on a short-term basis or buys some of the assets held by the discount market. Alternatively, the Bank of England may buy assets from the clearing banks (the large joint-stock banks), which then make the relevant moneys available to the market. On the other hand, if the discount market is oversupplied with funds, the Bank of England sells treasury bills, in this way mopping up the excess of cash. These transactions are known as smoothing-out operations.

Not all Bank of England transactions in securities are in treasury bills. Even for smoothing-out purposes, first-class acceptances and government bonds with a maximum maturity date of not more than five years are eligible. In addition, the Bank of England is also responsible for managing the national debt, and, whether the object is to influence the flows of money or not, such transactions in fact have monetary effects.

The Open Market Committee of the Federal Reserve

In the United States the Federal Reserve System regulates the money supply. Within the Federal Reserve System, the Federal Open Market Committee is the most important single policy-making body. It is presided over by the chairman of the Board of Governors, with the president of the Federal Reserve Bank of New York as its permanent vice chairman. The main responsibility of the Open Market Committee is to decide upon the timing and amount of open-market purchases or sales of government securities. Since open-market operations must obviously be consistent with other aspects of monetary and credit policy, it is in the committee that broad agreement is reached on matters such as changes in discount rates or reserve requirements.

One of the big differences between London and New York is that the central banking authorities in New York maintain direct relationships more or less continuously with the nonbank government securities dealers as well as with the commercial banks. The Federal Reserve Bank of New York may make temporary accommodation available to nonbank dealers under a repurchase agreement, whereby securities are sold to the bank under an agreement that they be repurchased after a stipulated time. These agreements are made only for the purpose of supplying reserves to the banking system, but from the dealer's standpoint they are helpful in financing portfolios. Since early 1966, the bank has also been prepared to mop up money by undertaking reverse repurchase agreements, in which the nonbank dealers act as intermediaries for large commercial banks with temporarily surplus money that they are prepared to place against bills, subject to the bank's repurchasing them a few days later; the commercial bank concerned lends the dealer the money to finance the holding of the bill. Similar arrangements are also made by the Federal Reserve directly with bank dealers.

All member banks of the Federal Reserve System have direct access to the discount service of their Federal Reserve Bank, of which there is one in each of 12 districts. This is a privilege, however, and not a right. In the early years of the system the banks would sell discountable paper to the Federal Reserve, but now they usually borrow against a pledge of government securities held in safe custody with the Federal Reserve Bank in question. The Federal Reserve lends for a number of purposes but always at a time of general stress. It is assumed that, as the pressure abates, borrowing banks will repay their indebtedness as quickly as possible. Under ordinary conditions, the continuous use of Federal Reserve credit by a member bank over a considerable period is not regarded as appropriate.

In the latter half of the 1960s, the Federal Reserve subjected its discount arrangements to an intensive three-year scrutiny and decided to make them more flexible. Open-market operations were to remain the main tool of monetary policy, but the proposed changes in discount operations were expected to lead to a generally higher level of borrowing by the commercial banks. This higher level of borrowing would not mean a corresponding increase in total reserves, since the increased borrowing would tend to be offset by correspondingly smaller Federal Reserve purchases of securities in the open market.

Direct controls. The so-called classical techniques of credit control—open-market operations and discount policy—can only be employed where there is a sufficiently developed complex of markets in which to buy and sell assets of the type that commercial banks ordinarily hold. Direct credit controls have a wider range of application. They may be used either as a substitute for the classical techniques or as a supplement to them. Direct controls are more likely to be resorted to when the money market is undeveloped, because then a central bank can only impose its authority by means of direct action. This is often the situation facing a newly established central bank. Rather than wait for the slow evolution of a money market, the authorities may provide the central bank from the start (as in Pakistan, the Philippines, Sri Lanka [formerly Ceylon], and Malaysia) with very full powers to control the banking system.

The aim in imposing a direct, quantitative regulation of credit is to curb inflationary pressures that may result from an expansion of commercial bank lending. This can be done in four main ways: (1) the commercial banks may be required to maintain stated minimum reserve ratios of cash to deposits, a stated liquid assets ratio, or some combination of both; (2) part of the cash resources of the commercial banks may be immobilized at the discretion of the central bank; (3) ceilings may be imposed on the amount of accommodation to be made available to the commercial banks at the central bank (sometimes referred to as "discount quotas"); and (4) a ceiling may be prescribed for commercial bank lending itself.

Variations in reserves

Minimum reserve requirements. The variation of minimum cash reserve requirements as a direct means of

quantitative credit control has become increasingly general in recent years. The practice has largely derived from experience in the United States. In its origin the U.S. insistence on stated minimum reserve requirements for commercial banks was simply a means of prescribing minimum standards of sound behaviour. Only later did such ratios come to be seen as a useful supplementary quantitative credit control. The history has been similar elsewhere (as in India and New Zealand).

The power granted by the Banking Act of 1935 to the Federal Reserve System to vary the cash reserves of the commercial banks in the United States was employed for the first time during the boom of 1936–37, and periodic variation of minimum reserve requirements subsequently came to be recognized as an appropriate technique for controlling the money supply. The Federal Reserve Board's decisions were sometimes subject to considerable criticism, but, as it became more experienced in the use of this technique, variation in reserve requirements combined with other measures came to be regarded as a useful means of cushioning the economy against a recession. The variation of reserve requirements did not prove as effective in preventing inflation, largely because of the government's insistence that the Federal Reserve simultaneously support the prices of government bonds through open-market operations. This insistence was abandoned after an "accord" between the Treasury and the Federal Reserve Board in March 1951. Since then, much greater emphasis has been placed on the use of open-market operations, which had become more effective, and the variation of minimum reserve requirements as a means of controlling the credit base has diminished in the U.S. The technique is still widely used, however, in many countries.

In some countries the authorities require the maintenance of minimum liquid assets ratios. This is often combined with minimum requirements for cash reserves, as in India, Pakistan, and West Germany, though not always (in France, for example, until 1967 there were no minimum cash reserve requirements). Where prescribed minima relate to liquid assets and not to cash as such, reserves are held in the form of earning assets—an important distinction from the point of view of the commercial banks.

Cash ratios and special deposits in England

After 1946 the clearing banks in England (but not the Scottish banks) observed a more or less fixed cash ratio of 8 percent; the convention of a 30 percent minimum liquid assets ratio (of which the 8 percent was part) became officially recognized in 1955. It was lowered to 28 percent in 1963. A new element was introduced in 1960, when the Bank of England launched its system of "special deposits" as a means of reinforcing other methods of credit control. Calls were now made from time to time on the London clearing banks to deposit with the Bank of England by a specified date some specified percentage of their gross deposits; similar arrangements applied to the Scottish banks, but the calls were smaller. In 1971, however, a new system was introduced whereby the banks were required to keep a uniform minimum reserve ratio of 12.5 percent of their "eligible liabilities" (primarily sterling deposits of up to two years' maturity, including sterling certificates of deposit) and, when called upon to do so, to place special deposits with the Bank of England. The assets representing the 12.5 percent included balances with the Bank of England (other than special deposits), treasury bills, company tax reserve certificates, money at call with the London money market, local authority bills eligible for rediscount at the Bank of England, commercial bills, and government securities. They did not include cash in till. (Member banks in the United States were allowed to include "vault cash" in their minimum reserves after 1959.)

The use of variable minimum reserve requirements as a means of credit control can, if carried far enough, produce results, especially when the requirements include the holding of cash balances. It is more useful as an anti-inflationary weapon than as a means of countering recession, since it cannot overcome a possible unwillingness of the banks to lend or of their customers to borrow. It is a somewhat clumsy technique, however, and cannot make adequate allowance for the special needs of different institutions.

Immobilization of cash resources. A second group of direct quantitative credit controls involves keeping a portion of the cash resources of commercial banks immobilized at the discretion of the central bank. Two leading examples of this technique were the use of the Treasury Deposit Receipt (TDR) in Great Britain during and after World War II and the "special account procedure" adopted in Australia in 1941. Both were means of immobilizing the increased liquidity deriving from wartime government expenditure.

The direct issue of Treasury Deposit Receipts at a nominal rate of interest to banks in the United Kingdom began in July 1940. They were not negotiable in the market nor transferrable between banks, but they could be tendered in payment for government bonds (and tax certificates); hence, during the war years they had a limited degree of liquidity. The Bank of England communicated to the banks collectively the amount of the weekly call, which was divided among them in proportion to their deposits. After the war, TDR's were replaced by treasury bills; in order to reduce the consequent high liquidity of the banks, there was a "forced funding" of £1,000,000,000 of treasury bills in November 1951, which were required to be invested in Serial Funding Stocks.

Special accounts in Australia

The special account procedure introduced in Australia in 1941 had a similar objective. The surplus investable funds of the Australian trading banks, defined as the amount by which each bank's total assets in Australia at any time exceeded the average of its total assets in Australia in August 1939, were required to be placed in special deposit accounts with the Commonwealth Bank (then the central bank) at a nominal rate of interest. A bank was not to withdraw any sum from its special account except with the consent of the Commonwealth Bank; during the war years, the bank generally directed the trading banks to lodge in their special accounts each month an amount equal to the increase in their total assets in Australia during the preceding month, although as a rule a lodgment was not required if it was known that a rise in assets would be followed by an early fall. Legislation in 1945 adopted the special account procedures as a means of general credit control (*e.g.*, to curb inflation), but the provisions were made more flexible. In 1953 a more complicated formula was introduced, and in 1960 the system was abandoned in favour of minimum reserve ratios.

Accommodation ceilings. Some countries have tried placing a limit on the amount of accommodation that the central bank may make available to the commercial banks. The difficulty in this type of quantitative credit control is to make it effective while at the same time allowing for changes in the economy; its most obvious use is as a means of checking inflation, but if the upward pressures on prices are strong, there is a temptation to increase the ceilings so that the restraint then becomes little more than a temporary check.

Usually, it is only when a control begins to be felt and to affect bank profits that the banks become really sensitive to changes in credit policy and the implementation of the control becomes truly effective. The postwar experience of France is a case in point. *Plafonds*, or "ceilings," were first introduced in France in 1948. Rediscount ceilings (or discount quotas) were fixed for each bank, though some categories of paper were excluded. Ceilings could be increased or (after 1957) reduced.

The French *plafonds*

From the authorities' point of view, the chief difficulty in operating this control was the persistent building up of pressure against the ceilings. This was met partly by upward revisions in the ceilings themselves and partly by instituting a number of safety valves. The degree of elasticity required constituted the chief weakness of the ceiling technique. The central bank was constantly under pressure to adjust the ceilings upward. Some upward revisions were unavoidable, but the problem was to decide which claims were legitimate and which not. Much bilateral bargaining took place between the Bank of France

and individual commercial banks, but the banks continued to complain that the strictness of the control was excessive and that the technique lacked flexibility.

The inadequacies of the *plafonds* technique in its original form became apparent when prices began to rise rapidly during the Korean War boom, and even the built-in safety valves failed fully to accommodate the pressures on bank liquidity. The need to strengthen the mechanism was obvious, and this was attempted in 1951. Previously, rediscounts had frequently exceeded the ceilings during the course of the month and were only brought within the *plafonds* by special action (*e.g.*, by purchase through the open market). The situation was brought under control by introducing a secondary ceiling to which a penalty rate of interest was applied. This was extended in 1958 to permit rediscounts even beyond the secondary ceiling, provided a further penalty was paid; each application, however, was scrutinized by the Bank of France. The system lasted until about the spring of 1964, though it did not finally disappear until 1968, when it was largely replaced by Bank of France operations in the open market. After early 1967, the banks were also subject to minimum reserve requirements.

Plafonds have also been employed in West Germany. They were introduced in 1952 and greatly strengthened in 1955. Quotas may be reduced from time to time (after 1964 they were also used to discourage institutions from borrowing abroad). Again there were safety valves (although less generous than in France) and the possibility of extra accommodation (Lombard credits) at a higher rate. In certain circumstances, supplementary quotas might be approved for periods of up to six months. A bank might also raise funds through the money market, though very likely at higher cost.

General ceilings on credit. Attempts have been made to prescribe a general ceiling within which the quantity of commercial bank lending must be held. This is even more difficult to achieve. One example of an attempt to place a ceiling on bank loans was the adoption of a "rising ceiling" by Chile in 1953. All banks were required not to expand the volume of their loans to businesses and individuals by more than 1.5 percent a month, using as their basis the average of a bank's advances on selected dates in 1953. Certain types of loans were forbidden, and bank resources were to be directed to productive and distributive activities that really contributed to the expansion of the national economy. The banks were also placed under an obligation to provide information on the destination of their loans. In succeeding years, adjustments were made on several occasions in the maximum permitted credit increase, expressed either as a percentage of advances or sometimes simply as a total for the banking system as a whole. In 1959 all quantitative credit restrictions were removed and banks were permitted to advance funds up to their financial capacity, provided they operated within the general banking law. There was no evidence that the controls had been effective, but the major problem in Chile was budgetary rather than monetary. A temporary ceiling on loans was imposed by agreement in Canada in 1951–52; in The Netherlands in 1957–58; and in France in 1958–59.

Ceilings on bank advances in Britain

Great Britain has had the most experience with this type of ceiling, introducing it as a temporary measure in 1955, when the banks were asked to bring their advances down by an average of 10 percent. Later an attempt was made to impose a true ceiling, requiring that bank advances not exceed the average of the period October 1956 to September 1957. This ceiling was continued until July 1958. Again, in 1961, the authorities indicated that the banks must aim at checking the rate of rise in bank advances; this came to be interpreted as a request that the level of advances at the end of 1961 be no higher than in the previous June. The banks were also not to encourage an increase in the volume of commercial bills. The request was modified in May 1962 and largely withdrawn in October, but it was made again in May 1965, when the clearing banks were requested not to increase their advances to the private sector at an annual rate of more than about 5 percent in the 12 months to mid-March 1966 (likewise with commercial bills). Other financial institutions were requested to observe a comparable degree of restraint. For 12 months after March 1966, advances and discounts, allowing for seasonal factors, were not permitted to rise above the levels set for March 1966. This represented an intensification of the credit squeeze because prices were rising. The credit restriction led to a falling off in business confidence, and, consequently, toward the end of 1966 bank lending was well below the official ceiling. In April 1967 the authorities announced a change in techniques, with an emphasis on making calls to special deposits (see above), but the ceilings returned again in November 1967. There was to be no increase in bank advances to the private sector (excluding exports and shipbuilding) except for seasonal reasons. In May 1968 a new ceiling was instituted for all such lending (including that for exports and shipbuilding); the clearing banks were now asked to restrict the total of this lending, after seasonal adjustment, to 104 percent of the November 1967 figure, with priority to be given to finance for exports and for activities directly related to improving the balance of payments. The restrictions also extended to other types of credit. Credit became even tighter (in March 1969) when the ceiling was reduced to 98 percent of the November 1967 level. The banks experienced considerable difficulty in meeting this new requirement and agreed merely to "do their best." Advances increased above the ceiling, and as a penalty the interest paid by the Bank of England on special deposits was halved. Not until late autumn 1969 did it become clear that the authorities were prepared to abandon their long campaign to get bank loans down to the target figure. By 1971 the ceiling had been replaced by the new reserve requirements described above. The system of quantitative credit control requires, for its successful implementation, the full cooperation of the banking community. In the United Kingdom, where banks base much of their lending on the overdraft technique under which customers can overdraw their accounts up to prearranged limits, the system was very unpopular.

Selective controls on credit

In addition to regulating the quantity of credit, central banks have sometimes attempted to influence the directions in which the commercial banks lend. A loose system of control prevailed in the United Kingdom during World War II and afterward, based initially on directives from the Capital Issues Committee and later on requests from the Bank of England. A highly formalized technique was employed in Australia during the war and earlier postwar years; detailed and specific instructions were given to the trading banks, marginal cases being referred to the central bank. The system of Voluntary Credit Restraint in the United States in 1951 was similar. The more formal controls seemed to be no more effective than the looser system employed in the United Kingdom.

Selective controls have been imposed on consumer installment finance in the United States and elsewhere (*e.g.*, by stipulating the percentage of deposit required and the length of the term over which repayments may be made). In the U.S., under the Securities Exchange Act of 1934, the Federal Reserve can vary the margins that purchasers of securities must pay in cash, thereby limiting the amount of credit available for this purpose.

BIBLIOGRAPHY

History of banking: JOHN H. CLAPHAM, *The Bank of England*, 2 vol. (1944); W.F. CRICK and J.E. WADSWORTH, *A Hundred Years of Joint Stock Banking* (1936); E.A. FEAVEARYEAR, *The Pound Sterling*, 2nd ed. (1963); MILTON FRIEDMAN and A.J. SCHWARTZ, *A Monetary History of the United States, 1867–1960* (1963); T.E. GREGORY and ANNETTE HENDERSON, *The Westminster Bank Through a Century*, 2 vol. (1936); BRAY HAMMOND, *Banks and Politics in America, from the Revolution to the Civil War* (1957); L.W. MINTS, *A History of Banking Theory in Great Britain and the United States* (1945); R.D. RICHARDS, *The Early History of Banking in England* (1929, reprinted 1958); R.S. SAYERS, *Lloyds Bank in the History of English Banking* (1957); JOSEPH SYKES, *The Amalgamation Movement in English Banking, 1825–1924* (1926); A.P. USHER, *The Early History of Deposit Banking in Mediterranean Europe* (1943, reprinted 1967); J.G. VAN DILLEN, *History of the Principal Public Banks* (1964).

Principles of banking: Two books that provide good general coverage of the principles of banking and finance are J.G. GURLEY and E.S. SHAW, *Money in a Theory of Finance* (1960); and R.S. SAYERS, *Modern Banking*, 7th ed. (1967). Much useful information on the workings of the financial system is contained in GREAT BRITAIN, COMMITTEE ON THE WORKING OF THE MONETARY SYSTEM, *Report* (1959, commonly known as the "Radcliffe Report"); and the U.S. COMMISSION ON MONEY AND CREDIT, *Money and Credit: Their Influence on Jobs, Prices, and Growth* (1961).

Banking systems: For a general survey of banking systems throughout the world, see BENJAMIN H. BECKHART (ed.), *Banking Systems* (1954). A survey of the United Kingdom, the United States, and the British Commonwealth countries is J.S.G. WILSON, *Monetary Policy and the Development of Money Markets* (1966). A study of U.S. experience is C.R. WHITTLESEY, A.M. FREEDMAN, and E.S. HERMAN, *Money and Banking: Analysis and Policy*, 2nd ed. (1968). Other titles include: H.W. ARNDT and C.P. HARRIS, *The Australian Trading Banks*, 3rd ed. (1965); A.Z. ARNOLD, *Banks, Credit, and Money in Soviet Russia* (1937); BANK OF JAPAN, *Money and Banking in Japan* (1964); W.F. CRICK (ed.), *Commonwealth Banking Systems* (1965); GEORGE GARVY, *Money, Banking, and Credit in Eastern Europe* (1966); BRANKO HORVAT, "Yugoslav Economic Policy in the Post-War Period: Problems, Ideas, Institutional Developments," *American Economic Review*, suppl., 61:69–169 (1971); S.A. MEENAI, *Money and Banking in Pakistan* (1966); R.S. SAYERS (ed.), *Banking in the British Commonwealth* (1952) and *Banking in Western Europe* (1962); P.B. WHALE, *Joint Stock Banking in Germany* (1930); and J.S.G. WILSON, *French Banking Structure and Credit Policy* (1957). For current articles on banking in most countries, see the London monthly *The Banker*.

Central banking: For general discussions, see C.H. KISCH and W.A. ELKIN, *Central Banks*, 4th ed. (1932); M.H. DE KOCK, *Central Banking*, 3rd ed. (1954); and R.S. SAYERS, *Central Banking After Bagehot* (1957). For specific countries, see the BANK OF JAPAN, *The Bank of Japan: Its Organization and Monetary Policies*, 3rd ed. (1971); BANK FOR INTERNATIONAL SETTLEMENTS, *Eight European Central Banks* (1963); H.A. DE S. GUNASEKERA, *From Dependent Currency to Central Banking in Ceylon* (1962); GERHARD DE KOCK, *A History of the South African Reserve Bank, 1920–1952* (1954); E.P. NEUFELD, *Bank of Canada Operations and Policy* (1958); RESERVE BANK OF AUSTRALIA, *Reserve Bank of Australia*, 2nd ed. (1969); RESERVE BANK OF INDIA, *History of the Reserve Bank of India, 1935–51* (1970); and the U.S. BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM, *The Federal Reserve System: Purposes and Functions*, 5th ed. (1963).

(J.S.G.W.)

Baptists

Baptists are Protestant Christians who share the basic beliefs of most Protestants and first received their name from their insistence on baptizing believers only and on Baptism by immersion only rather than by sprinkling or pouring. (This view is, however, shared by others who are not Baptists.) While Baptists do not constitute a single church or denominational structure, most of them adhere to a congregational form of church government. Some Baptists lay stress upon having no human founder, no human authority, and no human creed.

In the late 1960s there were nearly 28,000,000 Baptists in the world, with the vast majority of them concentrated in the United States, where they constitute the largest Protestant community. The 27 Baptist bodies in the United States have an inclusive total of 24,500,000 members. Of these, the large majority are included in four major conventions: the Southern Baptist Convention, with nearly 10,500,000 members; the National Baptist Convention, U.S.A., Inc., with 5,500,000 members; the National Baptist Convention of America, with over 2,500,000 members; and the American Baptist Convention, with nearly 1,500,000 members. The multiplicity of Baptist groups in the United States is accounted for in part by the 19th-century controversy over slavery, in part by racial and nationality differences, and in part by divergence of opinion on questions of doctrine and organization. Baptists also have a basic suspicion of super-congregational ecclesiastical organizations as valid expressions of the church.

Outside the United States, major Baptist communities are found in the U.S.S.R. (545,000 members), India (508,000), Brazil (243,000), Zaire (228,000), Burma (223,000), and Canada (177,000).

HISTORY

Origins. Some Baptists believe that there has been an unbroken succession of Baptist churches from the days of John the Baptist and the Apostles of Christ. Others trace their origin to the Anabaptist movement (16th-century radical Protestant movement) on the European continent. While differing in their estimate of the possible Anabaptist influence, most scholars agree that Baptists as an English-speaking denomination originated within 17th-century Puritanism (a church reform movement that attempted to "purify" the remaining vestiges of Roman Catholicism from the Church of England) as an offshoot of Congregationalism. There were two major currents in early Baptist life: the Particular Baptists adhered to the doctrine of a particular atonement—that Christ died only for an elect—and were strongly Calvinist (following the Reformation teachings of John Calvin) in orientation; the General Baptists held to the doctrine of a general atonement—that Christ died for all men and not only for an elect—and represented the more moderate Calvinism of Jacobus Arminius (a 17th-century Dutch theologian who advocated the priority of divine grace and free will).

Particular Baptists and General Baptists

The two currents were also distinguished by a difference in churchmanship related to their respective points of origin. The General Baptists emerged from among the English Separatists (see below), whereas the Particular Baptists had their roots in non-Separatist independency (see below). Both the Separatists and the non-Separatists were congregationalist. They shared the same convictions with regard to the nature and government of the church. They believed that church life should be ordered according to the pattern of the New Testament churches, and to them this meant that churches should be self-governing bodies composed of believers only.

The point at which they differed was with regard to their attitude toward the Church of England. The Separatists took what is commonly described as a sectarian position; they contended that the Church of England was a false church and insisted that the break with it must be complete and uncompromising. The non-Separatists, more ecumenical in spirit, sought to maintain some bond of unity among Christians. While they believed that it was necessary to separate themselves from the corruption of parish churches, they also believed that it would be a breach of Christian charity to refuse all forms of intercourse and fellowship with them. While many non-Separatists withdrew and established a worship of their own, they would not go so far as to assert that the parish churches were devoid of the marks of a true church.

Most scholars find no evidence of decisive influence having been exerted upon the English Baptists by the continental Anabaptists, but they acknowledge that the General Baptist wing of the English Baptists exhibits Anabaptist influence at several minor points.

Growth in England and abroad. Although the Particular Baptists were to represent the major continuing Baptist tradition, the General Baptists were first in the field. In 1608 religious persecution had induced a group of Lincolnshire Separatists to seek asylum in Holland. One contingent settled in Amsterdam with one John Smyth (or Smith) a Cambridge graduate, as their minister; another moved to Leiden under the leadership of one John Robinson. When the question of Baptism arose during a debate on the meaning of church membership, Smyth came to the conclusion that, if the Separatist contention that "the churches of the apostolic constitution consisted of saints only" was correct, then Baptism should be restricted to believers only. This, he contended, was the practice of the first New Testament churches, for he could find no scriptural support for the baptizing of infants. Smyth published his views in *The Character of the Beast* (1609) and in the same year proceeded to translate them into action by baptizing first himself and then 36 others who joined him in forming a Baptist Church. Shortly thereafter. Smyth became aware of the

The question of Baptism

existence of a Mennonite (Anabaptist) community in Amsterdam and began to question his procedure in baptizing himself. Such an act could be justified, he concluded, only if no true church existed from which a valid Baptism could be obtained. After some investigation, Smyth arrived at the conviction that the Mennonites did constitute a true church, and he recommended union with them. This was resisted by Thomas Helwys and other members of the group, who returned to England in 1611 or 1612 and established a Baptist Church in London. The parent group in Amsterdam soon disappeared.

The Particular Baptists stemmed from a non-Separatist church that was established in 1616 by Henry Jacob at Southwark across the Thames from London. In 1638 a number of its members withdrew under the leadership of John Spilsbury to form the first Particular Baptist Church.

The two decades from 1640 to 1660 constituted the great period of early Baptist growth, for the Baptist preachers found their great opportunity to win adherents around the campfires of the Puritan leader Oliver Cromwell's army. The greatest gains were made by the Particular Baptists, and the General Baptists actually suffered numerous defections to the Quakers. After the Restoration of the Stuarts in 1660, both groups were subjected to severe disabilities, being forced to go underground until the Act of Toleration of 1689—in which the idea of a comprehensive Church of England was abandoned and "Nonconformists" were permitted to have their own places of worship—granted them a measure of relief.

During the following decades, the vitality of the General Baptists was drained away by the inroads of skepticism, and their churches generally dwindled and died or became Unitarian. The Particular Baptists took an opposite course, retreating into a defensive, rigid hyper-Calvinism that prevented any effective evangelism.

Survival and expansion of Particular Baptists

Among the Particular Baptists in England, renewal came as a result of the influence of the Evangelical Revival, a new surge of growth initiated by the activity of the English Baptist clergymen Andrew Fuller, Robert Hall, and William Carey. Carey, in 1792, formed the English Baptist Missionary Society—the beginning of the modern foreign missionary movement in the English-speaking world—and became its first missionary to India. A New Connection General Baptist group, Wesleyan in theology, was formed in 1770, and a century later, in 1891, it united with the Particular Baptists to form the Baptist Union of Great Britain and Ireland.

By the end of the 19th century, Baptists, together with the other Nonconformist churches, were reaching the peak of their influence in Great Britain, numbering among their preachers several men with international reputations. Baptist influence was closely tied to the fortunes of the Liberal Party, of which the Baptist David Lloyd George was a conspicuous leader. After World War I, English Baptists began to decline in influence and numbers.

Baptist churches were first established in Australia (1831) and New Zealand (1854) by missionaries of the English Baptist Missionary Society. In Canada, Baptist beginnings date from the activity of one Ebenezer Moulton, a Baptist immigrant from Massachusetts who organized a church in Nova Scotia in 1763; Baptist work there, and in the 13 Atlantic seaboard colonies, was nurtured by the Philadelphia Baptist Association (see below). In Ontario, the earliest Baptist churches were formed by United Empire Loyalists who crossed the border following the American Revolution, while other churches were established by immigrant Baptists from Scotland and by missionaries from Vermont and New York. The Baptists of Canada are united in the Baptist Federation of Canada.

Development in the United States. Baptist churches in the English colonies of North America were largely indigenous in origin, being the product of the leftward movement that was occurring among the colonial Puritans at the same time as that in England. While some migrated to the new world as Baptists, it was more typical for Baptist views to be adopted after arrival in the colonies, as happened in the case of Henry Dunster, the first president of Harvard, and Roger Williams.

Colonial period. The first Baptist Church in the American colonies was established at Providence in 1639 by Roger Williams shortly after his banishment from the Massachusetts Bay Colony. While Williams' general Calvinist theological position was roughly analogous to that of Spilsbury, prior to becoming a Baptist, he had adopted the narrower Separatist view of the church. Williams soon came to the conclusion that all existing churches, including that newly established at Providence, lacked a proper foundation, and that this defect could be remedied only by a new apostolic dispensation, in which new apostles, divinely commissioned, would appear to re-establish the true church.

The defection of Williams left the church with no strong leadership and thus made it possible for it to be reorganized on a General Baptist platform in 1652. There was scattered General Baptist activity throughout the colonies, but the only real cluster of General Baptists was in Rhode Island, where the churches formed themselves into an association, or yearly meeting, in 1670. The early General Baptists never gained great strength. Most of their churches decayed, and some, including the Providence church, were reorganized as Particular Baptist churches. The half-dozen churches that survived never entered the main stream of American Baptist life and exerted no real influence upon its development.

The earliest strong Particular Baptist centre in the colonies was at Newport, Rhode Island, where, between 1641 and 1648, a church that had been gathered by the physician and minister John Clarke adopted Baptist views. Except for a church that had a brief existence at Kittery, Maine, there were only two other Particular Baptist churches in New England for the better part of a century. One of these was at Swansea, Massachusetts, where a church was formed by a group of Welsh immigrants under the leadership of John Myles, who had previously attempted to found a church near Plymouth but had been arrested, tried, and fined on a charge of conducting a public meeting without having first obtained permission to do so; the other was organized at Boston in 1665. Another Particular Baptist church was established at Charleston, South Carolina, in 1683 or 1684.

The great centre of Particular Baptist activity in early America was in the Middle Colonies. In 1707 five churches in New Jersey, Pennsylvania, and Delaware united to form the Philadelphia Baptist Association, and through the association they embarked upon vigorous missionary activity. By 1760 the Philadelphia association included churches located in the present states of Connecticut, New York, New Jersey, Pennsylvania, Delaware, Virginia, and West Virginia; and by 1767 further multiplication of churches had necessitated the formation of two subsidiary associations, the Warren in New England and the Ketochton in Virginia. The Philadelphia association also provided leadership in organizing the Charleston Association in the Carolinas in 1751, and this in turn fostered the formation of the Kehukee Association in North Carolina in 1765.

Philadelphia Baptist Association

While this intercolonial Particular Baptist body provided leadership for the growth that characterized American Baptist life during the decades immediately preceding the American Revolution, that growth was largely a product of an 18th-century religious revival known as the Great Awakening. Though they participated directly in the Awakening only during its last phase in the South, Baptists attracted large numbers of recruits from among those who had been "awakened" by the preaching of others. In addition to strengthening and multiplying the "regular" Baptist churches, the Awakening in New England produced a group of revivalistic Baptists, known as Separate Baptists, who soon coalesced with the older New England Baptist churches. In the South, however, they maintained a separate existence for a longer period of time. Shubael Stearns, a New England Separate Baptist, migrated to Sandy Creek, North Carolina, in 1755 and initiated a widespread revival that quickly penetrated the entire Piedmont region. The churches he organized

were brought together in 1758 to form the Sandy Creek Association. Doctrinally, these churches did not differ from the older "regular" Baptist churches, but what the older churches saw as their emotional excesses and ecclesiastical irregularities created considerable tension between the two groups. By 1787, however, a reconciliation had been effected.

In several of the colonies, Baptists laboured under legal disabilities of varying severity. The public whipping of one, Obadiah Holmes, in 1651 for refusing to pay a fine imposed for holding an unlawful meeting in Lynn, Massachusetts, caused John Clarke to write his *Ill News from New England* (1652). Fourteen years later Baptists of Boston were fined, imprisoned, and denied the use of a meetinghouse they had erected. Payment of taxes for support of the established church was a cause of continuing controversy in New England, while the necessity to secure licences to preach became an inflammatory issue in Virginia.

In the 19th century. The problem of travel had made it difficult for the Philadelphia association to serve as a bond uniting Baptists, and the rapid multiplication of churches made it impossible. It has been estimated that immediately before the American Revolution there were 494 Baptist congregations; 20 years later, in 1795, Isaac Backus estimated the number at 1,152. The initial expedient of the Philadelphia association had been to organize subsidiary associations, but during the war the churches, left to their own devices, proceeded to organize independent associations. By 1800 there were at least 48 local associations, and the great problem was to fashion a national body to unite the churches. The final impetus in this direction came from an interest in foreign missions. The first missionaries of the newly organized Congregational mission board were Adoniram Judson and Luther Rice, who had been sent to India. On shipboard they became convinced by a study of the Scriptures that only believers should be baptized. Upon arrival at Calcutta, Judson went on to Burma, while Rice returned home to enlist support among American Baptists. As a result of Rice's efforts, a General Convention of the Baptist denomination was formed in 1814. Its scope was almost immediately broadened to include, in addition to the foreign mission interest, a concern for home missions, education, and the publication of religious periodicals. In 1826 the General Convention once again was restricted to foreign mission activities, and in the course of time it became known as the American Baptist Foreign Mission Society. Other denominational interests were served by the formation of additional societies with similar specialized concerns, such as the American Baptist Home Mission Society and the American Baptist Publication Society.

Formation of General Convention

The unity that was achieved through these societies was partially disrupted as a result of the slavery controversy. During the decade prior to 1845, various compromises between the proslavery and antislavery parties in the denomination were attempted, but they proved to be unsatisfactory. As a result, a Southern Baptist Convention was organized at Augusta, Georgia, in 1845. Although its constitution provided for boards of home and foreign missions, education, and publication, its energies were devoted largely to foreign missions. Consequently, the American Baptist Home Mission Society and the American Baptist Publication Society continued to operate in the South after the Civil War and enjoyed a large measure of support from the churches. Toward the close of the 19th century, however, the Southern Baptist Convention began to develop its own home mission and publication work and to protest the intrusion of the older societies in the South. The final separation between Baptists of South and North was formalized in 1907 by the organization of the Northern Baptist Convention (after 1950 called the American Baptist Convention), which brought together the older societies and accepted a regional allocation of territory between the northern and southern conventions.

Development of Negro churches. Negro churches constitute an important segment of American Baptist life. Following the Emancipation Proclamation (1863)—an edict freeing the slaves of the United States—and the close of the Civil War, Negro Baptists began to organize their own churches. A state convention of Negro Baptist churches was formed in 1866 in North Carolina, and in 1880 the National Baptist Convention of America was organized. A dispute over the control of property and publications led to a division in 1916. The smaller of the two factions retained the original name, while the larger body became the National Baptist Convention, U.S.A., Inc.

National Baptist Convention

Developments in education. From the beginning, American Baptists have displayed an interest in an educated ministry, and their interest in higher education increased steadily as they grew in numbers. The Philadelphia association in the 18th century collected funds to help finance the education of ministerial candidates. Hopewell Academy was established in 1756, and in 1764 Brown University was founded in Rhode Island midway between Nova Scotia and Georgia. Eight other institutions were established before 1825, 25 more were established between 1825 and 1850, 39 between 1851 and 1875, and 70 between 1876 and 1900. The educational advance culminated in 1891 in the founding of the University of Chicago, which was intended to be a great national Baptist superuniversity that would tie together the smaller Baptist colleges.

In the North, regional education societies were the usual channels through which support was given to education, while in the South the institutions more often were sponsored directly by state conventions or by the Southern Baptist Convention. After the Civil War, the American Baptist Home Mission Society established a number of Negro Baptist colleges in the South. These came to be administered by Negro boards of trustees, with the cooperation of the American Baptist Home Mission Society and the Board of Education and Publication of the American Baptist Convention.

During the 20th century. After 1900 Baptists were troubled by theological controversies that led to the formation of several new Baptist groups. Some of the tensions arose over questions of structure of church organization, some arose over refusals to adopt an authoritative creedal statement, and some were the product of dissatisfaction with the affiliation of the American Baptist Convention with interdenominational and ecumenical bodies. Questions of organizational structure were involved in the formation of the American Baptist Association in 1905 by churches located primarily in Oklahoma, Texas, and Arkansas. Two other groups were products of the Fundamentalist controversy: the General Association of Regular Baptist Churches, organized in 1932; and the Conservative Baptist Association of America (1947).

A phenomenon of the post-World War II period was the abandonment by the Southern Baptist Convention of its regional limitations. Because of increasing mobility of population, the Southern Baptist Convention felt it necessary to follow its members to the growing urban centres of the North and West; by the second half of the century Southern Baptist churches were to be found in almost every part of the United States.

Growth outside the U.S. While Baptists have been troubled by divisive tendencies during the 20th century, a parallel tendency has been toward greater unity and cohesiveness through the Baptist World Alliance. The 19th century was the great period of Baptist missionary endeavour. The penetration of Asia was led by William Carey in India, Adoniram Judson in Burma, and Timothy Richard in China, and by the late 1960s there was a Baptist community of nearly 1,000,000 adherents in Asia, chiefly in India, Burma, and mainland China. The initial Baptist presence in Africa began in 1793 when David George, a former slave from South Carolina, reached Sierra Leone by way of Halifax, Nova Scotia. More organized activity was initiated in 1819 by Negro Baptists of Richmond, Virginia, who sent Lott Cary to Sierra Leone in 1821 and then shifted his base of operations to Liberia in 1824. By 1970 Baptists in Africa numbered a half million communicants, with major concentrations in the Republic of Zaire, Nigeria, and Camer-

oon. Of later origin is the Baptist community of 600,000 in Latin America.

The pioneer Baptist on the Continent of Europe was Johann Gerhardt Oncken, who organized a church at Hamburg in 1834. Oncken had become acquainted with Barnas Sears of Colgate Theological Seminary, who was studying in Germany, and with six others he was baptized by Sears. From this centre, evangelistic activity was extended throughout Germany, and missions were established in Austria, Hungary, Romania, Bulgaria, Switzerland, Belgium, The Netherlands, Denmark, Poland, and Russia. Baptist activity was initiated independently in France, Italy, and Spain. Swedish Baptist beginnings date from the conversion of Gustaf W. Schroeder, a sailor baptized in New York in 1844, and Frederick O. Nilsson, also a sailor, who was baptized by Oncken in 1847. From Sweden, Baptists penetrated Norway and Finland. Excluding the British Isles but including the U.S.S.R., there were nearly 870,000 European Baptists in the late 1960s.

It was this expansion of the Baptist community in Asia, Africa, Latin America, and Europe that led to the formation of the Baptist World Alliance at London in 1905. The purpose of the alliance is to provide mutual encouragement, exchange of information, coordination of activities, and consciousness of the larger Baptist fellowship. Periodic world congresses are held, and a headquarters secretariat is maintained in Washington and London.

Developments in Russia

The most notable growth occurred in Russia, where a Russian Baptist Union was formed in 1884 as the result of influences stemming from Oncken. Another Baptist body, the Union of Evangelical Christians, was organized in 1908 by a Russian who had come under the influence of English Baptists. Persecution of Baptists, which had been severe, was relaxed in 1905, and within the remaining disabilities a moderate growth occurred. The Revolution of 1917, with its proclamation of liberty of conscience, ushered in a period of astonishing advance: by 1927 the Russian Baptist Union numbered some 500,000 adherents, while the Union of Evangelical Christians embraced more than 4,000,000. The Soviet constitution of 1929 subjected them to pressure once again, however, and in the late 1960s the two groups, which had combined in 1944 to form the All-Union Council of Evangelical Christians and Baptists in the U.S.S.R., reported 545,000 baptized believers.

ORGANIZATION, WORSHIP, AND DOCTRINE

Organization. Baptists insist that the fundamental authority, under Christ, is vested in church life in the local congregation of believers, which admits and excludes members, calls and ordains pastors, and orders its common life in accord with what it understands to be the mind of Christ. These congregations, which are manifestations of the whole church of Christ, are linked together in cooperative bodies, to which they send their delegates or messengers—regional associations, state conventions, and national conventions. The larger bodies, it is insisted, have no control or authority over a local church; they exist only to implement the common concerns—missionary, educational, philanthropic—of the local churches.

The pattern of organization of the local church has been undergoing change during the 20th century. Traditionally, the pastor was the leader and moderator of the congregation; more recently, there has been a tendency to regard him as the employed agent of the congregation and to elect a lay moderator to act in his stead at corporate meetings of the church. Traditionally, the deacons' functions were to assist the pastor and to serve as agents to execute the will of the congregation in matters both temporal and spiritual; more recently, there has been a tendency to multiply the number of church officers by the creation of boards of trustees, boards of education, boards of missions, and boards of evangelism. Traditionally, decisions were made by the congregation in a church meeting; more recently, church meetings have become less and less frequent, and there has been a tendency to delegate decision making to various boards. The relationship of local churches to the cooperative bodies has been undergoing similar change, and this has occasioned continuing discussion among all Baptist groups.

Worship. Baptist worship is hardly distinguishable from the worship of the old Puritan denominations (Presbyterians and Congregationalists) of England and the United States. It centres largely around the exposition of the Scriptures in a sermon, and an emphasis upon extemporaneous, rather than set, prayers. Apart from the centrality of the sermon, hymn singing is one of the most characteristic features of worship. Communion, received in the pews, is customarily a monthly observance. Baptism is by immersion.

Doctrine. *History.* Initially, Baptists were characterized theologically by strong to moderate Calvinism. The dominant continuing tradition in both England and the United States was Particular Baptist. By 1800 this older tradition was beginning to be replaced by evangelical doctrines fashioned by the leaders of the evangelical revival in England and the Great Awakening in America and further elaborated by subsequent New England divines and frontier revivalists. By 1900 the older Calvinism had almost completely disappeared, and Evangelicalism was dominant. The conciliatory tendency of Evangelicalism and its almost complete preoccupation with heart-religion and the conversion experience largely denuded it of any solid theological structure opening the door to a new theological current, which in its later phases became known as Modernism. Modernism, which was an attempt to adjust the Christian faith to the new intellectual climate, made large inroads among the Baptists of England and the United States during the first two decades of the 20th century, and Baptists provided many outstanding leaders of the movement, including Shailer Mathews and Harry Emerson Fosdick. To many, these views seemed to pose a threat to the uniqueness of the Christian revelation, and they precipitated a counterreaction that became known as Fundamentalism (a movement emphasizing biblical literalism).

As a result of the controversy that followed, many Baptists developed a distaste for theology and became content to find their unity as Baptists in promoting denominational enterprises. By 1950 both Modernists and Fundamentalists were becoming disenchanted with the positions into which they had been forced in the heat of controversy, and it was from among adherents of both camps that a more creative theological encounter began to take place. While the majority of Baptists remained non-theological in their interests and concerns, there were multiplying signs that Baptist leadership was increasingly recognizing the necessity for renewed theological inquiry.

Contents. The unity and coherence of the Baptists is to be found in six distinguishing, although not necessarily distinctive, convictions they hold in common.

Central aspects of Baptist doctrine

1. The supreme authority of the Bible in all matters of faith and practice. Baptists are a non-creedal people, and their ultimate appeal always has been to the Scriptures rather than to any confession of faith that they may have published from time to time to make known their commonly accepted views.

2. Believer's Baptism. This is the most conspicuous conviction of Baptists. They hold that if Baptism is the badge or mark of a Christian, and if a Christian is one in whom faith has been awakened, then Baptism rightly administered must be a Baptism of believers only. Furthermore, if the Christian life is a sharing in the life, death, and resurrection of Christ, if it involves a dying to the old life and a rising in newness of life, then the act of Baptism must speak in these terms. The sign must be consonant with that which it signifies. It is for this latter reason that Baptists were led to insist upon immersion as the apostolic form of the rite, in addition to their initial insistence upon the Baptism of believers only.

3. Churches composed of believers only. Baptists reject the idea of a territorial or parish church and insist that a church is composed only of those who have been gathered by Christ and who have placed their trust in him. Thus, the membership of a church is restricted to those who—in terms of a charitable judgment—give clear evidence of their Christian faith and experience. The basis

of their church life is a church covenant wherein they covenant with God and one another to walk together in Christian obedience.

4. Equality of all Christians in the life of the church. By the doctrine of the priesthood of all believers, Baptists not only understand that the individual Christian may serve as a minister to his fellows, but they believe that it also confers upon each member of a church equal rights and privileges in determining the affairs of the church. The church officers—pastor and deacons—have special responsibilities, derived from the consent of the church, which only they can discharge, but they do not have a priestly unique status.

5. Independence of the local church. By this principle Baptists affirm that a properly constituted congregation is fully equipped to minister Christ and need not derive its authority from any source, other than Christ, outside its own life. Baptists, however, have not generally understood that a local church is autonomous in the sense that it is isolated and detached from other churches. The local church is but one manifestation of the Catholic Church, and as individual Christians are bound to pray for one another and to maintain communion with one another, so particular churches are under like obligation. Thus, the individual churches testify to their unity in Christ by forming associations and conventions through which they can seek counsel and advice and cooperate in common concerns.

6. Separation of church and state. From the time of Smyth, Baptists have insisted that a church must be free to be Christ's church, determining its own life and charting its own course in obedience to Christ without outside interference. Thus Smyth asserted that the

> magistrate is not by virtue of his office to meddle with religion or matters of conscience, to force and compel men to this or that form of religion or doctrine, but to leave Christian religion free to every man's conscience.

Baptists were in the forefront of the struggle for religious freedom in both England and America, they cherish the liberty established in early Rhode Island, and they played an important role in securing the adoption of the "no religious test" clause in the U.S. constitution and the guarantees embodied in the First Amendment.

Few Baptists have been willing to become so sectarian as to deny the Christian name to other denominations. With the exception of the Southern Baptists, the vast majority of Baptists cooperate fully in interdenominational and ecumenical bodies, including the World Council of Churches.

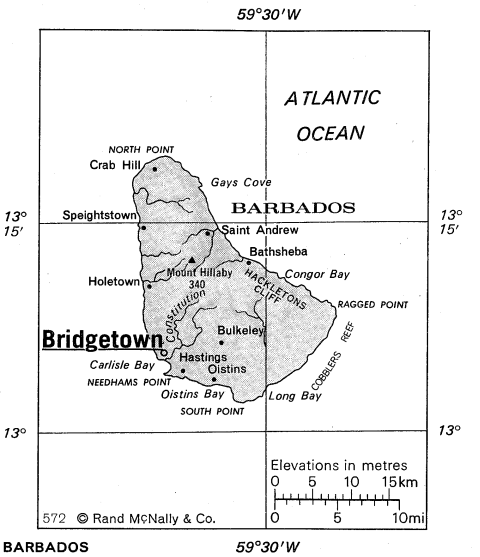
BIBLIOGRAPHY. R.G. TORBET, *A History of the Baptists*, rev. ed. (1963), is the most complete account of the Baptists. A.C. UNDERWOOD, *A History of the English Baptists* (1947), gives major attention to Baptist beginnings. W.L. LUMPKIN (ed.), *Baptist Confessions of Faith* (1959), is a compilation of documents. W.S. HUDSON, *Baptist Concepts of the Church* (1959), provides useful perspectives and interpretations. N.H. MARING and W.S. HUDSON, *A Baptist Manual of Polity and Practice* (1963), gives details of ecclesiastical organization. Additional references may be found in E.C. STARR (ed.), *A Baptist Bibliography* (1947–); statistical information in *The Baptist Handbook, 1969.*

(W.S.H.)

Barbados

Barbados is an independent island in the West Indies situated about 100 miles (160 kilometres) east of the Windward Islands. Approximately triangular in shape, it has an area of 166 square miles (430 square kilometres) and a population of about 240,000. Its capital is Bridgetown, the only seaport.

Strictly speaking, Barbados does not form part of the Lesser Antilles island chain, although it is sometimes grouped with this archipelago. Its geographic position has profoundly influenced the island's history and culture. In the era of sailing ships, the prevailing winds made the island difficult of access; if outward-bound ships from Europe did not gain the island while heading west, they found it difficult indeed to turn and reach its shores by sailing eastward. As a result of this circumstance, conquest of the island was impracticable, and it remained a British possession without interruption from the 17th century to November 30, 1966, when it attained independence and membership in the Commonwealth of Nations. It has also developed a proto-British culture, thus earning the nickname Bimshire, or Little England; to many "Bajans" (as Barbadians sometimes refer to themselves) the fact that Barbados lies about 4,750 miles southwest of London is more significant than the fact that it is about 300 miles northeast of Venezuela. Its many British customs, however, provide an attraction for tourists; tourism has strengthened the island's economy at a time when the production of sugar—its main crop—has been declining.



The landscape. *Geology.* The rocks underlying Barbados consist of sedimentary deposits, including thick shales (laminated sediments in which the particles are predominantly of clay), clays, sands, and conglomerates, laid down approximately 70,000,000 years ago. Above these are chalky deposits, which were capped with coral before the whole island rose to the surface. A layer of coral, up to 300 feet thick, now covers the island, except in the northeast (a physiographic region known as the Scotland District) where erosion has stripped off the coral cover.

The coral surface

Relief. Mt. Hillaby, the highest point in Barbados, rises to 1,115 feet in the north central part of the island. To the west the land drops down to the sea in a series of terraces, many of which are green and gently rolling. East from Mt. Hillaby, the land declines sharply to the rugged upland of Scotland District. Southward, the highlands again descend steeply to the broad St. George River Valley; between the valley and the sea the land rises to 400 feet to form Christ Church Ridge. Coral reefs surround most of the island.

Climate. The climate is generally a pleasant one. The temperature does not usually rise above 86° F (30° C) or fall below 72° F (22° C). There are two recognized seasons of the year: the dry season, from early December till May, and the wet season, which lasts the rest of the year. The average rainfall is about 60 inches a year, but, despite the smallness of the island, it varies markedly from district to district. Barbados lies in the southern border of the West Indian hurricane zone. Hurricanes in 1780 and 1831 caused great devastation, and others of lesser intensity have visited the island—including one in 1955 that was the first in 57 years.

Vegetation and animal life. Only about 40 acres of the original vegetation remain; the pale green of sugarcane is the characteristic colour of the landscape. Tropical trees—including the flamboyant tree, or poinciana, and the frangipani—are widespread, while flowering shrubs adorn parks and gardens.

The few existing wild animals—such as monkeys, hares,

raccoons, and mongooses—constitute a pest to local cultivators. Resident birds include the dove, hummingbird, and sparrow. The waters are alive with dolphins, barracuda, bream, and sprats (small European herring), and the tree frog punctuates the night with its constant piping or whistling.

The landscape under human settlement. Since over four-fifths of the farmland is the property of estates or of big landowners, there is little land that the poor can own. As a result, "tenantries" are much more common than villages. These are clusters of wooden houses—locally known as chattel houses—which are to be seen on the borders of the large estates; they are usually owned by the occupier but stand on rented ground from which they may easily be removed. Some of them have electric light; toilet facilities are outdoors, and water is obtained from a village pump, locally called a standpipe. Sheep, goats, and cows graze nearby. Children and young men often lounge around, talking, playing cards or cricket, and, on Sunday mornings, cutting each other's hair. A "cum rum" (a grocer's shop with a license for the sale of rum) is normally the only social amenity and is shared by tenantry and village alike. The villages are usually rows of houses—including the stone houses of the middle class—aligned along the main roads.

The only sizable town is Bridgetown, the national capital, which had a metropolitan population of about 97,000 in 1970. In its commercial and administrative centre, new multistoried buildings are replacing the 19th-century town, which once resembled a Victorian print. Donkey carts, bicycles, and old wooden buses, however, still dispute their right-of-way with English and American cars, while old women hucksters sit on the pavement selling mangoes and straw hats.

History. Although most of the island's early history is unknown, existing remains indicate that there was once a considerable population of Arawak Indians. The first contact with Europe probably came in 1518, when the Spanish landed to seek slaves for their colony of Hispaniola. By the mid-16th century no Indians appear to have remained, and Spanish claims to the territory had lapsed.

When the first English colonists landed in the early 17th century, their settlement was uncontested. Barbados became the third West Indian island to be occupied by the British—Bermuda having been occupied in 1612, and St. Kitts in 1623. The settlers brought in slaves to supply the large amount of labour needed for the production of sugarcane.

Effects of slavery

The decision in favour of slavery had momentous social consequences. In 1640 the population numbered 30,000, of whom only a few hundred were slaves. Five years later there were 6,000 slaves on the island, and by 1685 there were 46,000. The white population, however, began to decline; by 1700 it was reduced to 12,000—about what it was at the start of the 1970s.

Slavery was abolished in 1834. On other Caribbean islands emancipation all but ruined the sugar industry. In Barbados, however, sugar fared better, partly because there were fewer absentee landlords and partly because there were fewer places for the freed slaves to go. The central problem, however, was how to create a new community of free men at a time when the government represented the interests of a tiny oligarchy of planters and merchants. No immediate solution was found, but during the next 100 years democratic processes were gradually introduced.

At the end of the 1930s, rapid population growth and the worldwide economic depression led to rioting throughout the British West Indies. The Moyne Commission was sent by the British Parliament in 1937 to investigate. The enfranchisement of women, adult suffrage, the ministerial system, and the development of trade unionism and political parties were the fruits of this epoch. Barbados housed the headquarters of the Development and Welfare Organization for the West Indies from 1942 to 1958 and produced the first federal prime minister of the Federation of the West Indies in 1958. In 1965, however, Barbados broke away from the federation. On November 30, 1966, Barbados achieved independence, remaining a member of the Commonwealth of Nations at its own request.

Population. The population in 1970 numbered over 238,000, having increased annually at an average rate of less than 3 percent between World War II and 1960 and remained at a virtual standstill between 1960 and 1970 because of emigration losses. The density of over 1,400 persons per square mile was very high for a rural community. About 90 percent of the population was black, about 4 percent white, and about 6 percent mixed. There were about 4 percent more women than men. Marriage is statistically uncommon in the island, and more than 70 percent of the population is illegitimate.

The religion is Christian, with Anglicans accounting for more than half the population. Protestant sects combined make up nearly a quarter of the rest, while the Roman Catholic Church accounts for less than 3 percent.

Barbados, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census§
City				
Bridgetown	*	*	11,000	9,000
Parishes				
Christ Church	22	57	33,000	36,000
St. Andrew	14	36	8,000	7,000
St. George	17	44	17,000	17,000
St. James	12	31	14,000	15,000
St. John	13	34	11,000	11,000
St. Joseph	10	26	9,000	8,000
St. Lucy	14	36	9,000	9,000
St. Michael	15*	39*	82,000	88,000
St. Peter	13	34	11,000	11,000
St. Philip	23	60	17,000	17,000
St. Thomas	13	34	10,000	11,000
Total Barbados	166	430†	232,000‡	238,000‖

*Bridgetown is included in St. Michael. †Converted area figures do not add to total given because of rounding. ‡Census total excludes adjustment for underenumeration. Adjusted census total: 233,000. §Preliminary. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

Economy. Traditionally, sugar and its by-products—rum and molasses—have dominated the economy, constituting in 1970 about 50 percent of its domestic exports. In recent years, efforts have been made to diversify the economy and to provide a greater degree of employment. Fishing has been encouraged, and a number of light industries have sprung up, including food processing, textiles, the manufacturing of glass bottles and furniture, and local handicrafts. It is tourism, however, that is most likely to change the economy. Since 1968 the tourist trade has brought in more foreign currency than the production of sugar. Between 1965 and 1970 the amount of money spent by tourists more than tripled, land prices rose considerably, and numerous new hotels and apartment houses were built.

The tourist industry

The sugar industry, on the other hand, has been undergoing one of the recurrent crises which have marked its long history. It is short of the capital necessary to mechanize its operations, while strikes, malicious cane fires, and the growing disinclination of Barbadians to work on the land prevented the industry on three occasions in the 1960s even from fulfilling the annual quota of 163,000 tons for Britain and Canada. The situation is further complicated by a sharply rising cost of living and a national debt that quadrupled during the decade of the 1960s. Foreign trade—both import and export—is mainly conducted with the United Kingdom, the United States, and Canada. Commerce with fellow members of the Organization of American States is rising annually.

The monetary unit is the East Caribbean dollar. The government plans, however, to withdraw from the East Caribbean Currency Board and print its own currency.

Transport and communications. The island is provided with a network of good roads. The deepwater harbour at Bridgetown is the only port of call, but several international airlines and British West Indian Airways

run regular services to Seawell Airport near the south coast. The headquarters of the eastern Caribbean fishing fleet and receiving and transmitting stations of Cable and Wireless Ltd. are also located on Barbados.

Administration and social conditions. Queen Elizabeth II is the head of state. The British crown is represented by a governor general who acts on the advice of the island's Cabinet. Parliament consists of an upper and a lower house. The upper house, called the Senate, contains 21 government-appointed members; 12 are chosen by the prime minister, two by the leader of the opposition, and seven by the governor general. The House of Assembly consists of 24 members elected by universal adult suffrage. There is a prime minister's Cabinet, a High Court, and a Privy Council.

Education. The provision of state education compares favourably with that of Jamaica and Trinidad. The majority of the student population is educated at state schools, of which two grammar schools—Harrison College and Lodge School—enjoy reputations that extend throughout the eastern Caribbean. The government's expenditure on education is between a quarter and a fifth of the annual budget. According to the 1960 census, over 80 percent of the male population over 15 had completed primary school education, and 16 percent had obtained a secondary school education. There is a liberal arts college, founded in 1963 as part of the regional University of the West Indies, as well as a teacher-training college, a centre for multiracial studies financed by the University of Sussex, and an Anglican theological college.

Cultural life. The island has both a museum and a good public library. There is one daily newspaper, a serious literary magazine, and local radio and television stations. It is in sport, however, that Barbados can claim pre-eminence. Since 1950 the island has produced at least two cricketers of international reputation—Sir Frank Worrell and Gary Sobers—and a disproportionately large number of the members of the West Indian Test Match cricket teams.

Prospects. Political control is in black hands. Socially, however, Barbados is a rigidly stratified society, composed of a tiny planter and merchant class, mainly white; an equally small group of retired, rich foreigners—winter residents; and a professional group consisting of foreign doctors, engineers, hotel managers, and business people on contract. There is also a slightly larger black middle class of doctors, lawyers, teachers, and civil servants; the black working class represents 80 percent of the population. Economic power is in private—usually white and often foreign—hands. The economy is dependent upon sugar and tourism, both highly vulnerable industries. Granted reasonable good luck and a government that can preserve a reputation for honesty and prudent social spending, however, the two-party government system may survive, and the sharp distinctions of class and colour become softened with time.

BIBLIOGRAPHY. The *Barbados Census for 1960* is detailed and useful, but it is difficult to assess its accuracy. The *Report on Vital Statistics and Registrations* is concise and clear. Both the MINISTRY OF EDUCATION report, *Barbados 1967/68*, and the *Economic Survey 1969*, prepared by the Economic Planning Unit, are brief, but useful. JOHN H. PARRY and PHILIP M. SHERLOCK, *A Short History of the West Indies* (1958), is an orthodox, middle-of-the-road, general history that is mainly dependent on secondary sources; G.K. LEWIS, *Growth of the Modern West Indies* (1968), a stimulating, carefully researched, left-wing, idiosyncratic account; and JOHN MACPHERSON, *Caribbean Lands*, 2nd ed. (1967), an engagingly written, straightforward, geographical account of the area—well illustrated with photographs. F.A. HOYOS, *The Rise of West Indian Democracy* (1963), contains local history dealing with the rise of Sir Grantley Adams' party in Barbados. For ethnography, see M.G. SMITH, *The Plural Society in the British West Indies* (1965). Vegetation and animal life of Barbados are discussed in E.G.B. GOODING, A.R. LOVELESS, and G.R. PROCTOR, *Flora of Barbados* (1965); J.A. ALLAN, *The Grasses of Barbados* (1957); I. BAYLEY, "The Whistling Frogs of Barbados," *J. Barbados Mus.*, 17:161–170 (1950); and GARTH UNDERWOOD, *Reptiles of the Eastern Caribbean* (1962). For the economy, see JEANETTE BETHEL, *A National Accounts Study of the Economy of Barbados* (1960).

(C.S.J.)

Barcelona

The major Spanish port and commercial centre on the Mediterranean Sea, Barcelona, capital of the province of the same name in the region of Catalonia in northeast Spain, is a city of great individuality not only because of its physical surroundings, which, with its backdrop of mountains, are as fine as any on the Mediterranean, but also because of its Catalan cultural heritage and the proud independence of its people.

The city's geographical location, together with a mild and agreeable climate, its broad, busy avenues, and the attractions of its old buildings in the Gothic quarter—including the cathedral and the church of Santa Maria

Josip Ciganovic—FPG



Puerta de la Paz, Barcelona. The Columbus monument stands at left, and the Reales Atarazanas, a 14th-century arsenal housing the maritime museum, is at right.

del Mar—and its modern exemplars in the exuberant architecture of Antonio Gaudí, make Barcelona one of the most interesting and pleasant cities on the Mediterranean. In spite of residual pockets of poverty, the city's economic dynamism, cultural activity, and tourist attractions enhance a deeply rooted cosmopolitan character, joining old traditions with a good measure of modern progress.

History. *Foundation and medieval growth.* By tradition, Barcelona was founded by the ancient Phoenicians or Carthaginians and the name may be derived from the famous ruling Carthage family, the Barcas. During the Roman epoch it received the name Colonia Faventia Julia Augusta Pia Barcino, later called Barjalūnah by the Moors. Toward the end of the 3rd century the first city wall was built. After the domination (begun in AD 415) of Visigothic invaders, Barcelona fell, around 717, under the sway of the Islāmic world. In 801 the Carolingian emperor Louis I the Pious expelled the Arabs and added the city to his domains, ending centuries of instability. With the Muslim frontier kept at a prudent distance, the city prospered while the counts of Barcelona made themselves independent and, during the 10th and 11th centuries, consolidated their influence over the rest of Catalonia. The union, in 1137, of Catalonia and Aragon was followed by the rise to political importance of the wealthy merchant class, from which were elected the members of the municipal council, Consell de Cent (Council of 100). Barcelona now grew to be one of the foremost trading cities of the Mediterranean, the rival of Genoa and Venice, and its 13th-century maritime code, *Llibre del consolat de mar* (Consulate of the Sea), was for long widely recognized as authoritative.

Origin of name

Epidemics of plague, which decimated the population, brought a period of decline in the 14th century. Barcelona also lost ground when Naples became the political capital of the Catalan-Aragonese monarchy, in its stead, in 1442. Abroad, in the 16th century, the growing power of the Turks and the colonization of the New World caused Barcelona's former commercial influence in the Mediterranean to diminish. At home, from 1615, Catalonia's relations with the court in Madrid suffered progressive deterioration: Catalonia was excluded from commerce with America, and, until 1778, monopolistic control was granted to Andalusia.

The modern era. In 1705, during the War of the Spanish Succession, the archduke Charles III of Austria, contending for the Spanish throne, established his court in Barcelona. When, however, Philip V won the throne and his forces captured the city in 1714, he swiftly brought its self-government to an end.

By the time the city had recovered from its subjection under Napoleon (1808–13), industrialization in Spain was getting under way and the city's first steam-powered textile factory was put into operation in 1832. Because of its cotton industry, Barcelona became in this period the kingdom's most important city, attracting large numbers of the working class. Not surprisingly, its subsequent history was marked by episodes indicative of social, industrial, and political unrest. Especially serious were the uprisings of 1835, when 11 convents were destroyed, and of 1909, when more than 60 churches and religious buildings disappeared from the city's architectural inheritance.

19th-century industrialization

Barcelona was the stronghold of the Catalonian separatist movement and, following the outbreak of the Spanish Civil War in 1936, it became the seat of an autonomous Catalonian government and the main centre of Republican strength and suffered some damage from bombing. Its fall, in January 1939, led to the Republican government's final surrender. By the 1950s, however, Barcelona prospered again.

The contemporary city. *Layout and boundaries.* Barcelona is built on a gentle slope facing southeast to the Mediterranean. It lies in a fertile plain between the Rivers Besós, to the northeast, and Llobregat, to the southwest, and is backed by an amphitheatre of mountains that culminates in the Tibidabo (1,745 feet; 532 metres). The fortified hill of Montjuich (630 feet) rises from the sea and separates the city from the mouth of the Llobregat. The municipal area is 35 square miles (91 square kilometres) and the perimeter 27 miles. The city is a remarkably homogeneous urban unit, residential and commercial, centred on the core of the old city with its surrounding municipalities and its industries extended at the perimeter.

The main axis of the old town is formed by the Ramblas, a series of successive wide, tree-lined avenues, leading north to the Plaza de Cataluña, the largest square and commercial centre, and south to the Paseo Marítimo and seafront. The Ramblas, a favourite promenade renowned for its charm, has seats beneath the trees down both sides of the wide pavement and many stalls where birds and flowers are sold. To the north is the new town, the Ensanche (Extension), laid out in squares and crossed by wide avenues. The old town and the more modern part west of the Ramblas are enclosed in a semicircle by the Rondas, peripheral avenues occupying the ground where the fortifications once stood. At the northernmost part of the Rondas in the Plaza de la Universidad, through which runs the broad Avenida de José Antonio, with a bullring at either end. The other great thoroughfare, the Avenida del Generalísimo Franco is a residential street running southwest–northeast across the upper part of the town to the slopes of Pedralbes, where a 14th-century monastery stands on the hillside.

Architectural features. The oldest part of the city is built on a small hill, Monte Taber, the Roman walls being still visible in some streets. In the centre of the old city stands the cathedral. The present structure was built between 1289 and the late 15th century, and the west front was added in the 19th century. Excavations, however, brought to light the remains of a 6th-century basilica. Close by, in the Plaza del Rey, and the chapel and great hall of the Palacio Real Mayor, the royal palace of the counts of Barcelona (mainly 14th-century but with earlier parts) and the 16th-century palace housing the archives of the crown of Aragon. South of the cathedral are the archdeacon's house (16th century), containing the city's archives, and the episcopal palace. In the Plaza de San Jaime are the Provincial Diputación (congress hall), which was built in the 15th and 16th centuries and contains the Capilla de San Jorge and the Orange Tree Court. The Gothic town hall is a 14th- and 15th-century building with a modern facade.

The core of the city

In the Puerta de la Paz, the Columbus monument towers 197 feet high and overlooks the port; there also the Reales Atarazanas, a 14th-century arsenal, houses the maritime museum.

Perhaps the most remarkable of all Barcelona's monuments is the huge and elaborate unfinished church, Templo Expiatorio de la Sagrada Familia, begun in the 1880s, whose fantastic openwork spires dominate the city. It is the best known work of the Catalan architect and sculptor Antonio Gaudí, who designed many other notable structures in Barcelona, among them the Park Güell and the Casa Batlló.

The people. With more than 1,750,000 inhabitants in 1971, Barcelona proper has one of the highest population densities in the world. In relation to its suburbs, the city constitutes the main part of a metropolitan area (District of Barcelona) that comprises 27 municipalities with a total area of 188 square miles and a total population of over 2,650,000. It has been estimated that because of immigration from poorer regions of Spain, the population of the metropolitan area would continue to swell in the 1970s and beyond. The municipal birth rate and death rate were 17.07 and 8.76 per 1,000, respectively at the end of the 1960s.

Economic life. Three-quarters of the industrial establishments in Catalonia are concentrated in the Barcelona area, which contributes 20 percent of Spain's industrial output. The metal and chemical industries were, in the 1970s, gaining in importance over the once-dominant textile industry. Chief products are automobiles (including industrial vehicles), heavy machinery, office machines, and chemicals.

Decline of the textile industry

Barcelona's stock exchange is active, and the city is a

centre for Spanish and foreign banks, savings banks, and other financial institutions.

Government and services. The head of the municipal administration is a mayor appointed by the head of state for a six-year period, which is renewable. The mayor appoints six delegates who, with him and with six of the 36 councilmen and three deputy mayors, form the Executive Municipal Commission. The municipality is divided into 12 districts, with a District Board that serves as a coordinating link.

City transportation is provided by buses, subways, suburban railroads, and cable cars. City streetcars disappeared in 1971. The total number of passengers using collective transportation is about 600,000,000 annually. In 1970 some 350,000 private automobiles were in use in Barcelona. In 1969 the first toll freeways were opened, linking Barcelona with neighbouring towns. Later, an urban belt of rapid transit and three tunnels under the Tibidabo were under construction: these would permit the expansion of Barcelona beyond its surrounding mountain chain. The metropolitan subway (opened in 1924) had 19 route miles in 1971, with additional construction underway in the early 1970s to provide an eventual 49 mile, 108 station system by 1978. The Prat Airport, seven miles west of the city in the plain of Llobregat, serves international airlines. Railways connect Barcelona with the rest of Spain and with France.

The port of Barcelona, at first little more than an open roadstead, was improved in 1474 by the construction of a mole, the Moll de Santa Creu, but the harbour proper dates from the 17th century. It was subsequently greatly enlarged. The present port has 7.4 miles of docks and three special quays for gasoline and oil tankers; it also has silos, warehouses, and terminals for containers. In 1970 the traffic amounted to 9,100,000 tons, of which 2,500,000 tons were carried by oil tankers. There are over 130 regular shipping lines that link the port with other world ports.

The city receives electric energy from the zone of the Pyrenees, from other Spanish zones, and from the south of France. In 1970, Catalan Gas and Electricity set in operation a regional network of distribution for natural gas by installing a receiving terminal in the port of Barcelona. Nuclear energy was to come in the 1970s from the Vandellós Central Plant in the province of Tarragona.

Increasing consumption of water causes supply problems, since some sources and the Llobregat and Besós rivers do not have sufficient volume to cover the needs of the industrial and urban agglomeration. Water from the Ter River in the province of Gerona has been channelled to Barcelona.

The Catalan language

Cultural life. Barcelona is more than a cultural centre for Spain as a whole. It nurtures as well a long historic tradition that is given cohesion and a means of expression through the language of Catalonia: Catalan. This linguistic peculiarity, which has deeply marked the political, social, and cultural history of Catalonia, did not produce a closed provincial culture but one open to international currents, especially to those of Europe.

The outstanding museums are: the Fine Arts Museum of Catalonia (Romanesque and Gothic painting); the Federico Marés Museum (12th–18th-century sculpture); the Picasso Museum (enriched with important donations from the painter, who had spent nine years of his youth in Barcelona); and the Maritime Museum.

The University of Barcelona (founded in 1450 by Alfonso V the Magnanimous) and the technical colleges that occupy a large group of buildings on the edge of the city near Sarria had about 16,000 students by the early 1970s. In 1968 the Autonomous University of Barcelona was created, with more academic and administrative independence, to meet the growing demand for advanced education and research.

The general and specialized libraries (Central Library of Catalonia, the University Library, and the Municipal Periodical Library) and the scientific and artistic institutions are reinforced by many civil associations dedicated to varied cultural, recreational, or civic purposes.

Barcelona was one of the first Spanish cities to have a printing press and one of the first in Europe to publish a newspaper (*Diario de Barcelona*, 1792), which still appears. The city has ten daily newspapers and many periodicals. There are many radio stations, a television station connected to the state network, and three movie studios.

Opera and ballet have their headquarters in the Gran Teatro del Liceo (founded in 1847). The Barcelona International Festival of Music, celebrated in October, is the chief musical event of the year.

Recreation. There are many sports grounds, and among the events that are particularly popular are international automobile races, swimming contests, tennis matches, and association-football matches. The Barcelona Football (soccer) Club, founded in 1899, has more than 50,000 members and a stadium with a capacity for 100,000 spectators.

As a cultural centre Barcelona has nurtured the traditional Catalan dances, such as the well-known *sardana* round dance, the region's national dance. It also offers the attractions characteristic of a large city, such as amusement parks (Montjuich and Tibidabo and the Ciudadella Park, with its zoo), and all kinds of other entertainments (there are more than 130 movie theatres, 14 playhouses, two bullrings, and more than 100 dance halls).

BIBLIOGRAPHY. Three classic works deal with the history of Barcelona: ANDRES AVELINO PI Y ARIMON, *Barcelona antigua y moderna*, 2 vol. (1850–54), which describes Barcelona from its foundation to modern times; F. CARRERAS CANDI, *Geografía general de Catalunya*, vol. 6, *La ciutat de Barcelona* (n.d.); and ANTONIO DE CAPMANY, *Memorias historicas sobre la marina, comercio y artes de la antigua ciudad de Barcelona* (1779). A three-volume, re-edited edition, with commentary, was published in 1961. This latter book deals more with economic aspects of the city's development. ARMANDO SAEZ BUESA, *La Población de Barcelona en 1863 y 1960* (1968), is a comparative analysis of the urban demography of Barcelona, using as a base a series of unedited publications from 1863. *Ayuntamiento de Barcelona*, published by the Estadistica municipal, is an annual résumé (begun in 1936) and a primary statistical source. A more general view is found in ROLAND COURTOD and ROBERT FERRAS, *Les Grandes villes du monde Barcelona* (1969). The best guides to outstanding landmarks in the city include: CARLOS SOLDEVILA ZUBIBURU, *Guía de Barcelona* (1951); *Barcelona, la ciudad, los museos, la vida* (1962); and ALEXANDRE CIRICI, *Barcelona pam a pam* (1971). English-language works are: JAIME MIRAVALL, *Barcelona* (1951; Eng. trans. 1964); PASCUAL MAISTERRA, *Barcelona* (1967); CLIFFORD KING, *Barcelona* (1968); JAMES MORRIS, *Barcelona* (1967); and ROBERT GOLDSTON, *Barcelona, the Civic State* (1969).

(J.M.C.I.)

Barents Sea

A major outlying sea of the European sector of the Arctic Ocean, the Barents Sea lies to the northeast of the Scandinavian peninsula, separating the northern European Soviet Union from the great archipelagoes of Spitsbergen (and hence the Arctic Basin proper) and Franz Josef Land to the north, and the long, narrow, island of Novaya Zemlya (and hence the Kara Sea) on the east. The strait between Svalbard and Bjørnøya (Bear Island), halfway between Norway and Spitsbergen, leads to the neighbouring Greenland Sea, and the Norwegian Sea lies to the southwest. An irregular southern inlet forms the White Sea. The Barents Sea's area is 542,000 square miles (1,405,000 square kilometres)—larger than South Africa—and its average depth is 750 feet (229 metres), plunging to a maximum of 2,000 feet. The sea was known to Vikings and medieval Russians, its early frequenters, as the Murmean Sea and first appeared under its modern name in a chart published in 1853, thus honouring a 16th-century Dutch seeker of a northeast passage to India, Willem Barents. A German meteorologist, after an 1848 expedition, suggested that Atlantic waters penetrated the sea. The subsequent corroboration of this hypothesis was followed by the Russian academician A.F. Middendorf's important discovery of the warm North Cape Current, streaming in around the northern tip of Scandinavia. Further valuable research stemmed

History of exploration

from an 1898–1905 Russian expedition, charting warm and cold currents, the studies of the explorer F. Nansen (reported in *Northern Waters*, 1906), and the detailed reports of a number of Soviet research institutes. For related information see ARCTIC OCEAN; SOVIET UNION; and RUSSIAN SOVIET FEDERATED SOCIALIST REPUBLIC.

Physical characteristics. The Barents Sea covers a relatively shallow continental shelf fringing the Eurasian landmass. In Tertiary times (65,000,000 to 2,500,000 years ago) this region was land, with powerful rivers coursing over its surface. During the Pleistocene Ice Age (ending only 10,000 years ago) glaciers covered the surface, which later sank, though retaining older relief patterns. The floor—which is covered by sands, silts, and a sandy-silt mixture—is traversed from east to west by the major Bear Island Trench, 1,600 to 2,000 feet deep, and the smaller South Cape, Northern, and Northeastern trenches. The Central and Perseus elevations make for shallower relief in the north, and there are fishing banks and shallows to the southeast. The only island (apart from those on the periphery) is the oval Kolguyev Island, which lies in the southeast. The western portion of the mainland coast, as far as Cape Svyatoy Nos, is elevated, with abrupt shores pierced near the Soviet–Norwegian border by the deep, glacier-widened drowned valley inlets known as fjords. East of the Kanin Peninsula, which lies at the entrance to the White Sea, the coast is low-lying, with a number of shallow bays and inlets. Away from the mainland, the bleak coasts of the various archipelagoes are steep and high, with glaciers frequently plunging down to the sea and accumulations of moraines (glacier-carried debris) in the hollows.

The climate is sub-Arctic, with winter air temperatures averaging −13° F (−25° C) in the north and 23° F (−5° C) in the southwest; summer temperatures in the same regions are, respectively, 32° F (0° C) and 50° F (10° C). Absolute maxima and minima nevertheless range from −40° F (−40° C) off Novaya Zemlya to 87° F (31° C) off Murmansk. Annual precipitation is 20 inches (500 millimetres) in the south but only half this amount in the north. In years when Atlantic cyclones (low pressure areas) bring in cloudy weather, sunshine decreases and air and water temperatures exceed the norm; the converse applies when the clear, cold weather brought by Arctic anticyclones (high pressure areas) predominates.

Warm currents in the sea derive from the North Cape and Spitsbergen branches of the Norway Current, but heat is lost in mixing with colder waters. Salinity is high, at 34 parts per thousand; but ice forms in winter, although fields are thin and icebergs do not linger long. In summer, the edge of the ice is far to the north, sometimes beyond the sea, but some icebergs break from the great archipelagoes and drift as far south as Murmansk. The tidal amplitude varies greatly, especially in narrow coastal zones. Tidal current direction also varies; an important mixing of water layers thus occurs. Ice-free ports are the Soviet ports of Murmansk and Teribyorka, and the Norwegian port of Vardø.

Biological characteristics. Nutrients in the sea are facilitated by the influx of the North Cape and South Spitsbergen currents. The microscopic forms of phytoplankton produced feed, among others, a number of deep-sea invertebrates, a wide range of small, shrimp-like crustaceans, bivalves, and sponges. These in turn support fish (including cod, herring, salmon, plaice, and catfish), sea mammals (seals and whales), and land mammals (polar bears and Arctic foxes). There are also different types of sea gulls and, in warm weather, ducks and geese. Underwater flora is very rich in the shallow southern regions; and brown, red, and green algae are widespread. About 20 to 40 percent of the coastline contains shrubs, mosses, and lichens, whereas the rest is rock and stones. Grasses are rare, although vegetation increases on the continental mainland.

Sea nutrients

Fishing flourishes, with yield in the southern areas reaching a peak of 1,104,500 metric tons in 1967, followed by a sharp decline because of over-hunting. The hunting of harp seals in the southeast is now strongly regulated; fishermen from Norway and the Soviet Union take only about 30,000 specimens a year.

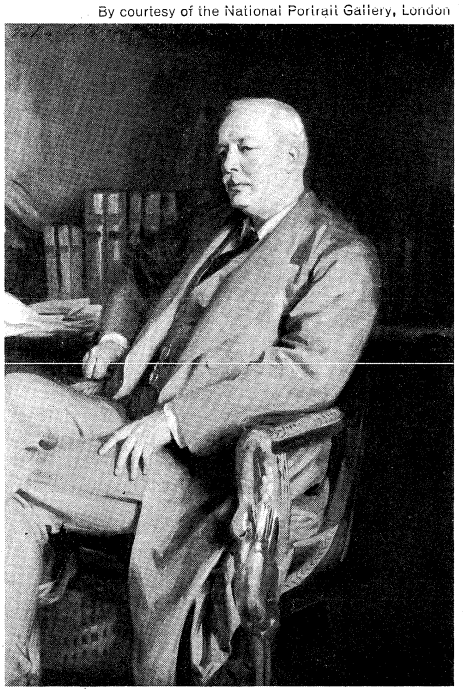
BIBLIOGRAPHY. The METEOROLOGICAL OFFICE, *Monthly Meteorological Charts and Sea Surface Current Chart of the Greenland and Barents Seas*, 2nd ed. (HMSO, 1959), is recommended as a textbook for studying the meteorological and hydrological regime of the Barents Sea. See also M.V. KLENOVA, *Geologiya Barentsova morya* (1960), a general study of the oceanography of the Barents Sea; KAZIMIERZ DEMEL and STANISLAW RUTKOWICZ, *The Barents Sea* (Eng. trans. 1966); and ANDREW W. GARCIA, *Oceanographic Observations in the Kara and Eastern Barents Sea* (1969).

(M.M.A.)

Baring, Evelyn, 1st Earl Cromer

An army officer, administrator, and diplomat, Evelyn Baring, 1st Earl Cromer, is best known as one of the last great proconsuls who served imperial Britain. For 24 years he represented his country as British agent and consul general in Egypt. He exerted a decisive influence on that country's development as a modern state, and his impressions of and attitudes toward Egypt were mainly responsible for Britain's imperial policy of a long-term commitment there.

By courtesy of the National Portrait Gallery, London



Baring, oil painting by John Singer Sargent, 1902. In the National Portrait Gallery, London.

Born on February 26, 1841, of a family distinguished in politics and banking, Evelyn Baring received his training at the Royal Military Academy, Woolwich, from which he graduated at the age of 17. He received a commission in the Royal Artillery and served in Corfu (where he met his first wife Ethel Errington Stanley), Malta, and Jamaica. He then entered the Staff College and a year later in 1869 he graduated first in his class. For a while he served in the War Office, but military life was not to his taste, and when in 1872 his cousin, Lord Northbrook, just named viceroy to India, offered to take him along as his private secretary, Baring accepted.

Career in India

In India Baring rapidly made his mark. His administrative qualities were obvious and highly appreciated by his superiors. His colleagues, however, dubbed him "Vice-Viceroy" and "Over-Baring," nicknames which clearly bespoke his self-assured efficiency and ability to command—traits invaluable in a leader of men, though not necessarily conducive to popularity among his equals. His manner was gruff to his equals, condescending and patronizing to his subordinates and to the people he chose to describe as the "subject races." Imbued with immense common sense and a profound belief in himself

and his country, he could not abide cant or hypocrisy. He was the typical Victorian colonial administrator, eminently fair and just but with little to endear him save an occasional flash of humour.

It was in India that Baring buried his early notions of self-determination for colonial peoples and decided that strong rule accompanied by reform programs was the only way to help the downtrodden peasant. His later experiences in Egypt strengthened his views on the tyranny of native rulers and the need for reform by the British. Reform became translated into one enduring principle which governed all his administrative actions—the need for a sound financial system.

Baring first went to Egypt in 1877 when he served as representative of the British holders of Egyptian bonds on the recently created Egyptian Public Debt Commission. The Commission was designed to help the Egyptian viceroy, the khedive Ismāʿīl Pasha, out of his financial difficulties, and also to safeguard the interest of the bondholders. Egyptian finances, however, were in a worse state than Baring had imagined, and he was the prime mover behind the creation of a Commission of Inquiry into Egyptian finances. When his advice was turned down by the khedive Ismāʿīl Pasha he resigned and returned to England, but when Ismāʿīl Pasha was deposed in 1879 he was invited to return to Egypt as British controller of the debt. In 1880 he became the financial member of the Viceroy's council in India where he remained for three years. After the British occupation of Egypt in 1882, he returned to Egypt once more in 1883 as British Agent and consul general with plenipotentiary powers, having in the interim been knighted.

Baring's mandate in Egypt

Baring's mandate in Egypt was to carry out wide-scale administrative reforms in a country that was bankrupt and had just gone through the upheavals of a popular revolution and a foreign occupation, and eventually to effect the evacuation of the British forces stationed there. Quite quickly he came to the conclusion that reforms and evacuation were incompatible, that reforms were of more lasting value to the mass of the Egyptians, and that evacuation should come only in the distant future when the Egyptians had been taught self-rule. He therefore instituted a form of government that came to be called the Veiled Protectorate whereby he ruled the rulers of Egypt, with the assistance of a group of English administrators trained in India, who were placed in key positions as advisers to the Egyptian government. Until his resignation in 1907 he remained the real ruler of Egypt. The system worked well during the first ten years, for the khedive Tawfīq Pasha was a weakling who abdicated all responsibility to the English. Egypt was made financially solvent by 1887, and after the British forced the Egyptian government to give up its attempt to reconquer the Sudan —wrested from its control by the religious rebellion of the *Mahdī*—there followed a period of peace and stability that allowed the country to recover from the chaos of the previous decade. Baring's parsimony in public spending and his encouragement of public irrigation works and other agricultural projects soon resulted in increased prosperity.

In 1892, a young new ruler, ʿAbbās Ḥilmī II, struggling to divest himself of the onus of the Veiled Protectorate, gave encouragement to a budding nationalist movement. Baring, who had been raised to the peerage as Lord Cromer, was as inflexible in his dealings with the young Khedive as he had been with his predecessor and succeeded in intimidating him quite thoroughly.

Throughout his years in Egypt, Cromer won the respect and admiration of the many men who occupied the foreign office and who usually deferred to his judgment in matters concerning Egypt. An exceedingly hard worker, his day began at sunrise and continued well after sunset with a two-hour break in the afternoon for physical exercise, which he pursued with as much deliberation as the rest of his duties. During his periods of relaxation he steeped himself in the classics he so admired. As a young officer he had learned Greek and Latin as well as French and Italian: he was later to learn Turkish, the language of the Turco-Circassian elite in Egypt. Yet in spite of his long stay in Egypt, he never attempted to learn Arabic and was never able to communicate either with the peasant whom he claimed to know so well, nor with the middle class that was to produce a new breed of nationalists. He had little liking for the Oriental mind, which he termed "slipshod," and even less understanding of it, despite his claims to the contrary. With age his aloofness increased and he dismissed the young nationalist movement as unimportant. Instead he worked hard at effecting the Entente Cordiale of 1904 with France, which set the seal on the permanent occupation of Egypt. His first wife, Ethel Errington, died in 1898. He married a second time in 1901, to Lady Katherine Thynne, the daughter of the 4th Marquess of Bath.

Retirement

In 1907 an incident in an Egyptian village, Dinshwai, in which a British officer was killed, resulted in brutal sentences being passed on the Egyptian peasants involved. Public outrage created a storm both in Egypt and in the British House of Commons and led the new Liberal cabinet under Prime Minister Sir Henry Campbell-Bannerman to adopt a more accommodating attitude toward Egypt. Cromer, who had little to do with the sentences since he was on home leave at the time, realized that a change was impending and, as his health had deteriorated, resigned office in 1907.

On his return to England he spent his time writing and in the House of Lords, where he was the foremost exponent of free trade. In 1916 he presided over the Dardanelles Commission, but the strain proved too taxing and he died on January 29, 1917.

BIBLIOGRAPHY. Among Baring's published works are: *Modern Egypt*, 2 vol. (1909), an account of his work and years in Egypt; *Abbas II* (1915), a somewhat snide and inaccurate telling of his relationship with the Khedive; and his annual reports on Egypt (1884–1906). LORD ZETLAND, *Lord Cromer* (1932), is his official biography. He is also mentioned in ALFRED M. MILNER, *England in Egypt* (1892), and showered with praise for his work in Egypt; also in RENNELL RODD, *Social and Diplomatic Memoirs, 1884–1893* (1922). AFAF LUTFI AL-SAYYID, *Egypt and Cromer* (1968), presents a modern Egyptian view of Cromer's work in Egypt; JOHN MARLOWE, *Cromer in Egypt* (1970), a modern British view of Cromer's life.

(A.L.al-S.M.)

Barmakids

The Barmakids (popularly known as the Barmecides) were a priestly family of Iranian origin, from the city of Balkh in Khorāsān, who achieved prominence in the 8th century as scribes and viziers to the early ʿAbbāsid caliphs. Their ancestor was a *barmak*, a title borne by the high priest in the Buddhist temple of Nawbahār. The Barmakids were also known for their patronage of literature, philosophy, and science and for their tolerant attitude toward various religious and philosophical issues. They promoted public works—such as canals, mosques, and postal services—but also squandered money on building magnificent palaces by the Tigris.

Origin of the family

When Balkh, the native town of the Barmakids, fell to the Arabs *c.* 663, Khālid ibn Barmak and his brothers moved to the garrison city of Basra in Iraq, where they converted to Islām.

Khālid ibn Barmak. Khālid ibn Barmak is the first Barmakid about whom much is known. He first appears in the mid-8th century as a supporter of the revolutionary movement that established the ʿAbbāsid caliphate. In 747 Khālid was put in charge of the distribution of spoils when the ʿAbbāsid army moved toward Iraq. Afterward, he was sent to Dayr Qunnā to administer the district. Under the ʿAbbāsid caliph Abū al-ʿAbbās as Saffāḥ, Khālid shared ministerial authority with Abū al-Jahm and was entrusted with the army and the collecting of the land tax.

Khālid's intimacy with the caliph reached the extent that the latter entrusted him with the upbringing of his daughter. During the reign of al-Manṣūr, Khālid was appointed governor of Fars, and in 765 he was among the delegates to obtain Prince ʿIsā's renunciation of succession to the caliphate. Khālid then was nominated governor of Ṭabaristān, where coins were struck in his name

between 767–771. There, he distinguished himself by capturing Ustūnā Wand and building a town called Manṣūrah. Because of political intrigues and rivalry, al-Manṣūr dismissed Khālid in 775 and imposed a heavy fine upon him. Al-Khayzurān, prince al-Mahdī's wife, helped him to raise the money. Subsequently Khālid was sent to Mosul to suppress Kurdish disturbances while his son Yaḥyā was put in charge of Azerbaijan. The Barmakids were endowed with more privileges during al-Mahdī's reign, when Khālid, helped by his son Yaḥyā, was appointed governor of Fars.

Yaḥyā. Khālid died in 781/782. Yaḥyā, well trained by his father and already undertaking various administrative jobs, was nominated in 778 as secretary-tutor to the caliph's son Hārūn. As secretary, he played a decisive role in ensuring the succession of his ward to the caliphate. In 779/780 the caliph appointed Hārūn, accompanied by Yaḥyā, to lead the expedition against the Byzantines. On his return Hārūn was put in charge of the western provinces, with Yaḥyā as his adviser. In 781 Hārūn was proclaimed second in succession after his brother Mūsā, but a little later—and due to al-Khayzurān's and Yaḥyā's influence—the caliph intended to deprive Mūsā of his rights as an heir apparent but died before accomplishing his scheme. Hārūn decided not to put up any opposition to the new caliph Mūsā al-Hādī. This wise decision, inspired by Yaḥyā, perhaps saved the empire from civil war.

Yaḥyā's influence on Hārūn ar-Rashīd

Al-Hādī, in turn, confirmed Yaḥyā's position with Hārūn. This was, no doubt, a tactical error by al-Hādī, for when he decided to nominate his own son to the caliphate, Hārūn would have given in had Yaḥyā not objected. Yaḥyā tried in vain to convince the caliph that the violation of an oath after so short a time would have disastrous consequences. Hārūn and Yaḥyā were jailed. At this point, however, al-Hādī suddenly died in obscure circumstances.

Thus Hārūn ar-Rashīd (786–809) was raised to power, not by his own efforts but by the machinations of the queen mother al-Khayzurān and Yaḥyā the Barmakid. It was, therefore, no surprise that he put the whole administration in the hands of Yaḥyā and his sons. Yaḥyā received the title of *wazīr*, and his sons al-Faḍl and Jaʿfar were placed in charge of the caliph's personal seal.

Al-Faḍl and Jaʿfar. Al-Faḍl and Jaʿfar also bore the title *wazīr*. Jaʿfar, the younger brother and ar-Rashīd's favourite, was known for his eloquence and for his love of pleasure and parties. He rarely left the court, but when, in 796, the caliph sent him to control a disturbance in Syria, Jaʿfar succeeded in quieting the situation. On his return, he was appointed director of the bureaus (*dīwāns*) of the post, textiles, and mint. In the latter office Jaʿfar minted coins in his name in various provinces. Al-Faḍl, unlike his brother, distinguished himself by his competence and seriousness. When in 792 the ʿAlid Yaḥyā ibn ʿAbd Allāh rebelled in Daylam, al-Faḍl, through diplomacy and promises, persuaded him to give in. In 793 al-Faḍl was appointed governor of Khorāsān; he was able to put an end to the disturbances in Kābul. In 797 al-Faḍl took over the central government from his father, who resided at Mecca. Al-Faḍl, besides, was a tutor to ar-Rashīd's elder son and heir apparent al-Amīn.

The fall of the Barmakids. The Barmakids' influence lasted 17 years, but they were extirpated at the peak of their power and fortune. Jaʿfar, only 36 years old, was executed in 803 and parts of his body displayed on the bridges of Baghdad. Other Barmakids, with the exception of Muḥammad ibn Khālid, were imprisoned and their property confiscated. Yaḥyā and al-Faḍl died in prison in 805 and 808 respectively. A number of their partisans were accused of heresy and executed.

Assessment

The Barmakids' fall was sudden and brutal. Many accusations were made against them at the time, but the Barmakids' disgrace is to be attributed, first, to their overmighty influence in the court, administration, and society. Second, they seized every opportunity to enrich themselves (which accounts for their ostentatious generosity). Thirdly, they showed a certain degree of liberalism toward various religious and political sects, which the caliph considered as a danger to his authority. The Barmakids' role ended but their fame survived. They became the subject of controversies among historians. Contradictory traditions, marred by the obvious flattery or prejudice by which they are inspired, represent an attempt by narrators to exalt or discredit the Barmakids' character, thus obscuring their true historical role. Late Muslim literature, especially Persian literature, is inclined to visualize the Barmakid period as an ideal period in the history of the caliphate. These traditions even consider the Barmakids' Zoroastrian by faith and trace their descent to the Sāsānid period. Be that as it may, their downfall was to be considered the end of the theory that ministers were initiators of policy and not merely heads of administration; it also marked the caliph's reaction against the liberal tendency current at the time.

The expression "Barmecide feast," for an imaginary banquet, comes from "The Barber's Tale of His Sixth Brother" (*The Arabian Nights' Entertainment*), where a Barmakid has a series of empty dishes served to a hungry man to test his sense of humour.

BIBLIOGRAPHY. The chief sources are the classical Arabic and Persian works such as AT-TABARI, *Tārīkh* (1903); YAʿQUBI, *Tārīkh*, 2 vol. (1883); AL-MASʿUDI, *Murūj* (French trans., *Les Prairies d'or*), 9 vol. (1861–77); IBN KHALLIKAN, *Wafayat al-aʿyān* (Eng. trans. by M. DE SLANE, 1961); and AL-JAHSHIYARI, *Kitab al-wuzarā* (1938; German trans., 1958). See also CHARLES HENRI SCHEFER, *Chrestomathie persane à l'usage des élèves de l'École Spéciales des Langues Orientales Vivantes*, 2 vol. (1833–85).

Modern works: Apart from general works on Islāmic history, see LUCIEN BOUVAT, *Les Barmécides, d'après les historiens arabes et persans* (1912); W. BARTHOLD, "Barmakids," in the *Encyclopaedia of Islam*, vol. 1, pp. 663–666 (1913); D. SOURDEL, "al-Barāmika," *ibid.*, new ed., vol. 1, pt. 2, pp. 1033–36 (1960); F. OMAR, "Hārūn al-Rashīd," *ibid.*, new ed., vol. 3, pp. 232–234 (1971); SYED NADVI, "The Origin of the Barmakids," *Islamic Culture*, 6:19–28 (1932); and HARRY PHILLBY, *Hārūn al-Rashīd* (1933).

(F.Om.)

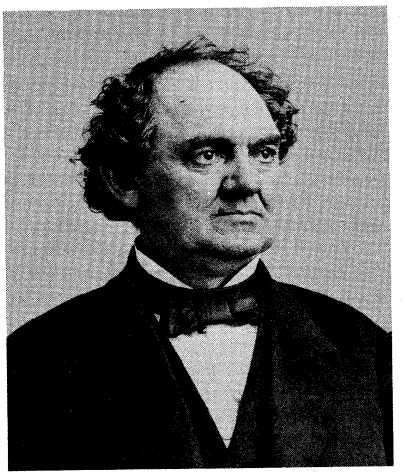
Barnum, P.T.

In a flamboyant career that spanned half of a century, Phineas Taylor Barnum established himself as the most innovative and celebrated showman ever to flourish in the United States. It was Barnum who popularized for mass audiences such amusements as the public museum, the musical concert, and the three-ring circus. In doing so, he created techniques of presentation and publicity that were to become widely imitated in vaudeville, motion pictures, and television variety shows of the 20th century. When he died, *The Times* of London echoed the press of the world in its final tribute. ". . . He created the *métier* of showman on a grandiose scale. . . . He early realized that essential feature of a modern democracy, its readiness to be led to what will amuse and instruct it. . . . His name is a proverb already, and a proverb it will continue."

Barnum was born in Bethel, Connecticut, on July 5, 1810. Although driven at the outset of his career by a desire for wealth and fame, Barnum may have been basically motivated by less selfish reasons. "This is a trading world," he wrote, "and men, women and children, who cannot live on gravity alone, need something to satisfy their gayer, lighter moods and hours, and he who ministers to this want is in a business established by the Author of our nature."

Early notoriety. Barnum was 15 years old when his father died, and the support of his mother and his five sisters and brothers fell largely upon his shoulders. After holding a variety of jobs, he became publisher of a Danbury, Connecticut, weekly newspaper, *Herald of Freedom*. Arrested three times for libel, he enjoyed his first taste of notoriety.

In 1829, at the age of 19, Barnum married a 21-year-old Bethel girl, Charity Hallett, who was to bear him four daughters. In 1834 he moved to New York City, where he found his vocation as a showman one year later when he successfully presented Joice Heth, a wizened black



Barnum.
By courtesy of the Library of Congress, Washington, D.C.

lady whom he advertised as the 161-year-old nurse to Gen. George Washington. This human relic, on her death, was exposed as a hoax.

Casting about for a legitimate undertaking, Barnum outmanoeuvred wealthier bidders to acquire John Scudder's American Museum, in New York City, a five-story marble structure filled with stuffed animals, waxwork figures, and similar conventional exhibits. The new owner rapidly transformed the museum into a carnival of live freaks, dramatic theatricals, beauty contests, and other sensational attractions. Playing upon the public's interest in the unusual and bizarre, Barnum scoured the world for curiosities, living or dead, genuine or fake. By means of outrageous stunts, repetitive advertising, and exaggerated publicity, Barnum excited international attention and made his showcase of wonders a landmark.

Barnum's museum

Between 1842, when he took over the American Museum, and 1868, when he gave it up after fires twice had all but destroyed it, Barnum's gaudy showmanship enticed 82,000,000 visitors into his halls, among them Henry and William James, Charles Dickens, and Edward VII, then prince of Wales.

Barnum's first successful exhibit in the museum was the Feejee Mermaid, which had a seemingly human head topping the finned body of a fish and was, of course, found later to be a fake. Among the genuine curiosities were Chang and Eng, Siamese twins who were connected by a ligament below their breastbones. It was, however, Charles S. Stratton, a midget only 25 inches tall who was discovered by Barnum, that proved to be his most profitable exhibit. Ballyhooing his midget as General Tom Thumb, Barnum sold 20,000,000 tickets to the museum. After being received by President Abraham Lincoln, Barnum and Tom Thumb enjoyed a triumphal tour abroad, during which the midget gave a command performance before Queen Victoria.

From promoter to impresario. Eager to change his image from promoter of freaks to impresario of artistic attractions, Barnum risked his entire fortune by importing Jenny Lind, a Swedish soprano whom he had never seen or heard and who was almost unknown in the United States. Dubbing Miss Lind "The Swedish Nightingale," Barnum mounted the most massive publicity campaign he had ever attempted. Jenny Lind's opening night in New York, before a capacity audience of 5,000, and her nine months of concerts across the United States earned immense sums.

At the peak of his career, Barnum's own appearance was nearly as familiar to the public as the exhibits he promoted. An impressive figure six feet two inches tall, semibald, with blue eyes, a bulbous nose, and potbelly, he called himself the "Prince of Humbugs." He dwelt in a three-story Oriental mansion, named Iranistan, on a 17-acre estate in Bridgeport, Connecticut, where he played host to such notables as Mark Twain, Horace Greeley, and Matthew Arnold. Close friends regarded him as good-natured, thoughtful, and kind, as well as parsimonious and egotistical.

His avocations were politics and writing. After serving two terms in the Connecticut state legislature, he was elected mayor of Bridgeport, in which post he fought prostitution and union discrimination against Negroes. In 1855 he published his autobiography, *The Life of P.T. Barnum, Written by Himself;* and because he frankly revealed some of the deceits he had employed, he was harshly taken to task by the majority of critics. Stung, Barnum continually modified the book in many revised versions, which, he claimed, sold a total of 1,000,000 copies. By 1884, more anxious for publicity than for profit, Barnum placed his autobiography in the public domain, allowing anyone to print and sell it without copyright infringement. Though the cynicism "There's a sucker born every minute" has long been attributed to Barnum, there is no proof that he ever wrote or spoke these words.

Barnum's family life was not entirely happy. One daughter had died in childhood; another was dropped from his will for committing adultery. Disappointed because he had no male heir, Barnum left a sizable bequest to a grandson on the condition that he agree to use the name of Barnum as part of his name. After 44 years of marriage, Charity Barnum died in 1873. The following year, Barnum, who was then 64, took the 24-year-old Nancy Fish, the daughter of a British admirer, for his second wife.

"The greatest show on earth"

Although his name has been popularly linked with the circus, Barnum did not, in fact, become a circus showman until he was past the age of 60. Barnum did not invent the modern circus, but, in partnership with the retiring, efficient James A. Bailey, he did give the American spectacle its gigantic size, its most memorable attractions, its widest popularity, attempting to make it what he called "the greatest show on earth." Barnum capped his circus career by purchasing a six-and-a-half ton elephant named Jumbo, who quickly earned back his purchase price during his first season under the big top.

In his 81st year, Barnum fell gravely ill. At his request, a New York newspaper published his obituary in advance so that he might enjoy it. Two weeks later, on the morning of April 7, 1891, after inquiring about the box office receipts of the circus, Barnum died in his Connecticut mansion.

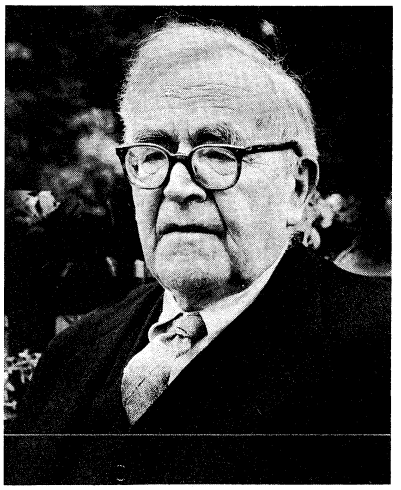
BIBLIOGRAPHY. IRVING WALLACE, *The Fabulous Showman: The Life and Times of P.T. Barnum* (1959); P.T. BARNUM, *The Life of P.T. Barnum, Written by Himself* (1855), *Barnum's Own Story*, ed. by W.R. BROWNE (1927), *Struggles and Triumphs: or, The Life of P.T. Barnum, Written by Himself*, ed. by G.S. BRYAN, 2 vol. (1927), last two works combined material from all versions of Barnum's autobiographies; JOEL BENTON, *Life of Honorable Phineas T. Barnum* (1891); H.W. ROOT, *The Unknown Barnum* (1927); M.R. WERNER, *Barnum* (1923).

(Ir.W.)

Barth, Karl

One of the most influential theologians of the 20th century, Karl Barth initiated a radical change in Protestant thought. He developed a "theology of the Word of God" in opposition to the anthropocentric (man-centred) theological writings of the 19th century that took their point of departure from man's reason and his innate religious and moral consciousness.

He was born May 10, 1886, in Basel, Switzerland, the son of Fritz Barth, a Reformed professor of church history and New Testament at Bern, and Anna Sartorius. He attended the Free Gymnasium at Bern, where as a young student he displayed a keen interest in history and military matters. Although there is no account of his conversion, the pastor under whom he was confirmed suggested that he take up the study of theology. At the age of 18 he began his studies: first at Bern, then at Berlin, Tübingen, and Marburg—the last three in Germany—during which time he was greatly influenced by the leaders of liberal theology, notably Adolf von Harnack and Wilhelm Herrmann.



Barth, 1965.
Horst Tappe—EB Inc.

For two years Barth served as an assistant minister in Geneva (1909–11) and then from 1911–21 as minister in the farming and working-class congregation in Safenwil (Aargau canton). In 1913 he married Nelly Hoffman, a talented violinist. Their children were a daughter, Franziska (1914); Markus (1915), a professor of New Testament; Christoph (1917), a professor of Old Testament; Mathias (1921), a theology student who died as a result of a mountain-climbing accident in 1941; and Hans Jakob (1925), a landscape architect. While at Safenwil, in close cooperation with Eduard Thurneysen, a lifelong friend and fellow theologian, Barth began to think through the situation of the church and theology burdened with the liberalism of 19th-century Protestantism. Shocked by the failure of the theology of his teachers in the face of social questions and World War I, he joined the Religious Socialist movement and sought to organize the workers of his congregation. Deeper reflection upon the real task of theology and the church led in 1919 to the publication of *Der Römerbrief* (*The Epistle to the Romans*), which, in six successive editions, shocked theologians of the early 1920s out of their complacency. During that period Barth stressed the "wholly otherness of God" in contrast to the rationalism, historicism, and psychologism that prevailed in liberal Protestantism.

The publication of *Der Römerbrief*

The sensation created by this book brought the young Barth, who had never taken an earned doctoral degree, to the attention of academic theologians, and he was subsequently appointed to the chairs of theology at Göttingen (1921), Münster (1925), and Bonn (1930). Another result of the attention gained by the publication of *Der Römerbrief* was the formation of the "Dialectical school," composed of Thurneysen, Rudolf Bultmann, Friedrich Gogarten, Emil Brunner, and Georg Merz, all theologians who became influential in Protestantism and beyond, and the founding of the periodical *Zwischen den Zeiten* ("Between the Times"). Differences concerning the basis of evangelical theology began to appear among the members of the school, and a crisis eventually occurred with the rise of Adolf Hitler to power in January 1933.

From the outset Barth was a vigorous opponent of National Socialism and of the "German Christian" party within the German Evangelical Church. Through his pamphlet *Theologische Existenz heute* ("Theological Existence Today"), the first in a series under that name, he clarified the basic theological issues and rallied churchmen to resistance. With Martin Niemöller, an anti-Nazi church leader, and others he organized the Synod of Barmen (May 1934) at which was adopted the Barmen Declaration that became the confessional basis of the Confessing Church, which claimed to be the "evangelical church in Germany," in opposition to the established church that did not oppose National Socialism. The text of the declaration was almost entirely Barth's work. Its first article epitomized his theological position:

> Jesus Christ, as He is attested for us in Holy Scripture, is the one Word of God which we have to hear and which we have to trust and obey in life and in death.

Refusal to take an unconditional oath of allegiance to Hitler led to Barth's suspension at Bonn and was the occasion of his accepting a chair of theology at Basel. From there he continued his fight against Nazism. Prior to and during World War II he wrote letters of encouragement and admonition to the churches and their leaders in many lands and voluntarily enlisted for service in the Swiss army.

At Basel, Barth continued to work on *Church Dogmatics*, which he had begun at Bonn. Although never completed, it runs to four volumes (13 parts) of over 9,000 pages. A truly ecumenical work, filled with new insights and a wealth of exegetical, historical, philosophical, and dogmatic material, it is regarded by many Protestant and Roman Catholic scholars as the classical theological work of the century. Basic to all his writings has been his concern with the task of the preacher who from Sunday to Sunday is called to proclaim not man's word but the Word of God. He delivered over 500 sermons, most of which are yet to be published posthumously. In his later years he preached almost exclusively in the Basel prison as "a prisoner among prisoners."

Political and social views

Even before the collapse of the Third Reich near the end of World War II, Barth was among the first to champion friendship with defeated Germany. Symbolic of this friendship were lectures he delivered in 1946 and 1947 amid the ruins of the University of Bonn, but he did not withhold the sharpest criticism of the development of German history from Frederick the Great, king of Prussia in the 18th century, to Otto von Bismarck, founder and first chancellor of the German Empire in the 19th century, and to Adolf Hitler, founder of the Third Reich. Indeed, he declared that the German people "suffers from the legacy of the greatest Christian German: from the error of Martin Luther with respect to the relation of law and Gospel, of temporal and spiritual power, by which its natural paganism has not been so much limited and restrained as it has been ideologically transfigured, affirmed and strengthened."

Soon thereafter, Barth's polemic was directed against those inside and outside the church who were advocating what amounted to an "anti-Communist crusade." He took a stand for peace, for the abolition of the "iron curtain" between East and West, against equating the totalitarianism of the Soviet Union with that of Nazi Germany, and against the use of nuclear bombs—not because he had a love of Communism but because he had a penetrating insight into the Phariseeism (legalism and belief in a moral superiority) of anti-Communism. He anticipated by 20 years a changed stance of the Roman Catholic and Protestant churches toward Communism and the political thawing of the cold war.

International reputation and influence

Indicative of his international reputation and influence, as well as of the ecumenical significance of his work, Barth's travels took him to France, Italy, The Netherlands, England, Scotland, Hungary, Romania, Czechoslovakia, and, in 1962, the United States. In 1948 he delivered one of the major addresses at the opening of the first meeting (in Amsterdam) of the World Council of Churches and, following the second Vatican Council (1962–65), made a special trip to Rome in order to be better informed about the renewal of the Catholic Church. Barth was the recipient of honorary degrees from the universities of Münster, Glasgow, Edinburgh, St. Andrews, Oxford, Budapest, Geneva, Strasbourg, Paris, and Chicago and was honorary senator of the University of Bonn, honorary professor of universities in Hungary and Romania, an honorary member of the British and Foreign Bible Society and of the Académie des Sciences Morales et Politiques of the Institut de France, and holder of the British King's Medal for Service in the Cause of Freedom.

Characteristics of the man were an uncompromising devotion to the Gospel of Christ and kindliness even toward those with whom he sharply disagreed, an insatiable intellectual curiosity combined with a probing, criti-

cal mind, and a humility and cheerfulness born of a sense of the goodness of the Creator and his creation—reflected in a childlike enjoyment of the music of Mozart. For Barth, theology was "a peculiarly beautiful science" and a joyful task because its object is the indescribably good news of "the beauty of the Lord our God" in the humiliation and exaltation of Jesus Christ. And—in the words of Keats—"a thing of beauty is a joy forever; its loveliness increases." On December 9/10, 1968, Karl Barth died in Basel, his native city.

MAJOR WORKS

THEOLOGICAL WRITINGS: *Das Wort Gottes und die Theologie* (1924; *The Word of God and the Word of Man*, 1957); *Die Theologie und die Kirche* (1928); *Fides quaerens intellectum: Anselms Beweis der Existenz Gottes* (1931; *Anselm: Fides quaerens intellectum*, 1960); *Kirchliche Dogmatik* (1932; *Church Dogmatics*, 1961); *Credo: Die Hauptprobleme der Dogmatik dargestellt im Anschluss an das Apostolische Glaubensbekenntnis* (1935; *Credo: A Presentation of the Chief Problems of Dogmatics with Reference to the Apostles' Creed*, 1936); *Evangelium und Gesetz* (1935); *Gotteserkenntniss und Gottesdienst nach reformatorischer Lehre* (1938; *The Knowledge of God and the Service of God According to the Teaching of the Reformation*, 1938); *Dogmatik in Grundriss* (1947; *Dogmatics in Outline*, 1949); *Christus und Adam nach Röm 5* (1952; *Christ and Adam: Man and Humanity in Romans 5*, 1957).

BIBLICAL EXEGESIS: His best known commentaries include *Der Römerbrief* (1919; *The Epistle to the Romans*, 1933); and *Erklärung des Philipperbriefes* (1927; *The Epistle to the Philippians*, 1962).

BIBLIOGRAPHY. KARL KUPISCH, *Karl Barth in Selbstzeugnissen und Bilddokumenten* (1971), although only 156 pages in length, is the fullest account of Barth's life and work extant and is supplied with copious quotations from his letters and books, and with documentary pictures. GEORGES CASALIS, *Portrait de Karl Barth* (1960; Eng. trans. with an introduction by ROBERT MCAFEE BROWN, 1963), is not so much a biography as a portrait of the man and his work. See also the articles on Barth by W. MATTHAIS in *Evangelisches Kirchenlexikon*, vol. 1 (1956); and by G. GLOEGE in *Die Religion in Geschichte und Gegenwart*, 3rd ed., vol. 1 (1957). T.H.L. PARKER, *Karl Barth* (1970), is an account of Barth's spiritual pilgrimage and of the development of his theological method. A definitive biography is being projected by the Karl Barth Foundation in Basel, together with some 40 volumes of unprinted material that Barth left to the executors of his literary estate. A complete edition of his works will run to some 70 volumes, exclusive of the *Church Dogmatics*.

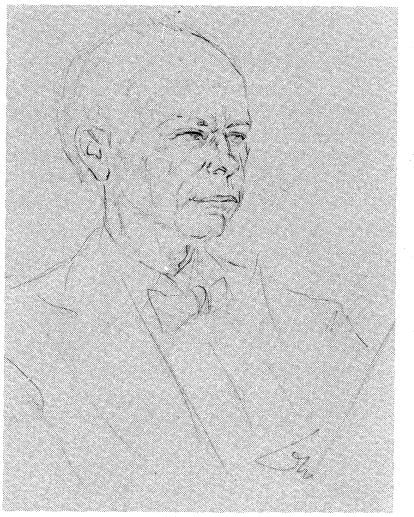
(A.C.C.)

Bartók, Béla

The significance of Béla Bartók lies in four major areas of music—composition, performance, pedagogy, and ethnomusicology. As a composer, of a stature equalled by few in the first half of the 20th century, he fused the essence of Hungarian and related folk music with traditional music to achieve a style that was at once nationalistic and deeply personal. As a pianist he gave concerts in Europe and America, disseminating the newer Hungarian music. As a teacher he helped train generations of pianists, both Hungarian and foreign. And as an ethnomusicologist he was one of the first to examine folk music with attention to its historical and sociological implications, and he helped to lay the foundations for the study of comparative musical folklore in Hungary.

Career in Hungary. Bartók was born in 1881 in Nagyszentmiklós, a small town then in Hungary and now in Romania. He spent his childhood and youth in various provincial towns. Studying the piano with his mother and later with a succession of teachers, he began to compose small dance pieces at the age of nine. Two years later he played in public for the first time, including a composition of his own in his program.

Following the lead of another eminent Hungarian composer, Ernő Dohnányi, Bartók undertook his professional studies at the Royal Hungarian Academy of Music in Budapest rather than in Vienna. He developed rapidly as a pianist but less so as a composer. After writing no music at all for two years, he resumed composing in 1902 under the stimulus of his discovery of the music of Richard Strauss. At the same time a spirit of optimistic



Bartók, portrait by B.F. Dolbin, 1944.
By courtesy of Andre Meyer

nationalism was sweeping Hungary, inspired by Ferenc Kossuth and his Party of Independence. As other members of Bartók's generation demonstrated in the streets, the 22-year-old composer wrote a symphonic poem, *Kossuth*, portraying in a style that was reminiscent of Strauss, though with a Hungarian flavour, the life of the great patriot Lajos Kossuth, Ferenc's father, who had led the revolution of 1848–49. Despite a scandal at the first performance, occasioned by a distortion of the Austrian national anthem, the work was received enthusiastically.

Shortly after Bartók completed his studies in 1903, he and the Hungarian composer Zoltán Kodály, who collaborated with Bartók, discovered that what they had considered Hungarian folk music and drawn upon for their compositions was instead the dilettante music of city-dwelling gypsies. A vast reservoir of authentic Hungarian peasant music was subsequently made known by the research of the two composers. The initial collection, which led them into the remotest corners of Hungary, was begun with the intention of revitalizing Hungarian music. Both composers not only transcribed many folk tunes for the piano and other media but incorporated into their original music the melodic, rhythmic, and textural elements of peasant music. Ultimately, their own work became suffused with the folk spirit.

Uncovering authentic peasant music

Bartók was appointed to the faculty of the Academy of Music in 1907 and retained that position until 1934, when he resigned to become a working member of the Academy of Sciences. His holidays were spent collecting folk material, which he then analyzed and classified, and he soon began the publication of articles and monographs.

At the same time, Bartók was expanding the catalog of his compositions, with many works for the piano, a substantial number for orchestra, and the beginning of a series of six string quartets that was to constitute one of his most impressive achievements. The first quartet (1908–09) shows few traces of folk influence, but in the others that influence is thoroughly assimilated and omnipresent. The quartets parallel and illuminate Bartók's stylistic development; in the second quartet (1915–17), Arab elements reflect his collecting trip to North Africa; in the third (1927) and fourth (1928), there is a more intensive use of dissonance; and in the fifth (1934) and sixth (1939), there is a reaffirmation of traditional tonality.

In 1911 Bartók wrote his only opera, *Duke Bluebeard's Castle*, an allegorical treatment of the legendary wife murderer, with a score permeated by characteristics of old Hungarian folk song, especially in the speechlike rhythms of the text setting. The technique is comparable to that used by the French impressionist composer Claude Debussy in his opera *Pelléas et Mélisande* (1902), and Bartók's opera has other impressionistic qualities as well. A ballet, *The Wooden Prince* (1914–16), and a pan-

tomime, *The Miraculous Mandarin* (1918–19), followed; thereafter he wrote no more for the stage.

Unable to travel during World War I, Bartók devoted himself to composition and the study of the collected folk music. During the short-lived proletarian dictatorship of the Hungarian Soviet Republic in 1919, he served as member of the Music Council with Kodály and Dohnányi. Upon its overthrow Kodály was removed from his position at the Academy of Music; but Bartók, despite his defense of his colleague, was permitted to remain.

Period of greatest activity

His most productive years were the two decades that followed the end of World War I in 1918, when his musical language was completely and expressively formulated. He had assimilated many disparate influences: in addition to those already mentioned—Strauss and Debussy—there were also the 19th-century Hungarian composer Franz Liszt and the modernists Igor Stravinsky and Arnold Schoenberg. Bartók arrived at a vital and varied style, rhythmically animated, in which diatonic and chromatic elements are juxtaposed without incompatibility. Within these two creative decades, Bartók composed two concertos for piano and orchestra and one for violin; the *Cantata Profana* (1930), his only large-scale choral work; and a number of important chamber scores, including the *Music for Strings, Percussion, and Celesta* (1936), and the *Sonata for Two Pianos and Percussion* (1937). The same period saw Bartók expanding his activities as concert pianist, playing in most of the countries of western Europe, the United States, and the Soviet Union.

U.S. career. As Nazi Germany expanded its sphere of influence in the late 1930s, and Hungary appeared in imminent danger of capitulation, Bartók found it impossible to remain there. After a second concert tour of the United States in 1940, he emigrated there later the same year. An appointment as research assistant in music at Columbia University enabled him to continue working with folk music, transcribing and editing for publication a collection of Serbo-Croatian women's songs, a part of a much larger recorded collection of Yugoslav folk music. With his wife, the pianist Ditta Pásztory, he was able to give a few concerts. His health, however, was never very strong and had begun to deteriorate even before his arrival in America. His last years were marked by the ravages of leukemia, which often prevented him from teaching, lecturing, or performing. Nonetheless, he was able to compose the *Concerto for Orchestra* (1943), the *Sonata for Solo Violin* (1944), and all but the last measures of the *Third Piano Concerto* (1945). When he died in New York on Sept. 26, 1945, his last composition, a *Concerto for Viola and Orchestra*, was left an uncompleted mass of sketches.

Bartók's heritage. During his life Bartók published several important book-length studies of Hungarian and Romanian folk music. His three-volume study of Romanian folk music was issued only in 1967; the first of three volumes on Slovakian folk music appeared in 1959. These contributions loom large in the field of musical ethnology, and it may be argued that they overshadow Bartók's legacy as composer.

Though his music was infrequently performed outside Hungary during his lifetime, many of his compositions, including the string quartets and the *Concerto for Orchestra*, later entered the standard concert repertory. Bartók's compositions within a quarter century after his death were ranked among the classics of Western music.

MAJOR WORKS

Orchestral music

Kossuth, symphonic poem (1903); *Suite No. 1*, op. 3, for large orchestra (1905); *Suite No. 2*, op. 4, for small orchestra (1905–07); *Music for Strings, Percussion and Celesta* (1936); *Divertimento*, for string orchestra (1939); *Concerto for Orchestra* (1943).

CONCERTOS: *Violin Concerto No. 1* (1907–08), *No. 2* (1937–38); *Piano Concerto No. 1* (1926), *No. 2* (1930–31), *No. 3* (unfinished, 1945); *Viola Concerto* (unfinished, but completed by Tibor Serly, 1945).

Chamber music

STRING QUARTETS: *No. 1*, op. 7 (1908–09); *No. 2*, op. 17 (1915–17); *No. 3* (1927); *No. 4* (1928); *No. 5* (1934); *No. 6* (1939).

MISCELLANEOUS: *Piano Quintet* (1904); *Sonatas for Violin and Piano No. 1* (1921), *No. 2* (1922); *Rhapsodies for Violin and Piano No. 1 and 2* (1928); *Forty-four Duos*, for two violins (1931); *Sonata for Two Pianos and Percussion* (1937; transcribed as *Concerto for Two Pianos and Orchestra*, 1940); *Contrasts*, for violin, clarinet, and piano (1938); *Sonata for Solo Violin* (1944).

PIANO SOLOS: *Rhapsody*, op. 1 (1904; also arranged for piano and orchestra and for two pianos); *Fourteen Bagatelles*, op. 6 (1908); *For Children* (1908–09); *Allegro barbaro* (1911); *Sonatina* (1915); *Suite*, op. 14 (1916); *Sonata* (1926); *Out of Doors*, suite (1926); *Mikrokosmos*, 153 progressive pieces for piano (1926–39).

Vocal music

STAGE WORKS: *Duke Bluebeard's Castle*, op. 11, an opera, libretto by Béla Balázs (Budapest, 1918); *The Wooden Prince*, op. 13, ballet, libretto by Béla Balázs (Budapest, 1917); *The Miraculous Mandarin*, op. 19, pantomine, libretto by Menyhért Lengyel (Cologne, 1926).

CANTATA: *Cantata Profana: The Nine Enchanted Stags*, for double mixed chorus, tenor and baritone soloists, and orchestra (1930).

SONGS: A large number of settings of Hungarian and other folk songs, including *Five Village Scenes*, for voice and piano (1924).

BIBLIOGRAPHY. A listing of the Bartók compositions and papers held by the Béla Bartók Archives in New York appears in VICTOR BATOR, *The Béla Bartók Archives: History and Catalogue* (1963). The composer's letters have been collected and edited by JANOS DEMENY in *Bartók Béla levelek* (1948), and *Bartók Béla levelei* (1951 and 1955); in *Ausgewählte Briefe* (1960); and in *Bartók Letters* (1971). HALSEY STEVENS, *The Life and Music of Béla Bartók*, rev. ed. (1964), is a biographical study and critical examination of all published works, with a catalog of compositions and an extensive bibliography.

(H.Ss.)

Baseball

Baseball (originally written "base ball") is a contest between two teams of nine players each. Four bases are laid out at angles of a 90-foot (27.4-metre) square at one end of a broad field of play. Teams alternate as batters ("ins") and fielders ("outs"), exchanging places when three of the batting side are "put out." As batters, players attempt to hit a ball served by a pitcher out of reach of the fielding side, and to run from base to base counterclockwise. A complete circuit counts one run, and victory goes to the side tallying the most runs at the end of nine innings (times at bat) for each team.

Baseball is played in most countries of the world, but is most popular in Latin American countries, Japan, and the United States, where it is, by tradition alone, accepted as the national game. As such, it is played informally by schoolboys on their playgrounds and in organized leagues, by teams representing community groups —schools and colleges, churches, department stores, employed personnel of factories, shipyards, mines, railroads, etc.—and by teams representing service units of the armed forces. While in the case of business firms some of the players receive financial compensation for playing, the term professional baseball player applies only to athletes employed as players by teams that are members of leagues subscribing to the authority and regulations governing what is known as organized baseball.

Baseball belongs to the extensive genus of civilized athletic competitions stemming perhaps from the primitive play urge to hit a fragment of rock, a piece of wood, or other object with a club. Two boys with one stick of wood and a rubber ball portray baseball in its essence when they mark two fixed points on their playground with a couple of flat stones as bases and take turns hitting the ball with the bat. One, armed with the bat, stands at one of the bases. The other stands an agreed distance away and tosses the ball to the batter, who drives it as far as he can. How many times the batter must run between bases before he is credited with a run, or a score, is usually agreed upon between the two contestants. Each keeps count of his runs and, after each has had the same number of turns at batting, the one who has scored the most runs is the winner.

Description of the game

Baseball in all the leagues, in the big city stadiums with

attendances of 50,000 or more, is but a refined and more intense form of the same competition. Nine-man teams oppose each other instead of just two individuals. There are four bases instead of two. In scoring a run, a batter must go to each of the other three bases before returning to his starting point to complete his run, but he may perch temporarily on a base on his way around until the opportunity is offered to continue his course.

Just as each batter has eight teammates to help him, there are nine players to pursue the batted ball instead of only the boy who in the playground miniature game had to be both thrower and retriever.

The official baseball rules, uniformly in force internationally and unchanged in any conspicuous feature since 1920, prescribe the dimensions and markings of the playing field, the conditions under which the teams function when at bat and when in the field, and the law of the game as it has evolved to cover every conceivable eventuality of the sport; variations in playing conditions in individual parks are provided for by local ground rules.

Innings and completion, termination, and suspension of games

The visiting team always has the first inning at bat in modern baseball, the home team's inning coming second. In common parlance, "inning" includes an inning at bat for each team. Originally a game was described as consisting of nine innings for each side. Usage has erased the last three words. A game, in modern baseball language, consists of nine innings, and a team's inning, for example, the sixth, is now its "half of the sixth." In written accounts of ball games a play is said to have occurred "in the first half of the sixth," or "in the top of the sixth," which conveys to the reader the information that the visiting team was at that time at bat in its sixth inning. If the home team was batting, the writer would place the play "in the last half of the sixth," or, more likely, "in the bottom of the sixth."

Whichever team has scored the most runs at the completion of a game is the winner. A game is completed at the end of nine innings if either team has scored more runs than the other, or at the end of 8½ if the home team is ahead at that point. If the score is even after nine full innings, play continues until one side has more runs than the other in equal innings at bat. Starting with the latter half of the ninth inning, if the home team acquires one run more than the visiting team, the game terminates at the moment the winning run is scored—unless that run is driven in by a home run, in which case any runners on base at that time may score ahead of the home run hitter, whose run becomes the final transaction of the game. A game may be terminated by the umpire because of rain, darkness, or other causes that in his judgment interfere with further play. If at least five equal innings have been played, the umpire may declare the game a tie or a decided game, depending on the score. A game may be suspended if there is light failure, if a curfew requires play to stop, or if league time limits are reached, and the suspended game is then completed at a later date.

The balance of this article deals primarily with the sport as played professionally in the United States. Following are the main sections and subsections of this article:

I. History of baseball
- Origin of the game
- Later history
- Organized baseball
- Competition
- Amateur baseball

II. Play of the game
- Grounds and equipment
- Conduct of the game
- Principles of play

I. History of baseball

ORIGIN OF THE GAME

Evolution from rounders. In the early days of modern U.S. baseball, no one who wrote of it seemed to doubt that it was an evolution from an English children's game known then and thereafter as rounders. A simple change in the rules, according to these authorities, transformed it into a man's game. In rounders, as then played, the fielder put out a runner running for a base, or caught off base, by throwing the ball at him and hitting him with it. This precluded the use of a hardball and, since a softball cannot be batted very far, limited both the size of the field and the activity of the players. Then, presumably about 1840, some American had the idea of putting out the runner by tagging him—touching him with the ball or with the hand holding it. It became immediately possible to use a hardball—at first, a kind of miniature cricket ball. And the game suddenly grew up. The rules of the pioneer Knickerbocker Baseball Club of New York, drawn up in 1845, constitute the earliest known documentary record of this change.

But many of the old-time players, and especially A.G. Spalding, who had made a fortune out of sporting goods, refused to entertain the thought that a foreign nation could have had anything to do with inventing a great U.S. institution such as baseball. Hence the Spalding Commission, composed not of skilled investigators but of baseball men assisted by a United States senator, was appointed ostensibly to investigate the origins of the game but really to prove its exclusively U.S. origin. In 1908 this commission reported in the official *Baseball Guide* that the game, including the essentials of the modern rules, the dimensions of the field, and even the name, was invented in 1839 at Cooperstown, New York, by Abner Doubleday—afterward General Doubleday, a hero of the Battle of Gettysburg—and that the foundation of this invention was an American children's game called one old cat. As most of the pioneer players were dead before 1908, this report caused so little controversy at the time that gradually even the standard reference books accepted its conclusions as seasoned history. Effective attempts to refute its data did not begin until 1939, the centenary of the alleged invention, when Robert W. Henderson, of the New York Public Library, issued a pamphlet embracing some scholarly researches of his own, all casting doubt on the Doubleday theory and strengthening the case for an English origin. Others have since added corroborative details.

The main points in support of this contention are as follows: the word baseball to designate some popular English game has been traced back to the first half of the 18th century. In the *Letters* of Mary Lepel, Lady Hervey, there occurs a passage under the date of November 14, 1748, referring to ". . . base-ball, a play all who are, or have been, schoolboys, are well acquainted with." In *Northanger Abbey* (written about 1798) Jane Austen remarks of her heroine, "It was not very wonderful that Catherine . . ., should prefer cricket, base ball, riding on horseback, and running about the country, at the age of fourteen, to books."

Evidences of English origins

But before this, the name and presumably the game were already known in America. In 1744 *A Little Pretty Pocket-Book* was published in England. Illustrated with crude woodcuts, it pictures and describes in doggerel quatrains 26 children's sports, one for each letter of the alphabet; and *b* is represented by "Base-Ball." The text records that the batter hits the ball and runs from base to base. The illustration, which comprises only part of the field, shows a player at the plate, holding a bat with a curious flat, fanlike end; a catcher behind him; a pitcher preparing to throw a small ball underhand; and two bases, marked by posts instead of bags, with a baseman beside each of them. This book, extremely popular in England, was reprinted twice in America—in New York City (1762) and Worcester, Massachusetts (1787).

Other later references present good documentary evidence of an American game of this kind. The journal of George Ewing, a soldier, written at Valley Forge in 1778, tells of "playing at base." Some boys "playing at ball" in the Wall Street region of New York abandoned their game to join one of the riots that preceded the American Revolution. In 1787 the faculty of Princeton College forbade the students to "play with balls and sticks in the back common of the college." Thurlow Weed, editor and politician of Rochester, New York, mentioned in his memoirs "a baseball club organized about 1825." Corroborating this, a newspaper item dating from the 1820s states that the Rochester Baseball Club, with about 50

members, was in practice for its season's activities. The elder Oliver Wendell Holmes (Harvard 1829) mentioned to an interviewer that he had played a good deal of baseball while in college at Cambridge, Massachusetts. These are only a few selected references from a number discovered incidentally and accidentally. Systematic search of books, documents, and records before 1830 would probably reveal many others.

Then came an unpretentious document that ties English rounders to American baseball in their primitive forms. *The Boy's Own Book*, a treatise on boys' sports and their rules, published in London in 1828, was so popular that it ran through many editions, the second of which includes a chapter entitled "Rounders," with a note that the game is called "feeder" in London. As there described, it bears a far closer family resemblance to modern U.S. baseball than does rugby football to its acknowledged offspring, U.S. intercollegiate football. It was played on a diamond with a base at each corner, the goal or fourth base being identical with the plate beside which the batter stood. The batter might run whenever he hit the ball across or over the diamond; if he hit it in any other direction, this constituted a foul, and he was not permitted to run. If he struck at it and missed it three times, he was out. A batted ball caught on the fly constituted an out. When a runner made the circuit of the bases, it counted one point or tally for his side. Then came the one vital difference that distinguished this from modern baseball. When a grounder was fielded, the fielder put the runner out by throwing the ball at him and hitting him with it. The same rule applied to a runner caught off base. The woodcut accompanying the article bears a strong resemblance to that illustrating the rhyme "Base-Ball" in *A Little Pretty Pocket-Book*. Even the bat is the same strange-looking implement. Objects that look like flat stones serve as bases, however, and there seems to be a second catcher to range for fly fouls.

The game in America. Many English immigrants to America in colonial times were from those south counties of which Mary Lepel and Jane Austen were residents. That the old game was called baseball in that region, as it was called feeder in London and rounders in western England, and that south-county immigrants took both the name and the game to America seems a tenable hypothesis. There would seem to be little doubt that modern-day baseball evolved from such sport just as so many other present pastimes did from games of bygone eras.

As for the "old cat" versions, the critics of the Spalding report hold that in the 18th century those games were what they are today—substitutes for baseball when the boys did not have enough players for two full teams. In one old cat there was one base, one pitcher, one catcher, and one batter. The more men available the higher goes the number of "old cats"—two, three, etc. "Town ball" seems to have been only another American name for rounders, feeder, or old-time baseball, and near Boston the sport was known as the "Massachusetts game" at the turn of the 19th century. Reminiscences in the sporting periodicals of the 1850s and 1860s prove that the game was played in New England even before 1833, when Philadelphia had a town ball club and drew up written rules. New England, however, unlike Philadelphia, distinguished between fair and foul balls. On the other hand, in part of New England, home base and batter's plate were not identical, but stood a few feet apart from each other, so that the lineup included a fourth baseman. Henderson's writings reveal that the first U.S. book on baseball was *The Book of Sports*, written by Robin Carver in Boston in 1834. In it the author credits much of his material to the London publication *The Boy's Own Book*, published six years earlier. Carver's book goes into details of the game of rounders, including an illustrated diagram for the placing of the posts (bases) in the shape of a diamond. Carver copied the English rules of rounders almost word for word, and yet he called his game base, or goal ball. Doubleday, Alexander J. Cartwright, a New York surveyor and amateur ballplayer, Harry Wight, who played as an amateur with the New York Knickerbockers and became a professional player and manager (Cincinnati and Boston), and Spalding have each been referred to as the "father of baseball." Doubleday no doubt did much to promote interest in the sport, probably while a cadet at West Point, for some sports historians place his years at the military academy as 1838–42. Unfortunately, the notes, statements, and affidavits gathered by the Spalding Commission were afterward lost in a fire, and only the finished report is extant. In this, an old resident named Abner Graves figured as the chief witness; and he testified that in the game Doubleday taught in Cooperstown, the fielder put out the runner by hitting him with the ball. This means, of course, that Doubleday's game was still much like the old English version and renders unlikely the statement that Doubleday laid out a diamond with the exact dimensions of the modern playing field, for any ball light and soft enough to be thrown full force against a player's head or body without danger of inflicting serious injury could scarcely have been batted out of an infield so large. Cartwright, who long had played the game with society friends about New York, became discouraged with the haphazard manner in which contests were conducted and organized a group to formulate a code of standard rules in 1845. Drawing heavily on Carver's book, the group submitted a set of regulations, which were adopted in September 1845; much of that original code is applicable today. It was then that tagging out a runner was adopted, which no doubt paved the way for the introduction of the hardball that is the key factor of modern baseball.

From a few of the 1845 rules it may be seen how closely they paralleled those in force in organized baseball of today; for example: (1) a ball knocked outside the range of first or third (outside foul lines today) is foul; (2) three balls being struck at and missed, and the last one caught is a hand out (one out), but if not caught, it is considered fair and the striker is bound to run; (3) three hand outs, all out (batting side retired); and (4) a player running the bases shall be out if the ball is in the hands of an adversary on the base and the runner is touched by the ball before he makes his base, it being understood, however, that in no instance is the ball to be thrown at him (a definite departure from rounders).

With the introduction of the new rules by the New York Knickerbockers, the sport gained quickly in popularity. The "Knicks," who were about to be ousted from their Manhattan playing field, found a new site across the Hudson River in Hoboken, New Jersey, and made it their home field for the 1846 season. Their rules soon were taken up by such clubs as the Gothams, Eagles, Empires, and Mutuals. The Olympic Club of Philadelphia, although it had been playing town ball since 1833, did not switch to the Cartwright code until 1860. At about the time the new version of baseball was being popularized around the New York area, the old softball variety had a sudden spurt in popularity in Boston. Until the U.S. Civil War, the game was called, respectively, Boston baseball and New York baseball. During the Civil War, however, the New York and New Jersey regiments taught their own versions of the sport to other Federal soldiers, and when the war ended the more rugged and adult New York game held the field as the sport for grown men.

The original code of rules

One old cat and town ball

LATER HISTORY

Emergence of organized clubs. From 1845 until 1854 baseball was played according to rules first written for the game. In 1854 a revision provided specifications for the size and weight of the ball. On March 10, 1858, the first attempt was made at organization of clubs, their number having greatly increased and the game having expanded throughout the territory about New York City, extending to Philadelphia. Massachusetts still played town ball. In 1859 Washington, D.C., organized a baseball club, followed in 1860 by Lowell, Massachusetts, Allegheny, Pennsylvania, and Hartford, Connecticut. From that time the game became more widespread, going to Maine, Kentucky, and in 1866 to Portland, Oregon. Baseball was played in towns and hamlets other than the cities mentioned during this period, but the clubs

First attempts at organization

were considered to be town or minor clubs as compared with organized clubs, a distinction that has since followed the progress of professional baseball. The National Association of Base Ball Players, organized in 1858, embraced 16 clubs in New York City. The Knickerbocker playing rules were amplified, and the ball and bat were made to conform to measurements approximating those of the present day. In 1863 the rules were further amplified. From 1861 to 1865 baseball languished except in the armies of the Civil War. In 1865 a convention was held in New York at which 91 clubs were represented, including those from the cities of St. Louis, Missouri; Chattanooga, Tennessee; Louisville, Kentucky; Washington, D.C.; Boston; and Philadelphia. It was strictly an amateur organization without schedules for games, and its purpose was to preserve the stability of the rules and the amateur status of the sport.

Appearance of professionalism. In 1865 and 1866 professionalism began to make an appearance. Even then, players did not derive their livelihood from baseball, but the more expert accepted sums of money as members of any club that would engage them for occasional games. This was a new development, somewhat unexpected, and it seriously perturbed the amateur players. Gentlemen players openly avowed their objection to what they expressed in caustic language as deterioration. In addition to the disposition of some players to accept hire for their services, open pool selling and bribery by gamblers, some of which was successful, outraged many amateur players and organizers of clubs. This conflict between amateurism and professionalism eventually led to a professional organization, the first of its kind, which was a puny affair compared with the later great associations of clubs in leagues.

In 1867 the Nationals of Washington made the first trans-Allegheny tour, going as far west as St. Louis. In that year the Rockford, Illinois, club began the practice of paying salaries to some of its players. In 1868 the Cincinnati team was organized on what were known as semiprofessional lines, a characterization of athletes peculiar to the United States. A semiprofessional does not play baseball for a living, but is hired for occasional games.

The first all-professional team

In 1869 the Cincinnati team was hired as an outright professional organization and made a successful tour of the United States from New York to San Francisco. The Cincinnati team did not lose a game that year and was undefeated until June 14, 1870. During the successful career of the Cincinnati team, rival clubs became imbued with an eager desire to win from it, throwing aside all restraint of policy relative to being amateurs in order to engage the best players available. Most of the important clubs abandoned every attempt to preserve an amateur standing, despite the fact that they were members of the National Association, which was an amateur body. Once professionalism had entered the game, it was impossible to keep the professional and amateur exponents of the sport reconciled. The organization of the professional Cincinnati team was followed by that of another professional team at Chicago, and in 1870 the National Association of Base Ball Players was disrupted, the pure amateurs withdrawing from the annual meeting. A new amateur organization was effected in 1872 but dissolved in 1874, the last of the concerted attempts to keep the control of the game within amateur influence. In 1871 the National Association of Professional Base Ball Players was organized in New York. This embraced the Athletics of Philadelphia; Bostons of Boston; White Stockings of Chicago; Eckfords of Brooklyn, New York; Forest Citys of Cleveland; Forest Citys of Rockford, Illinois; Haymakers of Troy, New York; Kekionagas of Fort Wayne, Indiana; Olympics of Washington, D.C.; and Mutuals of New York. The affairs of the organization were loosely conducted. The circuit was not preserved intact. Many scheduled games never were played. Gambling and contract breaking became so repulsive to players of higher principle and to certain owners that they withdrew. The association dissolved in 1876 when the National League of Professional Base Ball Clubs came into existence with the seceders from the association backing it. It was organized in New York City, February 2, 1876, with a membership made up of Philadelphia, Hartford, Boston, Chicago, Cincinnati, Louisville, St. Louis, and the Mutuals of New York City.

William A. Hulbert, who became president of the league in 1877, expelled four ballplayers found guilty of dishonesty from baseball for life; and from that time, confidence was established in the professional branch, and amateur baseball grew with the revived interest in professional baseball. Hulbert remained president of the league until his death in 1882. During his administration baseball developed sufficiently to be regarded as an institution.

Formation of leagues. In 1882 the American Association was formed in cities not members of the National League circuit. National League owners attempted to equalize salaries of players, regardless of cities and local conditions, by a uniform scale of hire. The players opposed the plan, and in 1890, after forming a league known as the Players' League, took the field against the National League. In one year the player organization was wrecked. In 1891, the American Association engaged the National League in open rivalry, a venture hopelessly destined to failure; finally, in the winter of 1891 the American Association was merged with the National League into a 12-club organization having a monopoly of major league baseball.

As "the" major league, the National operated with 12 clubs from 1892 to 1899, inclusive. Falling off of patronage in Baltimore and Cleveland, mainly due to the weakening of the two teams' lineups by trades, resulted in the return to the eight-club membership, starting with the 1900 season. Baltimore, Washington, Cleveland, and Louisville were dropped as National League cities, leaving Boston, Brooklyn, Chicago, Cinncinnati, New York, Philadelphia, Pittsburgh, and St. Louis. This resumption of an eight-club setup by the one and only major league in existence led to the rise of a rival circuit, the Western League, organized as a minor league in 1893, with membership of Midwest cities covering about the same territory as the present-day American Association.

In 1900 Charles A. Comiskey, then owner of the St. Paul, Minnesota, club of the Western League, moved his team to Chicago, and renamed it the Chicago White Sox, although Chicago was a member of the National League, with continuous membership from the league's inception in 1876, when its team had been known as the Chicago White Stockings. In the same year, the Western League shifted its Grand Rapids, Michigan, team to Cleveland, one of the cities abandoned by the National League after 1899.

The move to Chicago by Comiskey received the assent of the National League, but when permission was sought to place teams for the 1901 season in Baltimore and Washington, the National League refused.

The baseball "war" in the early 20th century

Two years of baseball "war" followed. Having changed its name to the American League, the militant former Western Leaguers moved their Indianapolis (Indiana), Kansas City (Missouri), Minneapolis (Minnesota), and Buffalo (New York) clubs to Baltimore, Washington, D.C., Philadelphia, and Boston for the start of the 1901 season.

The American League withdrew from the National Agreement, the body of rules governing relations between all professional leagues, major and minor, including their transfers of players from one league to another and their territorial restrictions in the operation of clubs. Its 1901 lineup of eight cities included—besides the newly invaded four eastern cities—Chicago, Detroit, Cleveland, and Milwaukee to the west. In 1902 the Milwaukee club was moved to St. Louis, invading National League territory there; and in 1903 the Baltimore franchise was shifted to New York, establishing the American League's roster of cities as it remained through 1953.

In the "war" years, the American League made inroads on the National's galaxies of star players to such a degree that it became firmly established as a major league. When peace was made between the two leagues in January, 1903, the new agreement entered into by both gave each

of the two leagues equal importance. The agreement forbade the consolidation of two clubs occupying one city into one ownership, prohibited shifting of teams from one city to another by either league without the consent of the other, set rules for transfers of players from one major league to another, and reestablished the rules dealing with securing minor league players.

The "war" brought into prominence two men whose names became baseball tradition, Connie Mack (Cornelius McGillicuddy) and John J. McGraw. In 1900 Mack, a former National League catcher and manager of the Pittsburgh Pirates in 1894–96, was manager of Milwaukee when the Western League changed its name. In 1901 he was placed at the head of the American League's new Philadelphia team, the Athletics. In 1902 he won the first of a long line of championships with his Athletics. The success of the American League's rise to major-league status was largely attributed to his leadership and inspiration.

McGraw, a star third baseman of the Baltimore Orioles —National League pennant winners three straight years in the middle 1890s—was manager of Baltimore's American League club starting the 1902 season. In June, a National League countermove to the raids on its ranks by the American League induced McGraw to jump back to the National as manager of the New York Giants. For the next quarter of a century, McGraw and his Giants were headliners in the National League comparable in prominence to Mack and his Athletics in the American League.

The Federal League

During the period when organized baseball was governed by the National Commission (1903–21), there was one attempt by outside interests to conduct championship baseball on a major-league level independent of organized baseball. This outlaw organization was the Federal League, which conducted pennant races in 1914 and 1915.

Baltimore, Brooklyn, Buffalo, Chicago, Indianapolis, Kansas City, Pittsburgh, and St. Louis were represented in the 1914 season, with Indianapolis winning the pennant. In 1915 the Indianapolis team was transferred to Newark, New Jersey. Chicago won the 1915 pennant. The Federal League not only placed teams in organized baseball territory but also offered contracts to the outstanding players and managers of the two major leagues. Its raids were not as successful in securing stars as those of the American League had been. After the 1915 season came another "peace treaty," and the Federal League passed out of existence. Two of the men who had heavily backed the outlaw league joined the major leagues as owners of the Chicago National League and the St. Louis American League clubs, respectively. Players developed by Federal League clubs and not previously the "property" of organized baseball found places in organized baseball. Those who had left organized baseball to join up with the "Feds" returned to the clubs from which they had jumped.

Survival and growth of the game. The administrative structure of baseball and its future as a national game were severely tested in the period from 1919 through 1921. During these years eight members of one of the game's greatest teams of all time, the Chicago White Sox of 1919, were accused of accepting bribes to throw the World Series of that year. The incident became known as the "Black Sox" scandal. In September 1920, all were indicted on a charge of fraud. The presiding judge for the grand jury was Kenesaw Mountain Landis, who one year later was named baseball's first commissioner. Comiskey suspended the eight players for the 1921 season. In August of that year all were found not guilty by a jury, but Landis, then commissioner, banned them from organized baseball for life.

During this period George Herman ("Babe") Ruth was becoming baseball's newest hero. He began hitting balls a greater distance than any other hitter the game had known, and was a tremendous factor in arousing the fans' desire to see a new idol who would go down in history as the greatest of all home-run hitters. About fifty years later, Ruth's career record of 714 major league home runs, not counting All-Star and World Series games, had been challenged by Willie Mays—who retired in 1973 after hitting 660 homers with the New York and San Francisco Giants and the New York Mets—and finally broken in 1974 by Hank Aaron of the Atlanta Braves, whose total reached 745 by the end of 1975.

Black players

Mays and Aaron were among the dozens of Negro players to rise to stardom in the 1950s and 1960s. In 1947 infielder Jackie Robinson had become the first Negro player in the major leagues since Welday and Moses Walker, brothers, played in 1884 for Toledo in the American Association, then a major league. Robinson, brought into organized baseball by Branch Rickey, president of the Dodgers, then in Brooklyn, helped the Dodgers win the pennant in his first year and received the National League's Most Valuable Player Award in his third. The Cleveland Indians followed Rickey's lead with Larry Doby, slugging outfielder, and, in 1948, with Leroy ("Satchel") Paige, then past 40, famous pitching star of the Negro leagues.

Within a few years all of the major-league teams had added Negroes to their rosters, and there were also many in the minor leagues. This integration was accepted by the players and fans alike; for example, there was virtually no mention in the press when the Pittsburgh Pirates started a game with Negroes at all of the nine positions on September 1, 1971. Moreover, Negro players had won every possible individual honour. In addition to Mays and Aaron, others to earn superstar status were pitcher Bob Gibson of the St. Louis Cardinals, first baseman Willie McCovey of the San Francisco Giants, first baseman Ernie Banks of the Chicago Cubs, and outfielders Roberto Clemente of the Pittsburgh Pirates and Frank Robinson of the Los Angeles Dodgers.

Negroes had been barred from the major leagues for more than a half century; nonetheless, there were many Negro baseball players. Some of them—such as Josh Gibson, Bill Yancey, John Henry Lloyd, Andrew ("Rube") Foster, and "Cool Papa" Bell—most certainly would have been major-league stars of the first magnitude if they had been white. Gibson, a catcher who died in 1947 when Jackie Robinson was a rookie with the Brooklyn Dodgers, was an extraordinary power hitter in the same class with Babe Ruth. Gibson was credited with hitting 89 home runs in one season and 75 in another, many of them against semipro competition, however. According to one legend, Gibson hit a ball so high and far that no one saw it come down. The umpire, after scanning the sky for a few minutes, called it a home run.

There had been several Negro players in the minor leagues in the 19th century, and in the 20th century many Negroes played winter ball in Cuba, Mexico, and Venezuela. The first professional Negro team, the Cuban Giants, of Long Island, was organized in 1885. A number of other Negro teams were formed, and the Negro National League was set up in 1920 and the Negro Eastern League in 1921, inaugurating a world series in 1924; but these failed in 1932 because of the depression. A second Negro National League, founded in the late 1930s, was dissolved in 1952. The Negro American League, founded in 1936, formed an eastern and a western division in 1952 to conduct the annual Negro East–West Game.

After World War II, the Negro leagues suffered the decline in attendance and revenue that affected the minor leagues of organized baseball, accentuated by the loss of outstanding players to formerly all-white teams.

ORGANIZED BASEBALL

The major leagues. Baseball's most skillful players are professionals who perform, under exclusive contract, for one or another of the clubs that make up the major leagues. In the early 1970s, salaries for the season, which begins in early April and closes in early October, ranged from $12,750 (minimum) to more than $225,000. While not more than nine players may be used at any one time in a game, each club carries a complement of 25 players during normal periods of the season.

Traditionally all clubs were named for the cities in

which they operated. Beginning in the 1960s some have been named for states—the Minnesota Twins (Minneapolis–St. Paul), the California Angels (Anaheim–Los Angeles), and the Texas Rangers (Dallas–Fort Worth). All are members of one of the two major leagues, the National and the American. New York and Chicago have clubs in both leagues, each with its own stadium for play and severely independent of the other. Other professional leagues, known as minor leagues, operate in smaller cities and towns and serve almost exclusively as player-development "farms" for major-league clubs.

Each major-league club fields a team whose goal is to win the league championship. Each league is split into two divisions, officially "East" and "West." Each team plays a 162-game league schedule, equally balanced between games at home and away from home, and including a specified number of interdivisional games.

Commercial operations

Operation of a major-league club is a highly coordinated commercial enterprise involving extensive investment in: real estate and accommodations for spectators (although many clubs now play in multi-million-dollar, municipally owned stadiums erected for the use of the club); park service employees; the operation of a nationwide scouting staff for the search, signing, and training in the minor leagues for the approximately 200 players who make up the club's reservoir of replacements for retiring major-league players; and spring-training camps at which the players are maintained in warm climates during the weeks before the season opens. There is also a large administrative staff, usually headed by a general manager.

Revenues are derived from admissions and from the sale of television and radio rights, of rights to vend food and beverages at the games, and of programs, car parking, and fence and scoreboard advertising.

Cities enter the major leagues through purchase or transfer of a franchise operating in another city or through the expansion of the leagues. Both leagues increased from eight clubs to ten in 1961–62 and from ten to twelve in 1969. A franchise refers to a club's official certificate of membership in a league and establishes the right of its team to compete for the league championship. It is frequently owned by one person or by a small group, though some franchises are publicly owned stock corporations and others are subsidiaries of large industrial organizations (Columbia Broadcasting System, Anheuser-Busch).

Divisional components in each major league at the start of the 1972 season were American East: Baltimore Orioles, Boston Red Sox, Cleveland Indians, Detroit Tigers, Milwaukee Brewers, and New York Yankees; American West: California Angels, Chicago White Sox, Kansas City Royals, Minnesota Twins, Oakland Athletics, and Texas Rangers; National East: Chicago Cubs, Montreal Expos, New York Mets, Philadelphia Phillies, Pittsburgh Pirates, and St. Louis Cardinals; and National West: Atlanta Braves, Cincinnati Reds, Houston Astros, Los Angeles Dodgers, San Diego Padres, and San Francisco Giants.

At the season's close, divisional winners, determined by best percentage of games won, meet in a best-of-five-games series for the league championships. These decided, American and National League champions meet in a best-of-seven-games series for the championship of the world (World Series).

The great interest of Americans in baseball is reflected in the fact that an estimated 63,000,000 people were tuned in for the fourth game of the 1971 World Series, which was played at night. Gate receipts for the seven World Series games in 1971, with admission prices ranging to $12 per ticket per game, exceeded $3,000,000. Under a schedule adopted in 1969, each player of the team winning the World Series is guaranteed a minimum of $15,000 in prize money, players of the losing team $10,000. Players of teams eliminated in league divisional playoffs receive $5,000.

While each club controls and sells television and radio rights of games played on its home field, the commissioner is vested with authority to sell such rights on a national network basis for the World Series, All-Star Game, and one game a week, selected by a network, during the regular season. In 1969 this package was sold for a three-year period for a total of $50,000,000, with $5,300,000 earmarked annually for the players' pension fund.

The constant influx of new talent comes principally from schools, colleges, and independent amateur teams of city and rural districts throughout North and South America. Twice annually, in January and June, the leagues hold a joint free-agent draft in which clubs (in prescribed order) select for contract eligible young players who have never played professionally. Players selected must be signed to contracts before the next draft session. If not signed, they become subject to selection in the next draft. Bonuses to youngsters for signing may include such compensation as an $8,000 college scholarship and, to exceptionally talented prospects, up to $100,000 in cash.

Canadian, Latin American, and Japanese leagues

Organized professional baseball in the United States includes clubs in both Canada and Mexico. The game is widely played in Cuba, Puerto Rico, Venezuela, and Panama. A number of Latin American players became stars in the U.S. major leagues. Cuba figured prominently in winter league play, and Havana operated a club in the International League until ties with the Americans were cut after the successful Castro revolution. Two independent 12-club leagues now operate in Cuba, and admission to all games is free. Baseball is also a major sport in Japan, and crowds upward of 50,000 have attended games at Koshien Stadium in Osaka. Two professional six-club leagues, the Pacific and the Central, play 130-game seasons climaxed by a Japan Series between pennant winners. All league clubs are owned by industrial concerns, and the Tokyo Giants, operated by the Yomiuri Shimbun newspaper, won the Japan Series for the tenth time in 1970.

The system of mutual agreement among the operators of professional teams representing cities and towns in leagues throughout the United States and in parts of Canada recognizes one central supreme authority, the commissioner, and subscribes to the body of rules administered by this authority. Most important, these rules regulate the transfer of players from one league to another and among teams in the same league. The provisions aim at protecting the individual player from exploitation, at the same time imposing restraint on his freedom of choice to the extent that at no time after signing up with any club in organized baseball can he, without the consent of his employer, obtain employment from any other club in organized baseball until he has received an unconditional release from the contract under which he plays in any given year.

The two major leagues and each of their member clubs are signatories to an agreement, called the Major League Agreement, the first article of which creates the office of commissioner and sets forth its functions. The two major leagues, acting as one party, and the National Association, representing the minor leagues as the other, are signatories to the Major–Minor League Agreement, the first clause of which recognized the office of commissioner as created by the Major League Agreement.

As an aftermath of the Chicago White Sox scandal in 1919, the Commissioner's Office in 1921 took the place of the National Commission, a three-man board of arbitration established in 1903 and made up of the presidents of the two major leagues and an elected chairman. The framework of the commission continued under the Commissioner's Office in the form of the Advisory Council, the two league presidents with the commissioner as chairman. Nearly all the authority formerly vested in the commission was transferred to the commissioner.

Realignment and expansion. The year 1953 saw the first break in a half century in the alignment of cities in the majors when Lou Perini, president of the Braves, succeeded in transferring his National League franchise from Boston to Milwaukee. The last previous change in the lineup of major-league clubs had occurred in 1903, when the Baltimore Orioles became the New York Highlanders and then the New York Yankees.

Transferring franchises

The St. Louis Browns figured in another major realign-

ment in September 1953, when the American League franchise of the Browns was sold to a Baltimore syndicate. Beginning with the 1954 season the former Browns became the Baltimore Orioles. Another change in cities was effected before the start of the 1955 season when the Philadelphia Athletics' franchise was shifted to Kansas City. The New York Giants and the Brooklyn Dodgers played in their respective cities for the last time in 1957. The National League approved a change in locations prior to the 1958 season: the Giants to San Francisco and the Dodgers to Los Angeles.

In 1961 the Washington Senators moved their American League franchise to Minneapolis–St. Paul and were renamed the Minnesota Twins. A franchise for a new team was granted in Washington and another in Los Angeles, giving the American League ten teams. In 1962 the National League also expanded to ten teams with franchises in New York and Houston. In 1966 the Braves moved from Milwaukee to Atlanta, Georgia, and in 1968 the Athletics moved from Kansas City to Oakland, California. Both moves evoked threats of congressional action to avoid the game's exemption from federal antitrust laws. A suit to block the move from Milwaukee was rejected by the Wisconsin Supreme Court. The move from Kansas City two years later evaded legal repercussions by an American League agreement to include that city in an expansion to 12 clubs in 1969. Seattle, Washington, was later awarded the 12th franchise. National League expansion followed soon after with admission of San Diego, California, and Montreal, Quebec, the latter being the first major-league franchise outside the United States. Following the 1971 season the Washington franchise was transferred to Dallas–Fort Worth, leaving the nation's capital without major-league baseball for the first time in 71 years.

Adding new teams

These two waves of expansion, from eight to ten teams and then from ten to 12, generally resulted in immediate attendance gains. Equally significant, they enabled the owners to enlarge their television and radio markets. The cities that were awarded new franchises or were given old ones being moved either had built new stadiums that were standing empty and ready for occupancy or had pledged such construction. Not all moves were successful. The Seattle franchise, a notable failure, was in operation for only the 1969 season. The owners insisted they had suffered large financial losses and asked the American League for permission to sell or move. Approval was granted, but the city of Seattle and the state of Washington immediately countered with injunctions and restraining orders. The owners finally declared bankruptcy, and the franchise was sold and transferred to Milwaukee, which had been abandoned four years earlier when the Braves moved to Atlanta.

The cost of the expansion franchises continually spiralled upward. In the first round of expansion, for example, the four new teams were obligated to spend approximately $2,000,000 for 28 to 30 players of lesser skill who had been placed in a pool by the eight existing major league teams. These player purchases, at inflated prices, were, in effect, the cost of the franchise. In 1968 the National League's new Montreal and San Diego franchises each required entry fees totalling $10,000,000. Simultaneously, the American League's Kansas City and Seattle franchises went for approximately $6,000,000, a comparative bargain, but nonetheless three times the cost of the 1961 expansion franchises.

With the exception of the California Angels, who finished in third place in 1962, their second year, the expansion franchises were unable to field representative teams. That had been anticipated because the players made available to the new clubs were largely either untested rookies or aging veterans in the twilight of their careers. Still, the teams stayed afloat. Many of them played in new stadiums such as the Astrodome in Houston, which itself was an attraction. Most of the teams also had the advantage of opening new major-league territories. California, for example, which was without a major-league club prior to 1958, now had five teams—in Los Angeles, Anaheim, San Francisco, Oakland, and San Diego.

The New York Mets, born in 1962, became the first expansion team to win a pennant and a World Series. Upon arrival, the Mets discovered an eager audience. Many of their rooters were old New York Giant and Brooklyn Dodger fans, starved for baseball since the Giants and Dodgers had fled to California. The Mets hired Charles Dillon ("Casey") Stengel, then 70 years of age, as their manager. He had been forced into retirement, against his wishes, a year earlier by the New York Yankees. In 12 seasons with the Yankees, Stengel had directed the team to ten pennants, including an unprecedented five in succession. The Mets, under Stengel, lost their first nine games, but the most inept players were lionized and the fans came out, not to see the Mets win, but to watch them lose. The Mets obliged. They solidified their comic image by losing a record 120 games in their maiden season. In 1965 Stengel was again forced into retirement by ill health. Gil Hodges, a one-time Brooklyn star and also one of the 22 original Mets (he was acquired in the expansion draft), was named manager prior to the 1968 season. One year later, Hodges completed the "Miracle of the Mets," directing them to the National League pennant and to a World Series victory over Baltimore, the first time a team soared from ninth to first place in one season. The Mets led the majors in home attendance in 1969, drawing 2,175,373 paid admissions. They led again in 1970 with a home gate of 2,697,479, only 57,705 short of the all-time one-team record set by the 1962 Los Angeles Dodgers.

Rise of the New York Mets

The minor leagues. A radical revision of the minor-league structure was effected following the close of the 1962 season. Once self-sufficient but now heavily subsidized by major-league clubs, the minors were classified by population into leagues designated AAA (or Triple-A), AA (Double-A), A, and Rookie leagues. Earlier there had been seven classifications, including B, C, and D leagues. Rookie leagues are limited to players in their first or second season of professional play. Also, the majors conduct winter instructional leagues for prospective players. In the 1940s about 60 minor leagues operated in more than 400 cities in the U.S., Canada, Mexico, and Cuba and attracted nearly 42,000,000 paid admissions. By the early 1960s the number had declined to 20 leagues in about 130 cities with 10,000,000 admissions. Player contracts, once the most valuable operating asset of minor-league clubs, were by the early 1970s virtually all owned or controlled by parent major-league clubs.

The Triple-A leagues are the International, Pacific Coast, and the American Association, each with eight clubs. Other minor leagues include teams from cities whose geographic locations are identified by their titles, such as Eastern League (AA), Southern League (AA), California League (A), Florida State League (A), and Appalachian League (Rookie). There are also four leagues in Mexico: Mexican (AAA), Mexican Center (A), Mexican Southeast (A), and Mexican Northern (A).

The grading of leagues from top to bottom is part of the mechanism in organized baseball for assuring a player's opportunity for advancement. The grading represents different levels of salary limit. An annual draft empowers higher classification leagues to select players at fixed prices from leagues lower in the rating. Transfers of players' contracts are on a buy-and-sell basis, not only among teams in the same league but also in interleague transactions. There are fixed prices applying at the various levels for the drafting process.

Drafting of players

Each year before the arrival of the draft deadline in October, players of unusual ability may be transferred to higher leagues by the sale of their contracts at prices higher than the fixed draft price that would be obtained if a player's club chose to keep his contract beyond the date of the draft deadline, on the chance that he would be overlooked by scouts seeking reinforcements for teams in higher rated leagues. This trading in players' contracts evolved out of the baseball wars of the 19th century. In the early years of organized baseball, rival clubs commonly raided each other's ranks and hired away each other's stars in free competition on an open market. In 1879, Arthur H. Soden, of the Boston club, effected the

adoption of a resolution by the National League by which five players of each club might be named who could not be approached for hire by any other team in the league. This was the beginning of the Reserve Rule (also known as the Reserve Clause), amplified in 1883 when Col. A.G. Mills, National League president, brought about an agreement among the National League and the American Association, the two major groups at the time, and the minor leagues then in existence.

The Reserve Rule. The Reserve Rule, requiring the observance of each team's rights to the services of the players on its reserve list, is said to be the cornerstone of organized baseball. The original list of five reserved players was increased to 11 as part of the 1883 agreement. The number of players permitted to be held on each club's reserve list was increased in all gradations of leagues in the intervening years. The modern maximum is 40. In an effort to eliminate excessive bonuses for promising new players, in 1964 the league adopted a free-agent draft with all college and other amateur players drawn from a common pool. The teams draft, by rounds, in reverse order of their final standings for the season.

In the 1940s major-league players chose representatives from their own ranks to discuss problems and air any grievances with the club owners and the league executive council. Higher travel allowances, pension plans, and other benefits are among the items handled by the athletes' spokesmen. The ownership of players resulted in a series of court fights in the 1950s stemming from the baseball clause that bans a player from bargaining on his own with any club while he is under contract to another team. Charges that baseball violated the federal antitrust laws went as far as the U.S. Supreme Court, which on November 9, 1953, reaffirmed a 1922 ruling that baseball was not a business within the purview of those laws. The House of Representatives Judiciary Committee, after an exhaustive survey of the monopolistic aspects of the sport, had issued a report in 1952 in which it found no grounds on which to recommend that the game be brought under antitrust regulations.

Challenges of the Reserve Rule

The club owners were confronted with still another antitrust challenge on January 16, 1970, when a $4,100,000 suit contesting the Reserve Clause was filed in New York City's Federal Court on behalf of Curt Flood, a star outfielder with the St. Louis Cardinals and a veteran of three pennant-winning Cardinal teams. Named as defendants were the baseball commissioner, the two major-league presidents, and the 24 major-league clubs. In essence, Flood objected to the fact that the Cardinals, in trading him to the Philadelphia Phillies without his knowledge or approval, had violated his rights as a citizen. Flood announced he would not report to the Phillies and that he would sit out the 1970 season. He did.

Unlike other previous court actions, this suit had the unanimous support of the Major League Baseball Players Association; moreover, the association provided financial help and hired Arthur Goldberg, the former U.S. Supreme Court justice and ambassador to the United Nations, to represent Flood. Judge Irving Ben Cooper heard the case without a jury. League officials and owners, predictably, expressed the belief that the Reserve Rule was necessary for an orderly arrangement, though one of the newer owners, Ewing Kauffman of Kansas City, admitted the reserve system did somewhat prohibit salary negotiations and that he, for example, would be willing to pay Flood $100,000 or $125,000 a year to play with his team. Judge Cooper, in a 47-page opinion, upheld the defense and ruled that the antitrust laws do not apply to baseball, citing the previous Supreme Court decisions. He did recommend, however, that modifications in the reserve system should and could be negotiated between the Players Association and the club owners. Flood's attorneys appealed, and in 1972 the U.S. Supreme Court heard the case. It rejected Flood's suit, reaffirming the 1922 and 1953 findings exempting baseball from the antitrust laws and calling on the Congress to correct any inequities under the existing system. In the meantime, Flood had been signed by the Washington Senators for the 1971 season, with the understanding that he would not be sold or traded without his consent; but in mid-season he left the team and the country to live in Spain.

Player revolts and unionism. By 1970, the players were organized as never before. The Players Association had 941 dues-paying members, including about 850 players, plus managers, coaches, and trainers. Only one eligible player had refused membership, not because of disagreement with association policies but because he did not believe in joining organizations of any type. Association dues were $2 a day for each day a player was on the active roster during the season, with a maximum payment of $344. The players received a like amount from the association's licensing and merchandising program, which, in 1971, produced revenues in excess of $250,000.

The first player union was the National Brotherhood of Base Ball Players, organized in 1885 by Billy Voltz, a sports editor and minor-league manager. It was originally a benevolent and protective association of about 200 players who were assessed $5 a month to aid sick and needy members and to provide death benefits. At first a secret organization, the brotherhood came out in the open and became a genuine bargaining union in 1886, when the National League adopted a $2,000 maximum salary rule. The leader of the brotherhood throughout its existence was John Montgomery Ward, a pitcher–shortstop who had paid his way through law school with his baseball earnings. Disappointed with conditions and low salaries, the brotherhood went into the baseball business and formed an "outlaw" Players' League. It collapsed in 1890 after one season.

History of players' associations

The Protective Association of Professional Baseball Players was organized in 1900 and lasted during the two years of the National–American League war. When National League owners ignored the requests of the Protective Association, the association drew up a long list of National League players and induced all but one of them to join the new American League. Later, the National League made concessions to the union in exchange for the promise that union members would honour their contracts, but the American League offered more money than the National did, and the union could not prevent its members from jumping. When the American was recognized as a major league in 1903, the Protective Association quietly died.

Next came the Baseball Players' Fraternity, organized in 1912 by David Fultz, a former player and at that time a lawyer in New York City. The fraternity grew out of an incident in the spring of that year when Ty Cobb of the Tigers was suspended indefinitely for punching a fan who had provoked him. Practically all of the major-league players and most of those in the higher minor leagues were members. Fultz once called a strike because of a squabble involving a minor-league player, but the dispute was settled the day before the strike was to begin. The Players' Fraternity lasted until World War I, when the War Department issued a "work or fight" order. It was never revived.

Raymond Cannon, a Milwaukee lawyer who represented some of the banned Chicago White Sox players in 1920 and later was a congressman, twice tried to organize the players after World War I, both times without success. Robert Murphy, a Boston attorney, organized the American Baseball Guild in 1946, which, in retrospect, was a successful failure. Though Murphy's guild did not grow to fruition, the very thought of its existence roused the owners to the adoption of a number of reforms the following season, including a players' pension fund.

Murphy claimed guild members in 12 of the 16 major-league teams and twice went before the National Labor Relations Board charging unfair labour practices against the players by the Pittsburgh and Washington clubs. The board refused to assume jurisdiction. Murphy called for a strike of the Pittsburgh players on June 7 because the club would not negotiate with him. While a game with the New York Giants was being held up, the players voted on the strike and decided against it. Later, the Pennsylvania Labor Relations Board ordered an election among the Pittsburgh players to determine if they wanted

guild representation. Prior to this vote, it was acknowledged that about 90 percent of the Pittsburgh players were in support of the guild; but the vote was 15–3 against the guild, with ten players refusing to vote.

The owners, cognizant of player unrest, approved reforms for the 1947 season. A minimum major-league salary of $5,000 was adopted, the first such minimum in history; it was agreed that no salary cut could be made on a player demoted to the minors during the course of a season; moving expenses up to $500 were allowed for players sold or traded; $25 in weekly spring-training expenses was provided (still referred to by the players as "Murphy money"); and annual cuts in player salaries were limited to a maximum of 25 percent. In addition, plans were made for the establishment of a players' pension fund and for player representation on the Executive Council.

Role of the pension fund

The players then began electing representatives from each team to discuss problems and air grievances, not only with the Executive Council but also with the management of the individual clubs. The pension fund, however, was to be the solidifying factor. It started growing to significant proportions in 1950 when the owners, guided by Albert B. ("Happy") Chandler, then the commissioner, signed a six-year contract for the All-Star and World Series television and radio rights, with much of this money going into the pension fund. The television contract alone was for $6,000,000, or $1,000,000 a year.

In 1953 the owners and players entered disputes about the pension money, and the players—then led by Ralph Kiner, a home-run-hitting outfielder, and pitcher Allie Reynolds—hired a legal adviser, J. Norman Lewis. With Lewis' help, and also because of the new Taft–Hartley legislation, it was agreed that thereafter the pension plan would be administered by a joint committee of players and owners. It was also determined that 60 percent of the World Series and All-Star Game television and radio income as well as 60 percent of the gate receipts from the All-Star Game would go to the pension fund. That formula assured the players of larger pension money as the cost of television and radio rights soared.

Lewis remained as the players' attorney for six years. Frank Scott, a former travelling secretary of the New York Yankees and later a business agent for most of the star players, was then chosen by the players as their general administrator and representative. In December 1959, the players appointed a new legal adviser, Judge Robert C. Cannon of Milwaukee, the son of Raymond Cannon, who had been unsuccessful in organizing the previous generation of players. He established rapport with both players and owners and was often mentioned as a potential successor to Ford Frick, who had followed Chandler as commissioner and was approaching retirement. Some of the veteran players, however, particularly those closely involved with association affairs, began expressing dissatisfaction with Judge Cannon, maintaining that association gains were small, sometimes insignificant. Nonetheless, in 1965 and in early 1966, after it was determined the association would hire a full-time counsel and negotiator, Judge Cannon was the favoured candidate and he was offered a $50,000 contract. But the judge soon disqualified himself. He rejected, accepted, and then a second time balked, requesting certain fringe benefits not previously included.

The players, many of them now financially sophisticated, formed a committee to find a successor and sought the advice of the chairman of the economics department of the Wharton School of Finance and Commerce of the University of Pennsylvania. Marvin J. Miller, then the assistant to the president of the United Steelworkers of America, was recommended. He had a thorough background in labour relations and had been appointed to the National Labor–Management Panel by Pres. John F. Kennedy in 1963 and reappointed the following year by Pres. Lyndon B. Johnson. Miller was the choice of the players' nominating committee and immediately toured the 20 major-league spring-training camps, speaking at hastily arranged clubhouse meetings. The players ratified Miller's nomination by a vote of 489–136, and he assumed office as the executive director of the Major League Baseball Players Association on July 1, 1966. Richard M. Moss, like Miller previously associated with the Steelworkers' Union, was appointed the association's chief counsel in January 1967. The owners retained their own negotiator, John J. Gaherin of New York City, previously a management representative for the Publishers Association of New York City and a negotiator for trucking and bus companies and railroads. Under Miller, the association established a bona fide collective-bargaining relationship with the owners. Rights and benefits of the players were set forth contractually, and grievance procedures, with unprecedented impartial arbitration, were established.

Establishment of collective bargaining

The minimum salary was raised 100 percent, to $10,000 in 1968, the first genuine across-the-board minimum increase in 20 years, and increased again to $12,000 in 1970, and to $12,750 in 1971. First-class travel and a liberal meal allowance, $16 daily in 1970, were assured. The players' financial gains, in salaries and benefits, were estimated at more than $11,000,000 during the first five years of Miller's tenure. On February 3, 1969, an estimated 125 players, including most of the top stars, met in New York City. It was believed to be the biggest mass player meeting in history and was a show of strength in a dispute with the owners over pension grants. The players threatened to boycott spring training and twice rejected owner proposals, first by a vote of 491 to seven, and then 461 to six. The owners subsequently increased their offer, and a compromise was reached several days before spring training was to open. The median player salary in 1969 was $20,000. On March 31, 1972, the Players Association, by a vote of 633 to 10, with two abstentions, voted to go on strike, also over a pension dispute. It was baseball's first general players' strike and lasted 13 days. The last few exhibition games of spring training were eliminated, along with the first 10 days of the championship season. The strike was settled when the owners agreed to release $500,000 of an estimated $1,000,000 surplus in the pension fund, for the purpose of widening player benefits eroded by a cost of living rise; and with the agreement there would be no attempt to re-schedule the 86 regular season games cancelled by the strike. The players also agreed to an approximate 5 percent salary loss because of these cancellations. The result was a shortened and imbalanced schedule for the 1972 season, with some teams playing several more games than others.

The umpires also began organizing. At a secret meeting in Chicago on September 19, 1963, the Association of National League Umpires was formed. John Reynolds, a Chicago attorney, was hired as counsel and administrator and won immediate financial gains. The American League umpires joined five years later, merging into the Major League Umpires Association. Two umpires, William Valentine and Anthony Salerno, were fired by American League president Joe Cronin, who announced they were dismissed for incompetency. Valentine and Salerno insisted they were fired for their union activities. The National Labor Relations Board, by a four to one vote, agreed to a hearing, an unprecedented action interpreted by some observers as an indication that baseball was no longer immune from the antitrust laws. The board, on November 19, 1970, dismissed the Salerno–Valentine charges, ruling insufficient evidence.

COMPETITION

The schedule. The regular major-league season usually opens the second Monday in April and closes on the last Thursday in September or the first in October. Divisional play-offs (best three of five games) begin the next Saturday and the World Series (best four of seven) a Saturday later. Minor-league seasons are shorter, usually ending on Labor Day (first Monday in September) or the Sunday following. Rookie leagues operate in July and August only, whereas Mexican leagues hold postseason series. Some split their loop into two divisions, winners playing off for the league championship. Others hold a Shaughnessy play-off series. In this, the top four clubs

The baseball season

in the final standing play a knockout tourney for a trophy, nominally a "governors' cup."

During the winter months there are league operations in Puerto Rico, Venezuela, the Dominican Republic, and frequently in Panama. These clubs are staffed with players on loan from the majors and minors as well as with native players. A Caribbean Series in February, to which the winter leagues send their championship teams, climaxes the winter season.

Preceding the opening of the championship season, each club has a period of spring training that lasts for a month or more, during which time its men are conditioned for playing almost every day over a period of months (5½ in the majors). It is during this training season that new candidates are tried out. The manager and his coaches extend their efforts to select and train the strongest lineup possible from their squad. Practice games are played between teams picked from the squad, and there are usually 30 to 35 exhibition games played against teams representing other organized baseball cities.

In 1904 the major leagues fixed on 154 games. With expansion of the major leagues to ten teams in 1961 (American) and 1962 (National), the figure was revised to 162. When games end in ties, the statistics of such games go into the records of the individual players, although subsequent dates for play-offs of the deadlocked contests are arranged. Postponements because of inclement weather, especially late in the pennant campaigns, often cause a reduction in the allotted number of encounters for the clubs involved because travel schedules of the teams might prevent the opportunity for a play-off before the scheduled closing day of the season.

In common practice, the play-off of ties and postponed games takes the form of the doubleheader, in which two games are played on the same date, with a 20-minute period intervening. The popularity of this bargain billing with the public has resulted in doubleheaders being placed in the schedule in its original draft before the season's start, usually on holidays and Sundays.

The games are arranged to take in 25 weekends and the three nationally celebrated holidays: Memorial Day, Independence Day, and Labor Day.

Starting in the 1930s, additional big days appeared in league schedules in the form of night games. Experiments in playing games under artificial lighting were reported in 1880 (Nantasket Beach, Massachusetts), but, because of the large area necessary to be illuminated, it was not tried in league games until 1930.

Baseball games at night

On April 28, 1930 (Muskogee at Independence, Kansas, Western League), the first night game that counted toward a league pennant was played. Since to most patrons of professional baseball weekdays are workdays, the opportunity to see a game on a weekday evening instead of waiting until Saturday or Sunday came as a boon. Night baseball became popular.

By 1935 engineers had solved the problem of lighting the extensive areas of large stadiums, and in May of that year the first major-league night game was played in Cincinnati (versus Philadelphia). By the 1960s most teams were playing approximately half of their games at night, and in 1971 a World Series game was played at night for the first time. All but one of the major-league parks were equipped to play games at night. The Chicago Cubs of the National League continued to play all their home games at Wrigley Field during the daylight hours. In 1965 the first indoor games were played, in Houston's air-conditioned Astrodome stadium.

Records and statistics. Each league maintains a bureau in which are filed the statistics of all its contests. After each game the official scorer, appointed by the league, drafts a record of the game, with its detailed columns showing each player's hits, runs, errors, etc. At the end of each season, the official averages for the entire campaign are compiled. With expansion of the major leagues in 1961–62, all previous averages and statistics, based on a 154-game season, became obsolete, and comparison between old and new records became a matter of controversy.

The biggest controversy occurred in 1961 when Roger Maris of the New York Yankees hit a total of 61 home runs in one season, one more than Babe Ruth's historic 60 in 1927, the most cherished of all baseball records. When it became apparent that Maris might surpass Ruth's record, Commissioner Ford Frick ruled that Maris would have to break the record within the confines of the Yankees' first 154 games to replace Ruth in the record books. This would not give Maris the benefit of the eight additional games created by the expanded schedule. Maris' last two homers were hit in the extra eight games, the 60th in the Yankees' 158th game and the 61st in the 162nd. Ruth also had the advantage of one extra game because of the replay of a tie. Both the Maris and Ruth marks are listed in the record books. Other records have toppled because of the longer schedule but have not caused concern.

The two home run records

Also affecting the records, though in a more subtle manner, were the new artificial playing surfaces. When it was apparent that grass would not grow in the Astrodome, Houston's indoor stadium, a synthetic grass, called Astroturf, was introduced in 1966. Three years later, the Chicago White Sox installed a similar artificial turf infield, setting a precedent for an outdoor stadium. By 1972 artificial grass had been installed at seven major-league parks. The ersatz grass has, in the main, helped the hitters. A ground ball skips through these infields with greater speed. For modern records, see under SPORTING RECORD in the *Ready Reference and Index*.

The championships. All scheduled games must have a winner. Games ending in a tie or unplayed because of inclement weather are scheduled for replay as soon as possible before the season closes. Games that remain unplayed at the close of the season, because of travel limitations or other cause, are cancelled, unless a club would lose a chance to become the divisional winner as a consequence. In such rare instances, these unplayed games are rescheduled after the close of the season.

Arranged in order of games-won-and-lost percentage, from the highest to the lowest, the list of clubs in a major-league division on any given day is called the club standing. The team at the top of the list (in first place) at the close of the season becomes the divisional winner and the final standing becomes the record for the regular season. When two clubs are tied for first place, a single postseason game is played between the two to decide the divisional winner. Divisional winners then meet in a best-of-five-games series for the league championship. The first two games are played on the home field of one club and the remaining game or games on the home field of the other.

The play-off winners become the league champions for the year, each winning the right to fly from a flagpole in its home field a banner or pennant, combining the red, white and blue of the U.S. flag in a design of its own selection, with lettering and numerals proclaiming the team as champion of its league for that year. This pennant is customarily raised at the champion's opening game at home the ensuing season and flown from the masthead throughout the new campaign. The pennant winners in each major league then meet in the World Series, a postseason competition in which about half of the total cash receipts of the first four games forms a players' pool.

The World Series. In the fall of 1903 the pennant winners of the two leagues met in a postseason series, won by the Boston Red Sox of the American League from the Pittsburgh Pirates, National League champions, five games to three. This postseason series was no new idea. In seven successive years, starting with 1884, postseason games had been played between the pennant winners of the National League and the American Association.

For the 1903 World Series, as in the World Series of the 1880s, the teams arranged matters between themselves, scheduling games and dividing the proceeds by mutual agreement. In 1904 the Boston Red Sox, who had won the 1903 world championship, won the American League pennant and issued a challenge to the National pennant winners, the New York Giants. The Giants refused, on the ground that there should be formal rules and central

supervision of a contest in which the prestige of the rival leagues was involved. Boston claimed the 1904 world championship by default. New York president John T. Brush offered a code of rules that was adopted starting in 1905. These rules were amended at various times but the main features remained unchanged. From 1919–21 the number of victories required to win the series was raised to five, but it reverted to four in 1922.

Schedule of World Series games

The scheduling of games originally involved choosing by lot or by tossing a coin. The major leagues now alternate in opening the series. In the even-numbered years, the series opens in the park of the National League pennant winner. In the odd-numbered years, the American League champion plays its first World Series games at home. Instead of 3–3–1, the distribution of the games is 2–3–2. After the first two games in the series-opening park, the team shifts to the park of the rival league's champions for three games. If neither has won four games after five games have been played, the scene shifts back to the field on which its first game was played and remains there for as many more games as are necessary to decide the winner.

During World War II, starting with 1943, transportation emergencies changed this order. Again, the system became three games in the series-opening park, then the shift to the other team's field, with the difference from the original Brush plan that regardless of how many games should become necessary, there would be no return trip to the first city. After the war the 2–3–2 system was again adopted. For World Series results, see under SPORTING RECORD in the *Ready Reference and Index*.

The All-Star Game. Starting with 1933, the major leagues have played an annual All-Star Game in July, one game between teams of players chosen from all the teams of both leagues. They oppose each other as league against league. No championship is involved; the players do not receive any part of the gate receipts, which are donated to the players' pension fund and to charitable purposes. The first All-Star Game, held in conjunction with the Century of Progress Exposition in Chicago, was played in Comiskey Park, home field of the Chicago White Sox (American League). In subsequent years the leagues have alternated as "home" teams. From 1959 to 1962 the major leagues played two All-Star Games each season.

The Hall of Fame. A National Baseball Hall of Fame and Museum is at Cooperstown, New York, where Abner Doubleday supposedly laid out the first baseball field. It contains relics, pictures, and documents, and its central chamber is the Baseball Hall of Fame, in which players and personalities who have made major contributions to baseball are memorialized. The first five players selected for membership in the Hall of Fame in 1936 were Ty Cobb, Babe Ruth, Walter Johnson, John P. ("Honus") Wagner, and Christy Mathewson. Additions to the roll of immortals are made from time to time by election in polls conducted by the Baseball Writers' Association of America and by the Hall of Fame Committee appointed by the commissioner of baseball. A special committee was formed in 1971 for the purpose of selecting one player each year from the Negro leagues, now extinct. The first Negro player to win election from this committee was Leroy ("Satchel") Paige, who pitched in the major leagues for five years when he was in the twilight of his career. For members of the Hall of Fame, see under SPORTING RECORD in the *Ready Reference and Index*.

AMATEUR BASEBALL

According to legend, in the United States almost every boy plays baseball or its variant, softball, during his school years. In the lower grades it is often played by both boys and girls during school hours, in the brief periods set aside for outdoor recreation. Records of competition between high schools date to 1886. Though the popularity of baseball in the rural and town areas has declined, it has grown at the high school level, from 150,000 participants in 1947 to more than 400,000 in the 1970s, with 13,000 high schools having at least one baseball team. In the warm-weather states, such as Arizona, Florida, and California, some of the high schools play a 50-game schedule, opening their season in January and finishing in June.

The largest organization offering a continuous amateur program without regard to age limit is the American Amateur Baseball Congress (AABC), founded in 1935. Originally designed for adults (unlimited age division), this program has grown to five divisions: Stan Musial Division (19 and older); Connie Mack (18 and under); Mickey Mantle (16 and under); Sandy Koufax (14 and under); and Pee Wee Reese (12 and under). By 1970, the AABC had approximately 3,000 sanctioned teams with an estimated 60,000 participants. National champions are crowned in each age division.

Organized ball for boys. The Little League, originally for the eight to 12 age group, is the best known of the boys' baseball programs. It started in 1939 with three teams in Williamsport, Pennsylvania, and had a phenomenal growth. By 1970 in the United States alone there were more than 60,000 Little League teams and 1,000,000 participants at the age 12 and under level. Subsequently, two older age groups were added—senior division for boys 13–15, and big leagues 16–18.

Little and junior leagues

Little League, which is now international, was granted a federal charter in 1964 by act of the U.S. Congress. The Little-Bigger League, for boys 13–15, was organized on a national basis in 1952 and at the end of the 1953 season changed its name to the Babe Ruth League. Competition supervised by the Police Athletic leagues in various cities and Boys' Baseball (formerly, PONY [Protect Our Nation's Youth] League) has been arranged for promotion of the game among boys.

The first national baseball program for boys was the American Legion Junior League, which was formed in 1926 for boys up to 17 years of age. In 1960 it was changed to include 18 year olds as well. There are American Legion teams in all 50 states and the Panama Canal Zone and Puerto Rico. In 1971 there were 3,200 teams. More than 60 percent of all the major league players participated in the Legion program as boys. When the Baltimore Orioles met the Pittsburgh Pirates in the 1971 World Series, 30 of the 50 eligible players for the series were graduates of the American Legion program.

College and international competition. Baseball is also widely played in colleges and universities. In 1967, the last time the National Collegiate Athletic Association (NCAA) surveyed its members, it was learned that more than 500 colleges fielded varsity teams, involving some 17,000 players. The NCAA, in 1963, provided the thrust in the formation of the United States Baseball Federation, which, in effect, is a collection of all the amateur groups. The federation sponsors clinics, is vigorous in promoting baseball, and, equally important, gives the United States a link with the International Federation, making American teams eligible for international competition such as the Pan-American Games and the World Amateur Tournament. The U.S. Federation usually selects all-star teams for international play, choosing the best players from the major and junior colleges and from the armed forces; in 1972 the federation announced an agreement with the Japanese Amateur Baseball Association to conduct a Collegiate World Series between the United States and Japan, beginning in July 1972.

Baseball tournaments are played on U.S. Army posts the world over. The game is a minor sport in many European countries (*e.g.*, The Netherlands, Italy, Belgium, England, Spain), as well as in Australia and Tunisia.

(Je.Ho.)

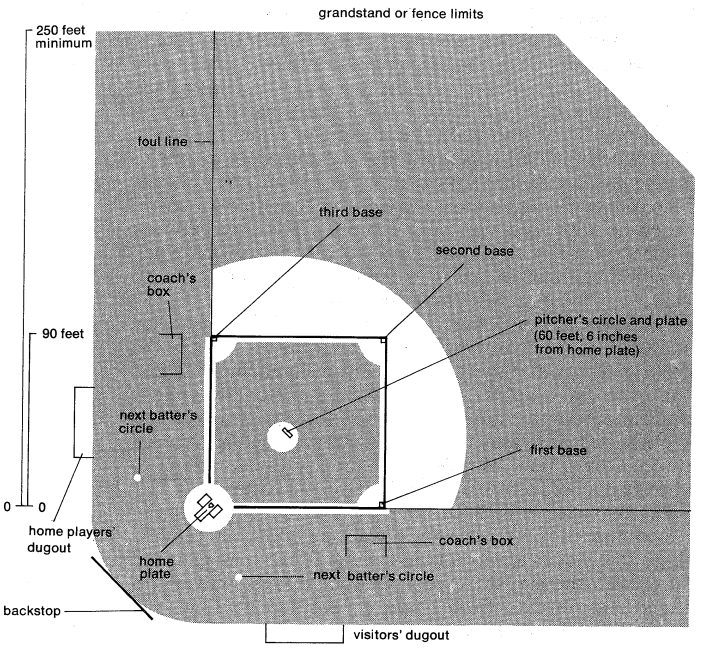
II. Play of the game

GROUNDS AND EQUIPMENT

The field of play. The baseball field, usually called the diamond, is a large, level area, most of which is covered with natural or artificial grass. Though the field may cover about two acres (one hectare), the actual dimensions vary from field to field.

The layout of the field is based on the layout of the bases, which form a square, also known as the diamond or infield, with the home base, or home plate, as the

point of orientation (see diagram). The sides of the square, and thus the distance between the bases which form its corners, are 90 feet (27.4 metres). White lines marked on the ground along two sides of the square



Layout of a representative baseball field.

The baseball diamond

(from home to first base and home to third) extend beyond first and third to the nearest fence, stand, or other obstruction. These are foul lines marking the left and right limits of the field of play; the outer limits are formed by a fence or the walls of stands. The distance along the foul lines from home plate to the limits of the field must be 250 feet (75 metres) or more; in ball fields built after 1958, it must be at least 325 feet (100 metres) and the distance to the centre-field fence at least 400 feet (120 metres). The entire area between the foul lines and extending roughly from the square to the outer limits of the field is known as the outfield.

The playing field is covered with natural or artificial turf except for a circular area around the pitcher's plate (see below), which is bare; traditionally, the paths between the bases and an area extending beyond the base lines from first to third (the size and shape of the area being determined by each club for its own ball field) also are left bare, as is the area around home plate. With the introduction of artificial turf, the tendency has been to cover the entire field except the areas around the pitcher's plate, home plate, and immediately around the bases. Although the infield is defined as the square delineated by the four bases, in common usage it also includes the area beyond the second- and third-base lines. There is no actual dividing line between the infield and the outfield.

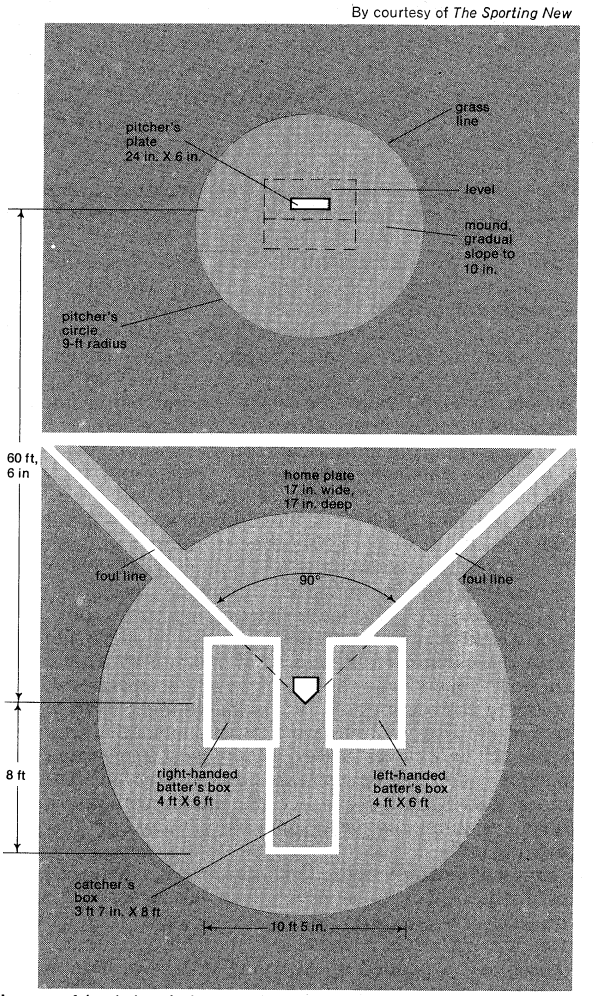
First, second, and third bases are marked by white canvas bags filled with a soft material (such as kapok) and securely attached to the ground by means of a metal peg or stake. Home base or, more generally, home plate, is a pentagonal white slab of rubber embedded in the ground at the intersection of the first- and third-base lines, its front edge facing the pitcher.

In the middle of the infield is the pitcher's mound, a circular area kept free of grass and rising in a gentle slope only slightly above the level of the base lines. On its summit, which is flat, the pitcher's plate, a slab of rubber that the pitcher must step on when delivering the ball, is embedded in the ground 60 feet six inches (18.5 metres) from home plate. The pitcher's mound and plate are sometimes referred to collectively by the term pitcher's box, a holdover from the 19th century when an oblong area was marked out on the ground for the pitcher to occupy in the act of throwing the ball.

Behind home plate is the catcher's box, the rectangular area within which the catcher must remain until the pitcher delivers the ball. On either side of the plate are marked batter's boxes, one for right-hand and one for left-hand hitters. The official rules recommend that the distance behind home base and the backstop screen and from the first and third base lines to the grandstand or fence limits be at least 60 feet (18 metres). Within this area, known as foul territory, and 15 feet (5 metres) off of first and third base are the coaches' boxes, rectangles, for the two men for each team assigned to guide base runners and communicate signals. Also in this territory and about 40 feet (12 metres) from each of the batter's boxes are two "on deck" circles—areas in which the next person to bat waits his turn.

The catcher's, batter's, and coaches' boxes

By courtesy of *The Sporting New*



Layout of (top) the pitcher's circle and plate and (bottom) home plate.

Uniforms. The uniform consists of a cap, blouse, pants, stockings, and shoes. Traditionally, the home team wore white blouses and pants, the visiting team gray; in the 1960s and 1970s, however, some teams began to wear coloured uniforms. An identification number is worn by each player.

The blouse is collarless and without sleeves or with short sleeves usually ending halfway from shoulder and elbow. The pants cover the player's knee, the elastic at the bottom pinching in just below the knee joint. Most players wear thigh pads under the pants and suspended from the waist for protection from abrasions when sliding along the ground into a base and in unintentional tumbles.

Players wear white cotton stockings underneath coloured stockings.

The shoes are oxford-style, light in weight, with flexible soles and heels. Each shoe has three prongs (spikes) on the sole and three more on the heel. The spikes, whether attached to the shoe individually or to triangular metal

spike plates, give the player a good grip on both the turf and the grassless dirt surface of the base lines. The pitcher also wears on his pitching foot (right foot for a right-handed pitcher; left foot for a left-handed pitcher) an additional pitcher's plate, a metal reinforcement at the inside of the toe to give him a better foothold in the act of pivoting to deliver the ball. Umpires wear shoes similar to those of the players. On artificial turf, some players wear a shoe that has up to three dozen small cleats, usually made of a synthetic and similar to the type worn by football players on synthetic grass.

The bat and ball. The bat is a smooth, rounded stick, made either of a single piece of hardwood or of laminated construction; the maximum length allowed is 42 inches (107 centimetres) and the maximum thickness is 2¾ inches (7 centimetres), decreasing in diameter at its handle end. There is no restriction of the bat's weight, except that no metal or any other reinforcement may be used in its construction. The gripping surface may have a covering of tape or other abrasive agent to give a firm hold for the batter's hands.

The ball is restricted to a weight of between five and 5¼ ounces (142 and 149 grams) and a circumference of from nine to 9¼ inches (23 to 23½ centimetres). It has a cork and rubber core, is wound tightly with woolen yarn, and is covered with two pieces of white leather shaped to fit together tightly when the two pieces are sewn together.

Gloves. In preleague baseball, the game was played bare-handed. One by one, starting with the catcher and first baseman, players began to wear gloves to protect whichever hand bore the brunt of the ball's impact. Every player now wears a glove when in the field.

Types of gloves

The gloves are leather, with some reinforcement (padding). The catcher's glove is the largest and heaviest, thickly padded at all points except the middle of the palm, in which the pitched ball embeds itself. It is all one piece, except for a cleft between the thumb and the index finger. The size of this glove is limited to 38 inches (97 centimetres) in circumference and cannot be more than 15½ inches (39 centimetres) from top to bottom. The first-baseman's glove, thinner and much more flexible than the catching mitt, is a solid expanse of leather for the four fingers. The thumb, however, is separate, with a webbing connecting it with the index-finger edge of the rest of the glove. Gloves worn by pitchers, infielders, and outfielders are finger gloves, usually with a separate compartment for each finger and for the thumb. Two or more leather straps connect the thumb with the index finger, enabling the player to get a tight hold on a swiftly moving ball.

Protective equipment. Except for the catcher, the player in the field is considered adequately armoured when he wears his regulation uniform and glove. The catcher wears a mask, chest protector, and shin guards as well. The mask, a padded metal frame with solid bars across its open front, fits the front half of his head, so that the catcher has full visibility but is protected from being hit in the face or about the head, ears, and throat by the ball. The chest protector is a solid, padded framework extending up over both shoulders and down between the legs, folding at about the waistline to allow stooping. The catcher's shin guards are of light metal over the shins, with padded leather extensions that fit over the knees, and at the bottom are shaped to the shoes so that a catcher blocking home plate has reasonable protection from the spikes of a player sliding feet first.

Protection for the catcher

At bat, a player wears a lightweight batting helmet in addition to his regulation uniform.

The umpire, likewise, wears mask, chest protector, and shin guards when on duty behind home plate. The mask is similar to the catcher's.

CONDUCT OF THE GAME

The umpires. Play is under supervision and control of one or more umpires, acting as judges and announcing decisions on whether pitched balls pass over the plate within or outside the strike zone, whether batted balls are fair or foul, and whether a base runner has been put out by a throw or by tagging or is safe (*i.e.*, entitled to hold the base he has reached by running).

Four umpires are on duty in major-league games, one stationed behind the catcher and the other three covering the plays that occur at first, second, and third bases. The home-plate umpire starts the game by calling play for the first inning. The ball goes into play when the pitcher pitches it to the catcher. It remains in play except when an umpire calls "time," which is done on all foul balls except those that are flies (*i.e.*, caught before touching the ground) and also upon the request of either team at a time when the ball is being held by a player of the fielding team and no base runner is attempting to advance. The request for time most frequently is made by a batter when he wishes to step out of the batter's box between pitches. Time is also called on the rare occasions when a batted ball strikes a base runner, and in other circumstances as the umpire's discretion dictates.

Managers and coaches. In professional baseball, the manager is the man who runs the team. He selects the players who will play in each position in each game and the order in which each will bat; he formulates plays; and he determines offensive and defensive strategy and tactics, and how and when they will be used or altered during the game. He decides, for example, when to change pitchers or use a pinch (substitute) batter.

The manager is assisted by a staff of coaches, usually accomplished players who have passed their playing days; they act as all-around assistants to the manager and work with the players, especially the younger players, to help them develop their pitching, batting, fielding, and other skills of the game.

Coaching functions

Another coaching function, which may be performed not only by a member of the coaching staff but also by any other member of the team, including the manager, is to give instructions (by voice or by means of signals) to base runners and batters of his team. When a team is at bat, one coach may stand within the coach's box off first base and one within the coach's box off third. The coaches, in turn, receive instructions from the manager, usually through signals, as to what strategy should be employed in a given situation.

PRINCIPLES OF PLAY

The bases are called first, second, and third base and home base or plate. All batting is done at home plate. Base running proceeds from home plate along the base lines (the boundaries of the square). A run is scored when a player, having successively touched first, second, and third, reaches home plate.

Seven of the nine players on the fielding team take their stations, prepared to capture (field) the ball as soon as possible after it is hit by the batter. The other two, the pitcher and the catcher, form the battery. The catcher stands behind and within stepping distance of home plate but not close enough to impede the batter's activities.

The pitcher stands near the centre of the infield. He puts the ball in play by throwing it to the catcher. The batter, standing on one side of home plate, gets a chance to hit the ball as it passes him on its way to the catcher. He stands facing the plate, swinging the bat with both hands, and putting the full force of his shoulders and arms into his swing at the ball.

The pitcher is the central figure in the game, which is a succession of "plays." Each play begins when the pitcher delivers the ball to the batter and lasts until the pitcher holds the ball again, standing with his foot against the pitcher's plate, or rubber, ready to toss the next pitch.

If the ball travels past the batter to the catcher and is then tossed directly back to the pitcher, it is a complete play in itself. If the ball is batted, the play goes on until the ball has been fielded, all base running has come to a stop, and the ball returned to the pitcher.

Players are termed right-handed batters or left-handed batters, according to whether they swing the bat from their right side or their left.

When the batter hits the ball inside fair territory (the area of the playing field between the foul lines), he is entitled to start running round the bases. Since it is only

a fair ball that entitles the batter to run, the fielding team's defensive arrangement places its seven movable men within the fair territory. Four of them (the infielders) patrol the ground between first and second and between second and third. The other three (the outfielders) play the outfield, the wide sweep of the field from foul line to foul line beyond the two base lines.

The fielders' playing positions

The outfielders are called right fielder, centre fielder, and left fielder with relation to a man standing on home plate and facing toward second base, but there are no visible boundaries separating the three fields. With the exception of the pitcher and catcher (and in their case only at the moment of putting the ball in play), there is no restriction on the movements of the members of the fielding team. The four infielders are named first baseman, second baseman, shortstop, and third baseman, stationed in that order from first base around the rim of the infield to third. Their fields of operation are nearly as flexible as those of the outfielders. The arrangement of the fielding team, at any given point in a game, is changed to meet the needs of the existing situation.

Offense. The objective of the offense is to score runs. The main offensive force at work in a baseball game, at any given point, is the batting team's ability to hit fair balls out of the reach of the defending fielders. Along with this batting goes the running speed of the batting team in its efforts to traverse the bases. Each team strives to advance its players around the bases to score as many runs as possible before the third out ends its half of the inning at bat. Each inning is a new game in the sense that it starts with the bases empty. The first man in the batting order is the first batter in his team's first inning. In subsequent innings, the first batter up is the man in the batting order after the last batter in the previous inning to complete his turn at bat.

Hitting the ball. The acme of successful batting is to drive a pitched ball inside the foul lines and beyond the confines of the playing field (usually into the stands among the spectators or completely out of the park). A ball so driven is called a home run. It has passed beyond the reach of any fielder and entitles the batter to run at any speed around the bases to score a run at home plate. Any and all runners who are on base when the home run is hit likewise make their way to the plate, in the order in which they reached base, and register a run each.

Hits and errors

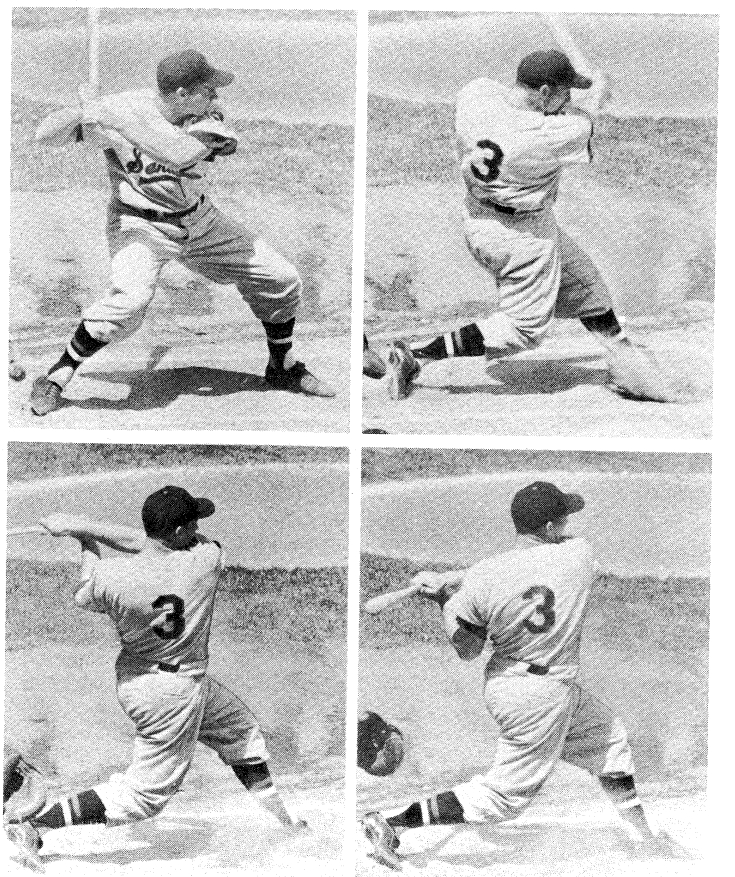
Failing such a hit, the batter aims to drive the ball so that it cannot be caught either before touching the ground in fair territory or soon enough after touching ground to be thrown to first or any other base before the batter or any other runner gets there. If he succeeds in getting on base before the ball gets there, he has made a hit. If he reaches first base and no farther, it is called a one-base hit. If he drives it far enough so that he can reach second base safely, it is a two-base hit. A hit long enough for the batter to reach third base is a three-base hit. A similar hit by a batter that enables him to touch all bases and score a run is known as an "inside-the-park" home run—a rarity.

With the ball thus hit by the batter into fair territory, any members of his team who are already occupying bases advance as far as possible toward home plate before the ball is returned to the infield. They must reach their objectives without being tagged out; thus, their judgment of how far they can go when the ball is hit safely is a major element in successful attack.

A batter may hit the ball inside fair ground and reach base safely because of a fielding lapse (*i.e.*, failure to catch the ball or inaccurate throwing). This is recorded as an error committed by the fielder. Failure to tag a runner with the ball when the fielder has a chance to make a putout in this way also counts as an error.

Base on balls. A batter may also reach first base without having hit the ball. This is nearly always the result of inaccuracy by the pitcher in his efforts to make the batter hit a bad ball (*i.e.*, a ball passing the batter at a point at which the batter cannot hit the ball with his full power). Whenever the batter does not swing at a pitched ball, and the ball does not cross the plate inside the strike zone, the umpire standing behind the catcher announces, "one ball." If four balls are thus declared in a single turn at bat—that is, before the pitcher has thrown three strikes—the batter is entitled to go to first base without a play being made on him, the recipient of a base on balls, also known as a walk. A game without a base on balls for any batter of either side is unusual.

Getting on base without a hit



Stages in the powerful swing of Harmon Killebrew, a leading home-run hitter.
Wide World Photos

Hit by pitched ball. The batter reaches first base, without hitting the ball, if a pitched ball at which he does not swing strikes any part of his person. The umpire then awards him the right to go to first base.

Interference. The batter also reaches first base if the catcher interferes with him by touching any part of his body or interferes with the swing of his bat as the pitched ball is on its way to home plate. The catcher in such a case is charged with an error.

When a batter is awarded a base on balls, for being hit by a pitched ball or for interference by the catcher, teammates on consecutive bases from first on advance one base; if all three bases are occupied (loaded), all advance one base, thus scoring a run for the batter's side.

Missed third strike. If, with two out or with first base unoccupied regardless of how many are out, the batter swings and misses the ball for his third strike and the catcher does not catch the pitched ball before it touches the ground, the batter is entitled to run for first just as if he had hit the ball along the ground in fair territory. The batter is also entitled to run to first base if the catcher misses a third strike that the batter does not swing at. The catcher must then get the ball and throw to first ahead of the batter in order to put him out. If such a pitched ball rebounds off the catcher out into the infield, the pitcher or any infielder may make the pickup and throw to first, just as if it were an infield grounder. This right to try for first base does not exist when first is occupied and there are fewer than two out because, if it did, the catcher might drop the ball directly in front of him and thereby create the opportunity for a double play, throwing to second to force out the man on first.

followed by a throw from second to first before the batter arrived at first. Missed or dropped third strikes are uncommon in the professional leagues.

The infield fly rule. Similar protection is afforded the batting side by the infield fly rule. This rule applies only if both first and second or first, second, and third bases are occupied, there are fewer than two out, and the batter hits a high fly that in the judgment of the umpire can readily be caught by an infielder or the pitcher or catcher inside fair territory. In such circumstances, the umpire immediately declares the batter out, whether or not the ball is actually caught. This law was enacted in 1895 because clever infield players would pretend to catch the fly but at the last moment would let it strike the ground, thus setting up a force play (see below) at both third base and second base, which could be made a double play by quick and accurate throwing. An attempt to bunt (a light tap of the ball rather than a full swing at it; see below) under the conditions noted above, which results in a fair fly, is not regarded as an infield fly.

Stolen bases. The base runner can wait on a base until a teammate drives the ball out of reach of fielders, enabling him to advance toward home plate to score a run. The rules do not, however, compel him to wait for the ball to be batted. He may advance on the bases at any time the ball is in play, his only restraint being the threat of being tagged out; that is, of being touched with the ball in the hand of a member of the fielding team when he is not on a base. Obviously, a dash for the next base when the ball is being held by the pitcher, the catcher, or an infielder makes the runner almost a certain out. When the pitcher delivers the ball to the batter, however, the runner can match his speed with the strength and accuracy of the catcher's arm. If the runner makes such an attempt and succeeds in reaching the next base without being tagged out, he is said to have stolen a base.

Stolen-base records

Major league stolen-base totals rose significantly in the 1950s and 1960s; there were 650 successful steals in 1950, 923 in 1960, and 1,908 in 1970, when 12 players, six in each of the two major leagues, stole 30 or more bases. Maury Wills, who was instrumental in the return of the stolen base, in 1962 set a modern (20th-century) major league record by stealing 104 bases, bettering the record of 96 set by the famous Ty Cobb in 1915.

The hit-and-run. Many attempts at stealing bases in modern baseball take the form of the hit-and-run play, in which the batter cooperates with the runner. The setup almost always calls for a runner on first, but on no other base. The runner starts for second as the pitcher begins his pitch. If the second baseman shifts to be in position to receive a throw from the catcher, the batter tries to drive the ball along the ground through the area ordinarily guarded by the second baseman. If it is the shortstop who has thus covered second, the batter tries to drive the ball through the shortstop's vacated sector.

If the batter succeeds in his effort, the ball goes to the outfield as a hit and the runner reaches third base easily. There are lesser degrees of success. The batter may hit the ball along the ground but not through the open sector. This usually results in an out at first base on the batter, with the runner reaching second safely, a moderately advantageous outcome because it has placed the runner on second, in position to score on an ordinary one-base hit to the outfield by a subsequent batter, and has removed the double-play menace that always exists when a runner is on first base with fewer than two out.

If the batter hits a fly, the runner must quickly retrace his path and regain first base; this is usually done easily. If the batter misses the ball entirely, the most he has accomplished is to offer a visual hazard to the catcher's throw to second base. The play in this instance takes the pattern of a plain attempt to steal second. In every major league season a large percentage of the stolen bases credited to players are the result of hit-and-run plays in which the batter fails to hit the ball.

The bunt. Nearly every time a batter tries to hit a ball he takes a full swing, aiming to drive it as fast and as far as he can. There are times, however, when he tries to bunt the ball—that is, tries to tap it lightly with the bat—to make it roll slowly along the ground in fair territory but off to one side or the other of a straight line from catcher to pitcher. In bunting, the batter usually relaxes his grip on the bat and, instead of swinging, merely holds out his bat so that the ball strikes it and drops to the ground with only enough force to send it out at a point in the infield. The batter tries to conceal his intent as long as possible so that neither the pitcher and catcher nor the infielders can get the jump on the play. The bunt is usually intended to be a sacrifice; that is, the batter expects to be thrown out at first base. His purpose is to enable one or more runners to proceed to their next base while the play is being made on his bunt, retiring him at first base. He thus sacrifices himself in the interests of advancing a potential run or two.

The squeeze play. The bunt is also used to sacrifice a runner home from third base, but the technique is somewhat different. On a prearranged signal, as the pitcher starts delivering his pitch, the man on third starts running toward home plate. This is an all-or-nothing play and was the original squeeze play. It is now known as the suicide squeeze because, if the batter misses entirely, the catcher, by catching the ball, has the runner trapped between third and home, a certain out, unless the runner resorts to dodging back and forth and some error of throwing or catching allows him to escape. If the batter does bunt the ball into fair territory, it means a certain run scoring for his side because of the runner's flying start. At best, the defensive side can only throw out the batter at first base.

Bunting to bring in a run

The suicide squeeze is the most spectacular and dangerous sacrifice bunt and was the forerunner of the safety squeeze, a similar manoeuvre but without as much risk. On the safety squeeze the runner at third holds until the hitter has bunted the ball on the ground. This delay makes it easier for the defense to adjust, but the play will succeed with a fast runner at third and a better than average bunter capable of pushing the ball about 20 feet (six metres) down either the first- or third-base line, but not much farther. If the bunt is shorter, the catcher can run it down and throw the ball to the pitcher who has rushed to cover the plate.

The batting order. Each of the nine players on a team must take his turn at bat. The strongest hitters are grouped conventionally in a certain order that has no relation to the positions they may play in the field. The first two positions, leadoff and number two, are assigned usually to players with keen eyes and brains, those fast afoot, and usually not so powerful in driving the ball for distance as batter number three and the cleanup man, number four.

The best leadoff is one who can judge whether the pitch will be a strike or a ball if he lets it pass him without swinging. By waiting out the pitcher, letting the bad ones pass, he draws many a base on balls. In his strategic position, first batter up in the first inning, and following the weak-hitting pitcher in the other innings in which he bats, a base on balls for him is usually just as advantageous to his team as making a one-base hit. Getting on base is the leadoff's main offensive function. Over a stretch of games his run-driving opportunities are relatively few.

For batter number two the manager selects his best hit-and-run man, the batter who is most skilled in driving the ball toward the right side of the field. When there is a runner on first, the first baseman must stay on or close to first, thus opening a wider sector through which the ball may safely be driven along the ground into right field. The ability to hit the ball in this direction is important, especially on a hit-and-run play with the runner rushing for second as the pitcher pitches. The number two man must also be a capable bunter.

Number three is usually the best all-around offensive player on the team, having running speed combined with batting power and skill. The greatest hitters of all time have been, in the main, number three in their team's batting order—Ty Cobb, Babe Ruth, Rogers Hornsby, Eddie Collins, George Sisler, Joe DiMaggio, Ted Williams, Stan Musial, Willie Mays, and Roberto Clemente.

The number-three hitter

Not all of those players hit third every season; sometimes they batted fourth, usually when their teams were without another outstanding home-run hitter.

Numbers four and five are the long-distance hitters, not likely to hit the ball safely as often as number three, but often with greater distance. The remaining positions in the batting order scale downward, and it is there the highly skilled defensive stars, players whose value to the team is their mastery of their fielding position, may be found. Number nine is almost invariably the pitcher, the most specialized of all players, whose hitting ability is usually negligible. Because a pitcher is called on for duty only about once in four or five days, even pitchers who can hit well do not have the chance to remain tuned up. Their batting ability is rarely considered; in every other position a man's worth always represents his combined batting and fielding skills.

In the positions requiring less skills, fielding ability is a less important factor. Batting ability overshadows fielding considerations in positions such as right field or left field, third base, and even first base; in the other positions the manager will weigh very carefully the fielding merits of the aspirants in making his selections. He usually wants for his shortstop the best possible fielder he can find, regardless of batting, and whatever such a star fielder can furnish in hitting power is sheer profit. His left and right fielders and first and third basemen, however, must be hitters, and their ability to run faster, to judge a fly ball, or to throw a ball straighter is secondary.

Substitutions. Substitutions may be made at any point in the game when time has been called by the umpire. A player taken out of the lineup cannot return in the same game. When the manager makes two or more substitutions at one time, he must specify them one at a time to the umpire, so that each substitute's place in the batting order is immediately established. Such substitutions may be made regardless of the positions played by the players involved. For instance, the manager may take out of his lineup his pitcher and second baseman, substituting a second baseman to bat in the pitcher's place in the batting order, the new pitcher batting where the second baseman formerly batted. He may send in as many substitutions at one time as he wishes, up to the limit of his nine names, but each substitute takes a fixed position in the batting order at the moment the change is made. Without making any substitution, the manager may at any time in the game shift his players from one fielding position to another. He may shift all nine positions in fielding, but he cannot change a man from one place to another in his batting order.

The use of a substitute as an offensive tactic most commonly involves sending in a pinch hitter; that is, taking a weaker hitter out of the lineup and substituting another player whose likelihood for driving the ball for a hit or a fly to the deep outfield is greater than that of the player next up in the batting order. Such a pinch hitter must be a player not already in the lineup nor in the batting order at any previous time in the game.

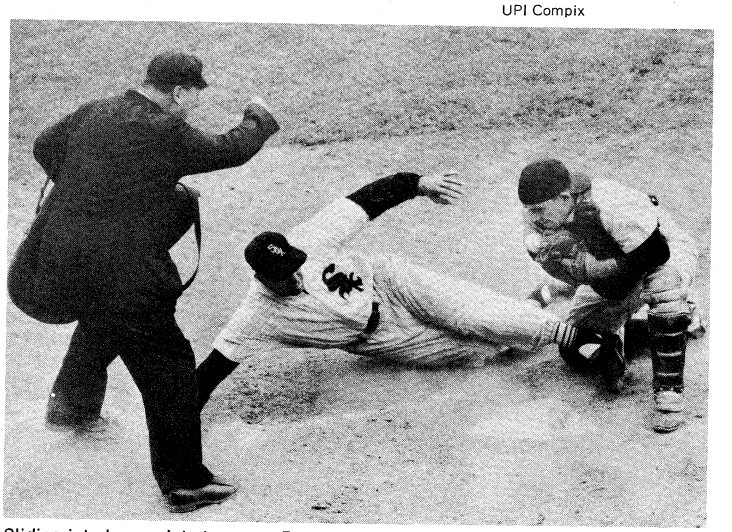
In 1973 the American League adopted experimentally the designated pinch hitter rule. The rule allows a pinch hitter, designated before the game, to bat for the pitcher without forcing the pitcher from the game. The pitcher may take his normal turn at bat, however, in which case the designated pinch hitter is taken out of the lineup and regular rules prevail. During a game the designated pinch hitter can replace a defensive man on the field. If this happens he continues to bat in the pitcher's spot, and the pitcher then bats in the spot vacated by the defensive player forced from the game.

Defense. To meet the offensive force of the team at bat, the rules provide the fielding team with ways of making putouts. A putout removes the player from further offensive play until his next turn at bat. The batting team's inning continues until three putouts are made; then the team goes into the field and the rival team comes in for its turn at bat.

Putouts. Most putouts are made by (1) striking out the batter; (2) catching a fly; (3) throwing him out; or (4) tagging out a base runner. The batter is struck out when the pitcher succeeds in preventing him from hitting the ball into fair territory within the limit of three strikes. Strikes are counted on the batter whenever he swings at a pitched ball and misses or when he does not swing at a pitched ball that passed him inside the strike zone.

The strike zone is an imaginary rectangular plane in front of the batter. Its top and bottom are in line with his armpits and knees, respectively. Its long sides are imaginary perpendiculars extending upward from both ends of the front rim of home plate (see diagram). The strike zone is thus a rectangle 17 inches wide, facing toward the pitcher, the length of its vertical sides depending on the height of the batter. When the ball passes through this rectangle at any point and the batter does not swing, it is called a strike, as if he had swung and missed.

UPI Compix



Sliding into home plate to score. Base runner Jim Rivera (Chicago White Sox), who knocks the ball from catcher Yogi Berra (New York Yankees), July 28, 1954.

A strike also is counted when the batter fouls the ball to the ground or out of the field of play. Fouling the ball means striking it with the bat but not driving it inside fair territory. This foul strike is not counted if the batter has two strikes against him, provided he swings at the ball in fouling it. If, with two strikes, he fouls in an attempt to bunt—that is, merely blocks the ball with his bat—his foul counts as a third strike and he is out.

Foul balls and force outs

A batter is put out if a member of the fielding team catches and holds a batted ball before it touches the ground, whether it is a fair ball or foul. There is the exception that a foul tip, a pitched ball that the batter merely flicks slightly with his bat, counts as a strike even if it is caught and held by the catcher. It does not count as a putout unless it occurs on the third strike.

A member of the offensive team is tagged out if, when running the bases and not in contact with a base, he is touched by the ball in the hand of a member of the fielding team. A member of the batting team is thrown, or forced, out if he bats a ball that touches the ground before being caught (usually by an infielder or the pitcher) and thrown to the first baseman, who touches first base, while holding the ball securely, before the batter can reach first base.

Only one runner may have title to a base at any given moment. It is therefore possible for a runner to be thrown out at second base, third, or even at home plate without being tagged. The batter becomes a runner entitled to try to reach first base safely the instant he hits a fair ball that strikes the ground. At that same instant, if there was a teammate on first when the ball was hit, that base runner is no longer entitled to first base and must run to second. If runners are on both first and second or runners are on all three bases, they are all forced to run when the batter hits a fair ball that strikes the ground. Any base runner thus forced to run can be put out, or retired, without tagging, by the throw of the ball to a

fielder who can touch the base before the runner reaches it.

The force play

This method of retiring base runners is called the force play. With first base occupied and the ball driven along the ground to the pitcher or an infielder, the ball can often be first thrown to second base for a force out of the man from first base, then relayed to the first baseman to retire the batter—two outs on one play. This is the usual form of the double play in baseball, one of the most effective and spectacular defensive tactics.

A runner can also be thrown out without being tagged if he has left his base before the ball was hit in a fly that is caught. With the catching of the fly, the runner must return to the base he just left before being eligible to advance. If the catcher of the fly throws the ball to that base before the runner returns and touches it, it counts as an out, retiring the runner as well as the batter by a double play of a different type than the force play, "double up," described above.

In very rare cases a runner can be tagged out while standing on a base if, through confusion of mind or misjudgment of the play in progress, he remains standing on a base to which he is no longer entitled. In almost all situations of baseball, however, the bases form anchoring posts (isles of safety) for the players of the side at bat in their progress toward scoring runs.

Intentional pass. Frequently (sometimes as many as three or four times in a single game), a base on balls is given not because of the pitcher's inaccuracy (see above *Base on balls*) but for tactical reasons. If, for instance, the team at bat has a runner in scoring position (*i.e.*, on second or third base, or both) but none on first, and especially if the man at bat is more likely to make a safe hit than the man who will follow him in the batting order, it is considered sound tactics to give the batter an intentional pass, or base on balls, to first base. The catcher moves from behind the plate to one side after the pitcher delivers the ball, and the pitcher delivers four pitches that the batter cannot reach while standing in the batter's box. The batter thus receives a base on balls and must go to first base, thus setting up a possible force play if the next batter can be induced to hit a ground ball into the infield.

Defensive positions and strategies. In the scale of defensive skills, pitcher is number one, shortstop is number two, and second baseman is number three, followed by catcher, centre fielder, first baseman, and third baseman, and, finally, the two flanking outfielders, whose relative importance is usually determined by the contour of their home playing field's outfield boundaries.

This scale of skills is the basis of the baseball expert's argument in rating one team defensively superior to another because it is stronger "down the middle." An outstanding team usually has a star player at each middle position—catcher; "the keystone combination," meaning the second baseman and shortstop; and centre fielder.

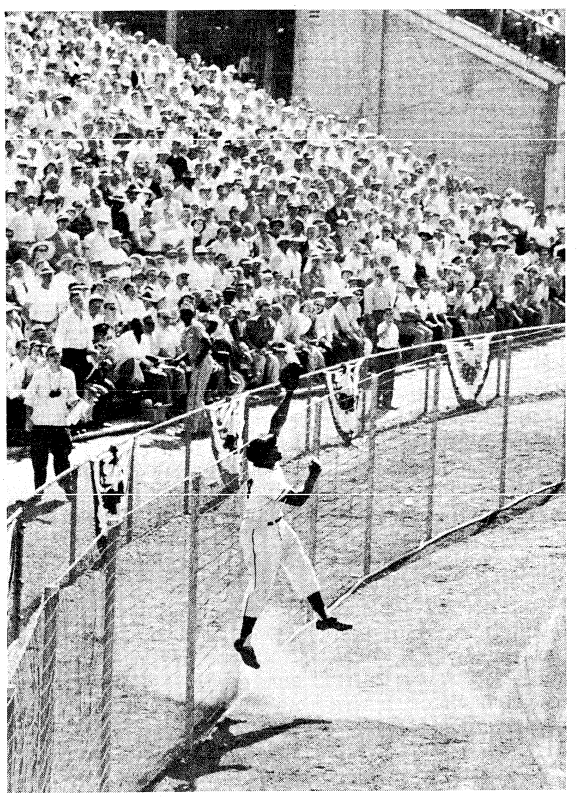
Outfielders. The three outfielders collectively require less special skill than in any of the other six positions. All the catching of thrown balls, all tagging of base runners, and all plays on ground balls to retire runners at bases take place in the infield and the immediately adjacent terrain. The outfielders are stationed in order to be able to catch batted balls driven reasonably high in the air beyond the infield. They must be able to judge the trajectory of such flies and have enough speed to run to the point where the ball will come down.

Batted or thrown balls that pass beyond the infielders along the ground must be run to earth and picked up by the outfielders. Strong throwing arms are essential, as is accuracy in throwing the ball to the right point in the infield. An alert outfielder can also add to his value by moving in as a backup of the infield when the ball has been batted to some point other than his own area and is being thrown about by his teammates in an effort to retire runners. On rare occasions an outfielder aids in a double play, when a base runner misjudges the fielder's ability to catch a well-hit fly ball, which then is returned to a base before the runner can return to it as the rule requires, or when a runner tries to advance after the catch and has misjudged either his own speed or the fielder's ability to throw.

Role of the centre fielder

Because of his central position in the outfield, the centre fielder must possess great speed and expert judgment of a batted ball's trajectory. The accomplished centre fielder makes a study of all the important batters of rival teams. He keeps a mental chart of each batter's power (*i.e.*, toward which point in the rim of the outfield the batter's longest drives usually are directed). The centre fielder will not only station himself at a strategic point to meet each threat but will also direct the playing positions of his outfield teammates on either side. Such a strategic shift of position often makes the difference between a three-base hit and a putout. Almost invariably the great defensive outfielders of baseball history have been centre fielders (Tris Speaker, Max Carey, Terry Moore, Joe DiMaggio, and Willie Mays, to name a few).

UPI Compix



A spectacular fly catch made by Willie Mays (New York Giants, now San Francisco Giants) during the All-Star Game, July 12, 1955.

Infielders. The outfielders form the outer ring of defense against the batting of the opposing team; the infielders form an inner ring, four strong, their function to field the ball at close range. They sometimes capture line drives on the fly, but mainly they pick up grounders skipping along the ground toward the outfield or shooting swiftly across the grass on one or more bounces. The fielding of a grounder is the most characteristic play in baseball, setting it apart from most other sports. When a batted ball strikes the ground, the game becomes a race between the batter's speed in running to first and an infielder's agility in gaining control of the ball and throwing it.

The four infielders shift positions to guard against each batter's individual power, as do the outfielders. They have the additional responsibility of guarding the bases when occupied. When a ball is batted along the ground, only one of them is called upon to gain control of it, but at least one of the others almost always covers a base to take the throw; sometimes two bases must be covered for a possible throw, and sometimes all three. If an attempted bunt is anticipated, the first and third basemen move in from their accustomed positions toward home

plate, ready, as are the pitcher and catcher, to rush toward the bunted ball and to throw it to second, third, or home for the force play or to first to retire the runner.

Infield strategy

In many situations arising in the course of one game, the infielders must adjust themselves correctly, rapidly, and in cooperation. The fielding of grounders—pickup and throw—is the common skill of all four infielders. Each position has its special fielding requirements.

The throw from shortstop is the longest and most difficult. The batted ball may be a swift one shooting toward the outfield to his right or left or straight at him, or it may be a tantalizing slow "hopper," for which he must rush toward the plate in order to field quickly enough to retire the batter.

The second baseman's function on grounders is the same, except that his shorter distance from first gives his throw an extra instant of time. In contrast to the shortstop, however, who always faces first base except when a ball is off to his right, the second baseman on most of his grounders must turn and throw "around the corner."

On a force play at second, whether or not a part of a double play, either the second baseman or the shortstop must cover second to catch the throw, with his foot touching the base, just as the first baseman does on an out at first. The force play is always a potential double play. The second baseman, coming to the base to take the throw, must turn to his left for the throw to first base to complete the double play. Thus, skill in pivoting is important in a second baseman.

The third baseman, playing nearer the batter than the shortstop or second baseman, is not called on to cover as wide a zone. The grounder aimed his way reaches him or is past him into the outfield sooner than at second or short. On his most difficult play, the fielding of a bunt or a dribbling roller halfway between home plate and his position, the third baseman has a throw to first that is shorter than those the second baseman and shortstop ordinarily make, and he often has more time in which to make it.

The first baseman's fielding of grounders is made easier because of his position near the base that the batter is running toward. Often the first baseman is able to touch base with his foot, with ample time to spare, after picking up the grounder. When there is no time for that, his throw may be a mere underhand toss to the player (often the pitcher) who moves over to cover first.

The battery. The pitcher and catcher together are known as the battery or as batterymen. The pitcher may function as an emergency first baseman when the first baseman fields a grounder too far from first to reach the base before the batter. This is not a difficult fielding play for the pitcher, provided he starts running promptly. The distance from the pitcher's box to first is roughly two-thirds as far as the batter must run. The pitcher then catches the ball tossed to him by the first baseman and touches first base with his foot. This is a fairly common play.

Fielding functions of the pitcher and catcher

The pitcher's other fielding functions consist of trying to field any ball batted in his direction. On bunts the pitcher functions as an infielder, covering his share of the zone between the foul lines and between the pitcher's box and home plate. Fielding skill is an important asset to a pitcher.

The catcher, as a fielder, is mainly a catcher of high flies, a thrower, and a guard at home plate. The flies usually soar almost straight upward from the bat and come down in foul territory someplace between the base lines and the grandstand or directly in front of home plate. Ability to perceive the angle at which the ball leaves the bat is necessary to get a quick start in the right direction from home plate.

The "good hands" essential to every player are especially important for the catcher. Throughout the game he must catch the pitched balls not hit by the batter and sometimes pitches that strike the ground near the plate. Sometimes he must jump for high ones or wide ones. The bat swinging directly in his visual line as he reaches for the pitch is no hazard to him. From long experience he has become accustomed to it. It is when the bat flicks the ball slightly in passing that the catcher's good hands are needed most. If the batter has two strikes, the ability of the catcher to catch such a tipped ball and hold it results in a strikeout, ending the batter's turn at bat with a putout; eluding the catcher's grasp, it becomes a mere foul ball, granting the batter one more chance to hit safely.

The catcher sometimes fields a bunt or a half-hit ball just in front of the plate. Agility in pouncing upon the ball and accuracy in throwing to the proper base are then required. The catcher's throwing arm is a valuable element in his team's defense. Base runners will be cautious of straying too far from their bases when the catcher has a quick and strong arm ready to shoot the ball to the baseman for a tag-out before the runner can scramble to safety. When a runner attempts to steal a base, the attempt usually begins at the start of the pitch. Unless the batter hits the ball, the play then becomes a match between the catcher's throwing arm and the runner's speed.

The catcher as counselor

Important as is his fielding, the catcher functions even more importantly as the counselor of the pitcher, as well as the rest of the team. As the only player in the defensive lineup who has the whole game in front of him at all times, the catcher is best placed for advising teammates when necessary.

In general, the catcher directs the pitching strategy. Veteran pitchers, learned in the ways of batters in general and with their own special systems of pitching to each, need little assistance from the catcher. For most pitchers, however, and particularly for those with less experience, the catcher, squatting behind the plate with his hands together between his knees, signals with the fingers of his bare hand what the next pitch should be. His knowledge of the pitcher's and the batter's abilities and peculiarities guides him. The catcher keeps his partner under constant study, alert to signs of weakening, even in a veteran pitcher. Occasionally, the catcher consults with the manager to replace one pitcher with another.

Pitching as the basis of defense. Until a batter hits the ball, the game is a duel between the pitcher (and catcher) and the batter, which is repeated as each batter comes to bat. Each batter that a pitcher strikes out or forces to hit a pop-up (an easily caught fly) or easily fielded grounder is a gain for the defense, preventing runs and bringing the team closer to its turn at bat and a chance to score.

In 19th-century U.S. baseball, up until about 1870, the pitcher was merely a player assigned to put the ball in play by pitching it to the batter to hit. One man generally did nearly all the pitching for a club all season, with a change pitcher to relieve him of some of the load at times. This change pitcher was usually an outfielder. The two would often merely exchange fielding positions without leaving the game. With the start of league baseball in the 1870s, the pitcher became a stronger factor in defensive play. His use of speed and curves became a deciding element in championship contests.

A major league club normally has 25 players on its roster; usually ten are pitchers. A team's infield and outfield lineup may remain intact for weeks on end, barring injuries and illnesses. Its regular catcher occupies that position in more than two-thirds of the season's games.

Starting and relief pitchers

In a staff of ten pitchers, the manager usually earmarks his best four or five at any given time as starting pitchers. These are the rotation starters. They take their turn every four or five days, resting in between. During rest days they do light exercise, run about the outfield during practice, and keep their pitching arms moderately exerted each day, but in general merely keep in shape until their next turn to pitch.

The remainder of the staff constitutes the bullpen squad or the relief pitchers. When the manager or pitching coach detects signs of weakening on the part of the pitcher in the game, these bullpen pitchers begin warming up. Near their bench, regulation-size home plates are embedded in the ground. At the correct distance from each of these plates is a pitcher's plate, or rubber. They warm up by throwing practice pitches to substitute catchers. The bench, together with the practice ground, is known as the bullpen. An effective relief pitcher is one

who can depend upon his ability to control the placement of his pitches, because he often enters the game at a stage in which the offensive team poses a scoring threat with one or more base runners and a dangerous hitter at bat.

Throughout the 1940s, usually one relief pitcher (sometimes two) was used regularly as the bullpen pitcher. The rotation starters, whether they were winning or losing, were accustomed to pitching a complete game; *i.e.*, the full nine innings. In the early 1950s, however, relief pitching began to grow in importance, and it became extremely difficult for a team to win a championship without a strong bullpen.

Control. Pitching demands more exact coordination of mental and muscular faculties and more continuous physical exertion than any other activity in the game. On each pitch the pitcher is aiming at the strike zone or a small part of it (*e.g.*, the lower right-hand corner), 60 feet 6 inches from the plate on which his foot pivots in the act of pitching the ball. His ability to throw the ball where he aims it is known as his control. Lapses of control become apparent to the catcher with a slowing up of fast pitches and a dulling in the break of the curve. To the spectators, they are seen in increasing bases on balls and safe hits. Less often, as a pitcher begins to tire, his loss of control may result in hit batsmen and wild pitches. Each base on balls puts a potential run on first base, as does each batter hit with a pitched ball.

The balk. A balk is a departure from the restrictions imposed upon the pitcher by the rules; it can only be committed with a runner or runners on base. A balk may occur under three circumstances: (1) in pitching the ball to the batter if the pitcher does not have his pivoting foot in contact with the pitching plate; (2) if the pitcher does not hold the ball in both hands in front of him at chest level before starting his delivery, or, once started, does not continue his motion so that the ball is delivered to home plate; and (3) if he starts to make a throw to first base when a runner is occupying that base but does not go through with the throw. When a balk is called by the umpire all runners on base advance one base each. With a runner on third, a balk thus results in one run for the team at bat. Under the old regulations the ball was declared dead as soon as a balk was called. The runners, if there were any at the time, automatically moved ahead one base each as a penalty for the infraction, but the batter was not involved and anything he might have done on the pitch was nullified by the balk. Under a clarification of the balk rule in 1953 the batter became involved. If he hit a balk pitch for a home run or for a shorter base hit, if he drove the ball through a fielder's position for an error, or was thrown a fourth ball or a wild pitch, his team could accept either the play or the penalty, whichever result was more favourable to its interest. Thus, the balk is declined in most cases in which the runners advance more than one base or the batter reaches base safely.

Clarification of the balk rule

The pitching repertoire. A pitcher's speed is usually the index of his general ability. Although the curve was developed in the late 1860s and other variations from the straight line of fast ball pitching developed in succeeding epochs, the fast ball continues to be the basis of pitching skill.

The fundamental or regulation curve is a swerving pitch that breaks away from the straight line downward and across home plate, in the direction of the catcher's left knee if the pitcher throws left-handed.

Some pitchers employ a curving ball that breaks in the opposite way from the regulation curve, a pitch known variously as the fadeaway (Christy Mathewson) or the screwball (Carl Hubbell), or some other name applied by the pitcher himself. The effect is to endow the pitcher with a reverse curve; thus Mathewson, a right-handed pitcher, used the conventional curve to break the ball away from right-handed batters, and his fadeaway to break it away from left-handed hitters. Hubbell, a left-hander, used his screwball to slant the pitched ball out and beyond the reach of right-handed hitters, using the regulation curve when facing a left-handed hitter. In the case of both curves and reverse curves, the ball reaches the batter at a slower rate of speed than the fast ball, and the deception is almost as much a result of the slower ball's falling away from the bat as to its swerving from the straight trajectory from the pitcher's hand to catcher's glove.

Kinds of pitches

A comparatively new pitch, called the slider, was introduced in the 1920s by George Blaeholder, who otherwise had an undistinguished major league career. The slider is a cross between the fast ball and the curve and includes the best features of both. It is thrown with the speed and the motion of the fast ball but, instead of the wide sweep of the conventional curve, has a short and mostly lateral break; in effect, it slides away from the hitter. The break, which is not easily discernible by spectators, is short and quick. The slider is perhaps the most difficult of all breaking pitches to hit and is widely used.

Relatively few pitchers use the knuckle ball, the power of which lies in its lack of axial rotation as it sails toward home plate, making it subject to air currents. The ball is wobbly as it approaches the batter, and so is harder to hit solidly than a spinning ball sailing along "on the beam." The knuckle ball is also difficult to catch; it can be dangerous to use with a runner at third base, from which he could easily score on a passed ball (one that gets past the catcher). The knuckler is thrown with an easy, almost lob, motion and requires not much more effort than a ball thrown when playing "catch." Because of the minimal arm strain, knuckle-ball pitchers have remarkable longevity and can pitch almost every day, if necessary. The most celebrated knuckle-ball pitcher was Hoyt Wilhelm, who, in 1972, at the age of 49, was in his 21st major-league season; he had appeared in more than 1,000 major-league games—a record.

When a player comes to bat, the pitcher and catcher operate against him as a unit. The pitcher is equipped with a strong arm, able to fire the ball across home plate so rapidly as to offer the batter only an instant to decide whether or not to swing at it.

Besides sheer speed, the pitcher has curves, sliders, or other ways of throwing the ball to divert from the usual line of a fast ball. This causes the batter to go off his stride, to swing too early or too late, to strike it merely a glancing blow, or to miss it entirely.

Most batters get their greatest power into a drive when the ball comes to them a little above waist level. A minority are low-ball hitters, preferring a pitch between the belt line and the knees. The batter usually prefers the ball to be either inside or outside; that is, near the edge of the plate that is closer to him or farther from him. Most younger batters can more readily hit a fast ball—that is, a straight one—than a curve or any of the breaking balls. Some batters, usually veterans, have more success when the ball breaks; the breaking ball is always less swift in its course than the fast one. (G.P.L.)

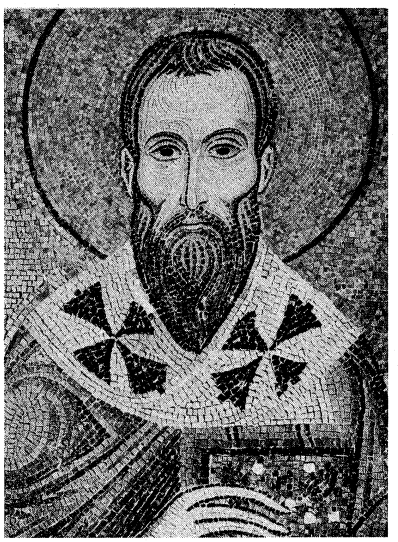
BIBLIOGRAPHY. The *Official Baseball Guide*, containing records and a narrative review of the previous season, is published annually by *The Sporting News*. The standard reference work covering the records of professional players since 1871 is HY TURKIN and S.C. THOMPSON, *The Official Encyclopedia of Baseball*, 5th ed., rev. by SUZANNE TREAT (1970). A good general survey of the modern game is LEONARD KOPPETT, *A Thinking Man's Guide to Baseball* (1967). Early history is covered in HAROLD SEYMOUR, *Baseball: The Early Years* (1960) and *The Golden Age* (1971); and LAWRENCE S. RITTER, *The Glory of Their Times* (1966). For a history of Negro players and teams, see ROBERT W. PETERSON, *Only the Ball Was White* (1970). The life of the professional player is well depicted in JIM BROSNAN'S autobiographical journal, *The Long Season* (1960); JIM BOUTON, *Ball Four* (1970); and GIL HODGES, *The Game of Baseball* (1969). The inside business and politics of the game are covered in BILL VEECK'S autobiographical *Veeck–As in Wreck* (1962) and *The Hustler's Handbook* (1965). Baseball serves as a medium for philosophical concepts in novels by BERNARD MALAMUD, *The Natural* (1952); and ROBERT COOVER, *The Universal Baseball Association, Inc., J. Henry Waugh, Prop.* (1968). A successful baseball fantasy is DOUGLASS WALLOP, *The Year the Yankees Lost the Pennant* (1954), later a hit musical comedy and motion picture, *Damn Yankees*. See also Wallop's *Baseball: An Informal History* (1969), an anecdotal account from 1869.

(Je.Ho.)

Basil the Great, Saint

Basil the Great, as one of the three Cappadocian Fathers —along with his brother Gregory of Nyssa and their friend Gregory of Nazianzus—helped to consolidate Greek Orthodox thought between AD 360 and 390. Along with the two Gregories, he played an important part in securing the final victory of the orthodox supporters of the Nicene faith (defined at the Council of Nicaea in 325) over the heretical Arians, who denied the divinity of Christ. The piety, devotion, and learning of these three theologians won for them a place among the saints and doctors (teachers) of the church. Basil was the leader and inspirer of the group, even though his life's work was unfinished at his death.

Alinari



Saint Basil, detail of a mosaic, 12th century. In the Palatine Chapel, Palermo.

Early life and ecclesiastical career

Basil was born about AD 329, coming from a distinguished family of Caesarea, the capital of Cappadocia, which was a province of Asia Minor of special importance in the 4th century due to its position on the military road between Constantinople and Antioch. The family had been Christian since the days of the persecutions of Christians, which ended early in the 4th century. One of Basil's uncles was a bishop, as later were two of his brothers (Gregory and Peter of Sebaste). He received a literary education, however, which would have fitted him to follow in his father's footsteps as lawyer and orator. He studied at Caesarea and Constantinople and, finally (*c.* 351–356), at Athens, where he developed his friendship with Gregory of Nazianzus. On returning home he began a secular career, but the influence of his pious sister Macrina, later a nun and abbess, confirmed his earlier inclination to the ascetic life. With a group of friends, he established a monastic settlement on the family estate at Annesi in Pontus. In 357 he made an extensive tour of the monasteries of Egypt, and in 360 he assisted the Cappadocian bishops at a synod at Constantinople. He had been distressed by the general acceptance of the Arian Creed of the Council of Ariminum the previous year and especially by the fact that his own bishop, Dianius of Caesarea, had supported it. Shortly before the death of Dianius (362), Basil was reconciled to him and later was ordained presbyter (priest) to assist Dianius' successor, the new convert Eusebius. Basil's abilities and prestige, as well as Eusebius' dislike of asceticism, led to tension between them, and Basil withdrew to Annesi. In 365 he was called back to Caesarea, when the church was threatened by the Arian emperor Valens. His theological and ecclesiastical policy thereafter aimed to unite against Arianism the former semi-Arians and the supporters of Nicaea under the formula "three persons (*hypostases*) in one substance (*ousia*)," thus preserving both unity and the necessary distinctions in the theological concept of the godhead. On Eusebius' death in 370, Basil became his successor, although he was opposed by some of the other bishops in the province.

Anti-Arian activities

As bishop of Caesarea, Basil was metropolitan (ecclesiastical primate of a province) of Cappadocia, and his own diocese covered the great estates of eastern Cappadocia, where he was assisted by a number of "country bishops" (*chorepiscopi*). He also founded charitable institutions to aid the poor, the ill, and travellers. When Valens passed through Caesarea in 371, Basil dramatically defied his demand for submission. But in 372 Valens divided the province, and Basil considered this a personal attack, since Anthimus of Tyana thus became metropolitan for the cities of western Cappadocia. Basil countered by installing supporters in some of the border towns—Gregory of Nazianzus at Sasima and his own brother Gregory at Nyssa. This tactic was only partially successful, but Basil escaped the attacks that Valens launched on orthodox bishops elsewhere. Meanwhile, Basil tried to secure general support for the former semi-Arian Meletius as bishop of Antioch (one of the five major patriarchates of the early church), against Paulinus, the leader of the strict Nicene minority, since he feared that the extreme Nicenes at this point were lapsing into Sabellianism, a heresy exaggerating the oneness of God. During Basil's lifetime, however, this was prevented by the recognition of Paulinus by the bishops of Alexandria and —in spite of a series of negotiations—after 375 by Pope Damasus of Rome.

Basil's numerous and influential writings stemmed from his practical concerns as monk, pastor, and church leader. The *Longer Rules* and *Shorter Rules* (for monasteries) and other ascetic writings distill the experience that began at Annesi and continued in his supervision of the monasteries of Cappadocia: they were to exert strong influence on the monastic life of Eastern Christianity. A notable feature is Basil's strong preference for the monastic life, in which brotherly love can be practiced, as opposed to that of the hermit. Basil's preserved sermons deal mainly with ethical and social problems. One of the best known, the *Address to Young Men*, defends the study of pagan literature by Christians (Basil himself made considerable critical use of Greek philosophical thought). In the *Hexaëmeron* ("Six Days"), nine Lenten sermons on the days of creation, Basil speaks of the varied beauty of the world as reflecting the splendour of God. *Against Eunomius* defends the deity of the Son against an extreme Arian thinker, and *On the Holy Spirit* expounds the deity of the spirit implied in the church's tradition, though not previously formally defined. Basil is most characteristically revealed in his letters, of which over 300 are preserved. Many deal with daily activities; others are, in effect, short treatises on theology or ethics; several of his *Canonical Epistles*, decisions on points of discipline, have become part of the canon law of the Eastern Orthodox Church. The extent of Basil's actual contribution to the magnificent series of eucharistic prayers known as the *Liturgy of St. Basil* is uncertain. But at least the central prayer of consecration (setting apart the bread and wine) reflects his spirit and was probably in use at Caesarea in his own lifetime.

Basil's health was poor, perhaps because of the rigours of his ascetic life. He died on January 1, 379, soon after Valens' death in the Battle of Adrianople had opened the way for the victory of Basil's cause. Vigorous and firm and sure of his own position, in his own time he seems to have been admired rather than loved, even by his intimates. But he was widely mourned and was soon numbered among the saints. In the Greek Church January 1 is still observed as St. Basil's Day and is marked by the use of the Basilian Liturgy.

BIBLIOGRAPHY. The Benedictine edition of Basil's works is reprinted in J.P. MIGNE (ed.), *Patrologia Graeca*, vol. 29–32 (1857). There are modern editions in *Sources Chrétiennes* of *On the Holy Spirit*, 2nd ed. by B. PRUCHE, vol. 17 (1968); and the *Hexaëmeron*, 2nd ed. by S. GIET, vol. 26 (1968), with French translations; and of the *Letters* and *Address to Young Men* in the *Loeb Classical Library;* the *Hexaëmeron* and *Letters* are in *Nicene and Post-Nicene Fathers*, ser. 2, vol. 8

(1895), trans. by BLOMFIELD JACKSON. *The Ascetic Works* were translated by W.K.L. CLARKE (1925).

The basic source for the life of St. Basil is the eulogy by Gregory of Nazianzus (Oration 43). E. VENABLES, "Basilius of Caesarea," in *Dictionary of Christian Biography*, vol. 1, pp. 282–297 (1877), is still important. Among modern sketches, see J. QUASTEN, "Basil the Great," in *Patrology* vol. 3, pp. 204–235 (1960), with bibliography; J.W.C. WAND in *Doctors and Councils*, pp. 31–46 (1962); and HANS VON CAMPENHAUSEN in *The Fathers of the Greek Church*, pp. 84–100 (1963).

(E.R.Ha.)

Basil I the Macedonian

The Byzantine emperor Basil I, who reigned from 867 to 886, was the founder of the Macedonian dynasty that lasted until 1056. He also inaugurated a much-needed legal reform, heralded during his own reign by two small handbooks, the *Procheiron* and the *Epanagoge*, and completed under his son and successor, Leo VI, with the appearance of the Greek code known as the Basilica.



Basil I, coin, 9th century. In the British Museum.
By courtesy of the trustees of the British Museum

Basil's rise to power

Basil came of a peasant family that had settled in Macedonia, perhaps of Armenian origin. He was a handsome and physically powerful man who gained employment in influential official circles in Constantinople and was fortunate enough to attract the imperial eye of the reigning emperor, Michael III. After rapid promotion he became chief equerry, then chamberlain, and finally, in 866, co-emperor with Michael. Quick to sense opposition, he forestalled the Emperor's uncle, the powerful Caesar Bardas, by murdering him (866) and followed this by killing his patron, Michael, who had begun to show signs of withdrawing his favour (867).

From the mid-9th century onward, the Byzantines had taken the offensive in the agelong struggle between Christian and Muslim on the eastern borders of Asia Minor. Basil continued the attacks made during Michael III's reign against the Arabs and their allies, the Paulicians, and had some success. Raids across the eastern frontier into the Euphrates region continued, though Basil did not manage to take the key city of Melitene. But the dangerous heretical Paulicians on the borders of the Armenian province in Asia Minor were crushed by 872, largely owing to the efforts of Basil's son-in-law Christopher. In Cilicia, in southeast Asia Minor, the advance against the emir of Tarsus succeeded under the gifted general Nicephorus Phocas the Elder. Though Constantinople had lost much of its former naval supremacy in the Mediterranean, it still had an effective fleet. Cyprus appears to have been regained for several years.

Basil's plans for Italy involved him in negotiations with the Frankish emperor Louis II, the great-grandson of Charlemagne. The Byzantine position in southern Italy was strengthened with the help of the Lombard duchy of Benevento, and the campaigns of Nicephorus Phocas the Elder did much to consolidate this. The region was organized into the provinces of Calabria and Langobardia. But key cities in Sicily, such as Syracuse in 878, still continued to fall into Muslim hands, an indication of the strength of Arab forces in the Mediterranean.

Relations with Rome

Another arm of Byzantine policy was the attempt to establish some measure of control over the Slavs in the Balkans. Closely allied to this was the delicate question of ecclesiastical relations between Constantinople and Rome. During Basil I's reign, the young Bulgar state accepted the ecclesiastical jurisdiction of Constantinople (870). This had significant results both for the Balkan principalities and for the Orthodox Church, as well as greatly strengthening Byzantine influence in the south Slav world. Basil had succeeded to a quarrel between Photius and Ignatius as to which was to be patriarch of Constantinople. This had international implications, since appeals had been made to Rome. Immediately on his accession, Basil attempted to win support at home and to conciliate Rome by reinstating the deposed patriarch Ignatius and excommunicating Photius. Eventually, Photius was restored by Basil on the death of Ignatius (877) and recognized by Rome in 879. Contrary to the belief that used to be held, no "second schism" occurred. Basil successfully resolved the tension between liberal and strict Byzantine churchmen and managed to maintain a show of peace between East and West despite Rome's displeasure at the marked extension of imperial influence in the new Balkan principalities.

Toward the end of his life, Basil seemed to suffer fits of derangement, and he was cruelly biassed against his son Leo. He died on the hunting field in 886. The 11th-century historian Psellus wrote of his dynasty as "more blessed by God than any other family known to me, though rooted in murder and bloodshed." But Macedonian historians were understandably biassed in favour of the existing dynasty, to the detriment of the rulers it had supplanted. Recent historical research has raised the stature of Basil's predecessor, Michael III, and his regents. It is now generally agreed that the "new age" in Byzantine history began with Michael III in 842 and not with the Macedonian dynasty in 867. Basil's policies were largely determined, both at home and abroad, by factors not of his own making.

BIBLIOGRAPHY. A good short outline of Bzyantine history is G. OSTROGORSKY, *Geschichte des byzantinischen Staates*, 3rd ed. (1963; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968). Recent views on Basil I, particularly in relation to his predecessor Michael III, are presented in H. GREGORIE, "The Amorians and Macedonians 842–1025," *Cambridge Medieval History*, new ed., 4:105–192 (1966). A lively account of ecclesiastical problems and Byzantine relations with the Slavs is G. EVERY, *The Byzantine Patriarchate 451–1204*, 2nd ed. rev. (1962).

(J.M.H.)

Basil II Bulgaroctonus

The reign of Basil II, widely acknowledged to be one of the outstanding Byzantine emperors, admirably illustrates both the strength and the weakness of the Byzantine system of government. His indomitable and forceful personality and his shrewd statesmanship enabled him to expand imperial territory in the Balkans and in the regions of Mesopotamia, Georgia, and Armenia. At home he increased imperial authority by ruthlessly attacking the powerful landed interests, both of the military aristocrats and of the church. Under him the Byzantine Empire reached its peak, but his reign demonstrates the inherent weakness of an imperial autocracy that depended so much on the character of the ruler.

Basil II (*c.* 958–1025), son of Romanus II and Theophano, was crowned co-emperor with his brother Constantine in 960, but as minors both he and his brother remained in the background. After their father's death in 963, the government was effectively undertaken by the senior military emperors, first by Nicephorus II Phocas, their stepfather, and then by John I Tzimisces. On the latter's death (976) the powerful great-uncle of Basil II, the eunuch Basil the chamberlain, took control. His authority—and that of Basil II—was challenged by two generals who coveted the position of senior emperor. Both related to emperors, they belonged to powerful landed families and commanded outside support from Georgia and from the Caliph in Baghdad. After a pro-



Basil II, illumination from the frontispiece of his psalter; showing Bulgars prostrate before the Emperor. In the Biblioteca Nazionale Marciana, Venice (MS. Gr. 217).

By courtesy of the Biblioteca Nazionale Marciana, Venice

longed struggle both were defeated by 989, though only with the help of Russians under Vladimir of Kiev, who was rewarded by the hand of Basil's sister Anna on condition that the Kievan state adopted Christianity. Certain Russian soldiers remained in Basil's service, forming the famous imperial Varangian guard. Eventually, Basil asserted his claim to sole authority by ruthlessly eliminating the dominating grand chamberlain, who was exiled in 985.

Consolidation of imperial authority

Basil aimed solely at the extension and consolidation of imperial authority at home and abroad. The main fields of external conflict were in Syria, Armenia, and Georgia on the eastern front, in the Balkans, and in south Italy. He maintained the Byzantine position in Syria against aggression stirred up by the Fāṭimid dynasty in Egypt and on occasion made forced marches from Constantinople across Asia Minor to relieve Antioch. By aggression and by diplomacy he secured land from Georgia and from Armenia, with the promise of more to come on the death of the Armenian ruler. He is, however, best known for his persistent and ultimately successful campaigns against a revived Bulgarian kingdom under its tsar Samuel. This ruler centred his activities in Macedonia and established his hegemony in the west Balkans. From 986 until 1014 there was warfare between Byzantium and Bulgaria, interrupted from time to time by Basil's intermittent expeditions to settle crises on the eastern front. Basil enlisted Venetian help in protecting the Dalmatian coast and Adriatic waters from Bulgarian aggression. Year by year he slowly penetrated into Samuel's territory, campaigning in winter as well as summer. Finally, holding northern and central Bulgaria, he advanced towards Samuel's capital, Ochrida, and won the crushing victory that gave him his nickname, "Slayer of the Bulgars" (Bulgaroctonus). It was then that he blinded the whole Bulgarian army, leaving one eye to each 100th man, so that the soldiers might be led back to their tsar (who died of shock shortly after seeing this terrible spectacle). Thus the revived Bulgarian kingdom was incorporated into the Byzantine Empire. Basil then looked further west and planned to strengthen Byzantine control in south Italy and to regain Sicily from the Arabs. He attempted to establish a Greek pope in Rome and to unite in marriage the German (though by birth half Byzantine) ruler Otto III with his favourite niece Zoe. Both schemes failed, but he was more successful in south Italy, where order was restored, and at his death preparations were being made for the reconquest of Sicily.

The ruthlessness and tenacity that served Basil in his military and diplomatic activities were displayed in his domestic policy as well. Its keynote was the strengthening of imperial authority by striking at his overpowerful subjects, particularly the military families who ruled like princes in Asia Minor. The by-product of this policy was the imperial protection of the small farmers, some of whom owed military service to the crown and paid taxes to the central exchequer. Title to land was rigorously inspected, and vast estates were arbitrarily confiscated. Thus, in spite of his costly wars, Basil left a full treasury, some of it stored in special underground chambers.

Domestic policy

Both in near-contemporary history and in manuscript illustrations, Basil is pictured as a short, well-proportioned figure, with brilliant light-blue eyes, a round face, and full, bushy whiskers, which he would twirl in his fingers when angry or while giving an audience. He dressed plainly and even when wearing the purple chose only a dark hue. An abrupt speaker, he scorned rhetoric yet was capable of wit. He has been described as mean, austere, and irascible, spending most of his time as though he were a soldier on guard. He knew only too well the danger of any relaxation. He showed no obvious interest in learning, but he did apparently commission works of religious art, and he had churches and monasteries rebuilt or completed in Boeotia and in Athens, though this may be accounted for by conventional piety. He seems never to have married or had children. On his death there was no able military aristocrat or other leader to take the situation in hand, and thus Basil II's work was rapidly undone.

BIBLIOGRAPHY. G. SCHLUMBERGER, *L'Épopée Byzantine*, vol. 1–2 (1896–1900), the classic narrative account, now in need of some revision; H. GREGOIRE, "The Amorians and Macedonians 842–1025," *Cambridge Medieval History*, new ed., vol. 4, pt. 1, pp. 105–192 (1966); G. OSTROGORSKY, *Geschichte des byzantinischen Staates*, 3rd ed. (1963; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968), a good short outline with bibliography; P. CHARANIS, "The Monastic Properties and the State in the Byzantine Empire," *Dumbarton Oaks Papers*, no. 4 (1948); H. AHRWEILER, *Byzance et la mer* (1966), for organization of naval resources.

(J.M.H.)

Basin and Range Province

The unique topography, climate, and drainage of a great natural region of the western United States combine to make the area one of the most distinctive surface features of the entire North American continent and of considerable interest to a variety of natural scientists. The term Basin and Range Province is used by geographers and geologists to describe this 190,000 square mile (492,000 square kilometre) expanse, stretching from the Sierra Nevada Range on the west to the fringes of the Rocky Mountains—specifically the Wasatch Mountains of central Utah—on the east, and tapering to the Mojave Desert of California in the south. The term Great Basin, also in use, is probably more common among the people who inhabit the region (for related information see ROCKY MOUNTAINS; NORTH AMERICA; NORTH AMERICAN DESERT).

Environmental characteristics. The ranges of the region have been likened, in an old survey report, to a group of caterpillars, all crawling irregularly northward. Each range generally has a linear character and is separated from adjacent ranges, east and west, by wide valleys. The ranges are from 60 to 120 miles in length, and from three to 15 miles in width. The valleys are usually somewhat wider than the ranges and are for the most part deserts, with floors 1,000 to 6,000 feet (300 to 1,800 metres) above sea level. The ranges have peaks commonly reaching 9,000 or more feet above sea level, and where this occurs they catch a moderate amount of precipitation and support conifer stands. Some of the higher ranges have small permanent streams, but most of these disappear underground when they reach the basin. The Sierra Nevada Range otherwise blocks rain-bearing winds from the Pacific, forming a "rain shadow" over the entire region, which has an annual rainfall of 10 inches (254 millimetres) or less, and supports little more than a sparse desert or semidesert vegetation.

The Great Basin is particularly noted for its internal drainage system, whereby rain falling on the surface

leads eventually to closed valleys and does not reach the sea. The Humboldt River of northern Nevada, for example, rises in ranges in the northeast of the state, drains a number of small valleys on its way westward, and ends in a closed basin called Humboldt Sink. The Great Salt Lake (*q.v.*) lies in the final and lowest catchment basin of western Utah and gathers much of the drainage of the region that has not evaporated or seeped underground en route. Some of the smaller closed basins may be draining underground to adjacent, lower, basins, and thus may contain temporary lakes, which hold water only during the spring run-off from the ranges, or after flash-flood storms. The Colorado River (*q.v.*) cuts across the Great Basin in southwestern Utah, southern Nevada, western Arizona, and southeastern California, and, since it is deeply incised, drains the tributary valleys on either side, forming an exception to the internal drainage pattern.

Salt Lake drainage catchment

The Great Basin is recognized as extending to the Sonoran Desert area of southern Arizona and to north-central Mexico, although there the ranges are subdued and more circular than linear, with the intervening valleys very wide and the surface drainage better developed than it is further north. The Gila River drains most of the region into the Colorado River.

Exploration. The arid Great Basin for a long time thwarted cross-country travel to California, and hence impeded significantly the development of the American nation. Jedediah Smith (1798–1831), a great explorer of the West, made the first journey across the Basin in 1824, but did not document his travels. He was followed by John C. Frémont, who surveyed an eastern swathe of the Great Basin in 1846, but did not cross it. In the summer and fall of 1846 the ill-fated Donner Party crossed the Basin from what became Salt Lake City to the Sierra Nevada, following the Humboldt River, but most members lost their lives the following winter in the dreadful snows of the Sierras. The California Gold Rush brought thousands westward in 1848 and 1849, many of them reaching Salt Lake City and then attempting alternate routes across the Great Basin. Most of them proceeded northward into southern Idaho, around the Great Salt Lake Desert, entering the Humboldt River route in northwestern Nevada.

A survey made in 1867–78 produced the first federally-sponsored scientific account of the climate, travel conditions, and resources of the Utah-Nevada region. With the subsequent discovery of many mineral deposits in western Utah, Nevada, and southern Arizona, the U.S. Geological Survey made many studies of the specific mining districts and published a variety of reports. In the 1950s a flurry of interest in oil resources resulted in a number of valuable regional geological studies, adding to the work produced by the appropriate state geological surveys.

Geological background. Many scientists have characterized the ranges and valleys of the Great Basin as huge blocks of the earth's crust, which have been uplifted, dropped, and tilted. Enormous cracks, or faults, bound the blocks, and the uplifted parts have been eroded, over geologic time, with the debris accumulating over the depressed parts. Several such blocks are to be found in both western Utah and western Nevada. The blocks are 15–30 miles across and follow an approximate north-south direction. There are about 30 major fault-bounded blocks between the Wasatch Range and the Sierra Nevada. The movement in the faults—a response to stresses in the earth's crust—has been in an up and down direction, 1,000 to 15,000 feet in extent, although toward the western edge of the province some horizontal movement has been observed.

In many places volcanic rocks have been cut and displaced by the block faults, and since the volcanic rocks are about 30,000,000 years old, the faulting is obviously younger than that. Since the faulting occurs in small steps of a few feet each, and since most of the faults have total displacements of several thousand feet, it is believed that, in general, the process took an enormous period of time. Furthermore, many of the faults exhibit fresh surfaces, indicating recent movement, while there are historical records of earthquakes and constant contemporary micro-earthquakes, indicating that faulting has continued to the present day.

The Great Basin is nevertheless geologically youthful, and it is quite likely that it obscures older mountain systems that were, respectively, eastward extensions of the Sierra Nevada and westward extentions of the developing Rocky Mountains.

Resources. Minerals have proved to be the greatest resource of the Great Basin. There are large copper mines at Bingham, Utah, and Ely, Nevada, and a number of major copper mines in Arizona make it the nation's leading copper producing state, with a good share of the nation's molybdenum (a rare metal) produced as a by-product, along with gold. Silver, gold, lead, zinc, and copper have been found in many of the widely scattered mining districts of the region and result from the widespread penetration of sedimentary rocks by hot, mineralizing solutions associated with former volcanic activity. Many valuable nonmetallic minerals, such as limestone, dolomite, gypsum, and pumice, are also widely quarried.

Most of the population in Utah is located along the west base of the Wasatch Mountains focussing on Salt Lake City, with sustaining water supplies coming from streams in the mountains and also from wells tapping the great underground water reservoir trapped beneath the adjacent valley.

Similarly, on the other side of the Great Basin, a good part of the population of western Nevada, centering on Reno, is found along the east front of the Sierra Nevada, which supplies most of the water. To the south, in the Las Vegas, Tucson, and Phoenix areas of Nevada and Arizona, water comes largely from wells. In the large Las Vegas Valley the water table has not yet been lowered noticeably, but in the Tucson and Phoenix population centres the water table has been drawn down hundreds of feet, and by the 1970s the problem of future water supply had become a serious consideration.

Development. The Mormons settled the eastern part of the Great Basin beginning in 1847, and now about 1,000,000 people live there. The early prospectors were transient inhabitants from elsewhere in the Great Basin, but from their early towns and supply points a number of lasting communities have evolved. These are sustained by livestock raising, mining, and railroading. The dry, warm climate has attracted many people to Arizona and southern Nevada. The cities there have had a rapid expansion rate, especially in the last 30 years. The desert floors of south central Nevada have been used for testing of nuclear devices since 1945.

BIBLIOGRAPHY. A.J. EARDLEY, *Structural Geology of North America*, 2nd ed., ch. 31 (1962), a general review of the geological characteristics of the Basin and Range Province; JAMES GILLULY, "Volcanism, Tectonism, and Plutonism in the Western United States," *Spec. Pap. Geol. Soc. Am. 80* (1965), a coherent summary of the general physical history of the region; R.J. ROBERTS, "Tectonic Framework of the Great Basin," *UMR Journal*, ser. 1, no. 1 (1968), chiefly an outline of the pre-Basin and Range geological history, but also proposes a theory for the origin of the block faulting.

(A.J.E.)

Basketball

The only major sport strictly of U.S. origin, basketball was invented by James Naismith (1861–1939) on or about Dec. 1, 1891, at the International Young Men's Christian Association Training School (now Springfield College), Springfield, Massachusetts, where Naismith was an instructor in physical education. His invention of the game was motivated by a desire to relieve the boredom of students in gymnasium classes, where marching, calisthenics, and apparatus work were the usual activities.

After considerable trial and error and much meditation, Naismith, borrowing, modifying, and sometimes inserting ideas from football, soccer, hockey, and other outdoor games, prepared a set of 13 simple rules embodying five principles that still govern today's game:

1. There must be a ball—large, light, and handled with the hands. 2. There shall be no running with the ball. 3. No mem-

Principles of the game

ber of either team shall be restricted from getting the ball at any time it is in play. 4. Both teams are to occupy the same area, yet there is to be no personal contact. 5. The goal shall be horizontal and elevated.

Originally, Naismith planned to nail square boxes as targets at opposite ends of the overhead gymnasium board track, but as boxes were not immediately available he used instead two half-bushel peach baskets, which gave the sport its name.

In making his first test of the game, Naismith explained the rules to a class of 18, asked two of them, Frank Mahan and Duncan Patton, to choose sides of 9 men each, and the game was underway. To everyone's amazement, the students became enthusiastic. After much running and shooting, William R. Chase chanced to connect on a midcourt shot and that historic contest ended in a 1–0 score.

In marked contrast to their former attitude, the Springfield students awaited their daily gym classes with immense interest. When all 18 went home for the Christmas vacation, they told their friends and the local YMCA people about the newly invented game. Numerous associations wrote Naismith for a copy of the rules, which were published in the Jan. 15, 1892, issue of the *Triangle*, the campus paper.

THE EARLY YEARS OF BASKETBALL

In the early years, the number of players on a team varied according to the number in the class and the size of the playing surface. At Cornell University in 1892, Ed Hitchcock, Jr., the physical director, divided his class of 100, but the results were dismal since everyone converged upon the ball. In 1894 teams started to experiment with five on a side when the playing surface was less than 1,800 sq ft (167.2 sq m); the number rose to seven when the gymnasium measured 3,600 sq ft, and back to nine when the playing areas exceeded that. In 1895 the number was occasionally set at five by mutual consent and, two years later, the rules stipulated five, and this number has remained.

Since five of the original players were Canadians, it is not surprising that Canada was the first country outside the U.S. to play the game. It was introduced in France in 1893, in Australia, China, and India soon thereafter, in Japan in 1900, and in Iran in 1901. A missionary is said to have tested the game in São Paulo, Brazil, in 1896. The game was demonstrated in London in June 1894 on the 50th anniversary of the founding of the YMCA.

While basketball helped swell the membership of YMCA's because of the availability of their gyms, within five years the game was outlawed by various associations because the outcome of the contests created ill-feeling among the members; where the gyms formerly were occupied by classes of 50 or 60 members, the areas were now monopolized by only 10 to 18 players—thus inactivating the others. Since the physical directors were largely judged by the size of their classes, they sensed their positions were in jeopardy. The banishment of the game induced many members to terminate their YMCA membership and to hire halls to play the game, thus paving the way to the professionalization of the sport (see below *Professional Basketball*).

Early uniforms and equipment

Originally, players had the choice of three styles of uniforms: knee-length football trousers; jersey tights, as commonly worn by wrestlers; and short padded pants, forerunners of today's uniforms, plus knee guards. The courts often were of irregular shapes with occasional obstructions such as pillars, stairways, or offices that interfered with play. In 1903 it was ruled that all boundary lines must be straight.

In 1892 Lew Allen, of Hartford, Connecticut, conceived the idea of fashioning a cylindrical basket of heavy woven wire to replace the peach baskets, and the following year the Narragansett Machinery Co., of Providence, Rhode Island, marketed a hoop of iron with a hammock style of basket. Originally, a ladder was used to remove the ball from the net after each successful throw. Later, a pole was used to retrieve the ball, then a chain fastened to the bottom of the net with a cord that extended within reach of the official. Nets open at the bottom, thus permitting the ball to drop through, were adopted in 1912–13. In 1895–96 the points for making a basket (field goal) were reduced from three to two, and for making a free throw from three to one.

In the first years, spectators delighted in occupying positions behind the basket so that they could lean over the railings and deflect the ball to favour one side or hinder the other, and in 1895 teams were urged to provide a 4 × 6 ft screen to eliminate such unsportsmanlike interference. Soon after, wooden backboards proved more suitable. Glass backboards were legalized by the professionals in 1908–09 and by colleges in 1909–10. Fan-shaped backboards were made legal in 1940–41, and transparent backboards, although commonly used in public auditoriums since the mid-1930s, were finally authorized in 1946–47. In 1920–21 the backboards were moved two feet, and in 1939–40 four feet, from the end lines to reduce frequent stepping out-of-bounds.

The soccer ball was used for the first two years. In 1894, the Overman Wheel Co., bicycle manufacturers of Chicopee Falls, Massachusetts, marketed the first basketball, laced, close to 32 in. or about 4 in. larger than the soccer ball in circumference and weighing less than 20 oz. By 1948–49 when the laceless molded ball was made official, the size had been set at 30 inches (76 centimetres). The weight of the ball is 20 to 22 oz. (565–625 g).

Introduction to colleges

The distinction of becoming the first college to play the game belongs to either Geneva College (Beaver Falls, Pennsylvania) or the University of Iowa in 1892. At the former institution, C.O. Bemis chanced to return to Springfield, heard about the new sport and tried it out with his students. At Iowa, H.F. Kallenberg, who attended Springfield in 1890, wrote Naismith for a copy of the rules and similarly presented the game to his undergraduates.

A friendship that had developed in 1890–91 at Springfield between Kallenberg and Amos Alonzo Stagg, who became athletic director at the new University of Chicago the following fall, led to the playing of the first college basketball game with five on a side, at Iowa City, Iowa, on Jan. 18, 1896. The University of Chicago won, 15–12, with neither team using a substitute. Kallenberg refereed that game, a common practice in that era—and some of the spectators, although attending their first game, took exception to some of his calls.

The first official rules committee had been formed in 1896. The colleges formed their own rules committee in 1905. In 1913 there were as many as five sets of rules: the collegiate, YMCA–Amateur Athletic Union, state militia groups, and two varieties of professional rules, and often teams agreed to play under a different set for each half. To establish some measure of uniformity, the colleges, AAU, and YMCA formed a Joint Rules Committee in 1915, in 1933 renamed the National Basketball Committee (NBC) of the United States and Canada. It meets each March to formulate rules for the Canadian Intercollegiate Union, the Canadian Amateur Basketball Association, the National Collegiate Athletic Association, the National Federation of State High School Athletic Associations, National Junior College Athletic Association, and the YMCA. Rules for women and professional and international rules vary slightly from those of the NBC.

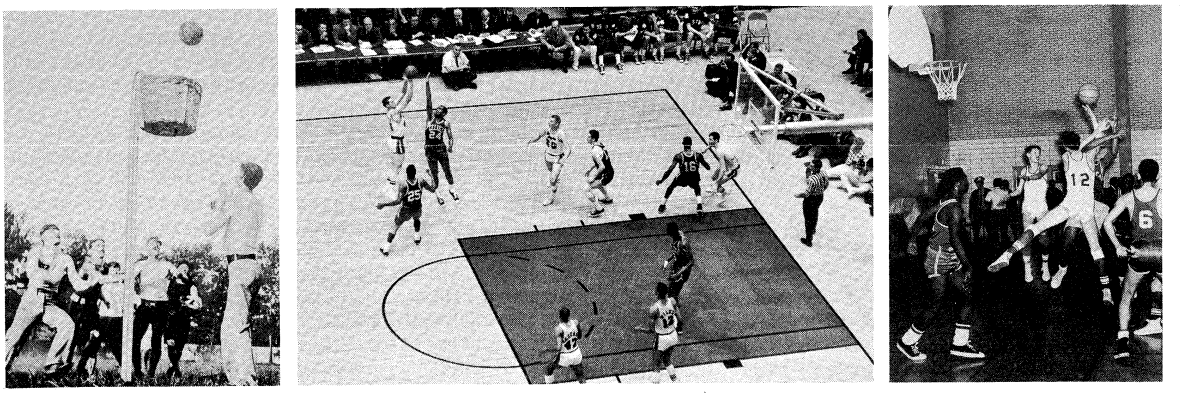
Moves toward uniform rules

GROWTH OF THE GAME

In the 20th century, and especially since World War II, basketball has grown into a major sport, not only in the country of its origin but also in international competition including the Olympic Games. The history of that growth is discussed below as it relates to the game's four major branches: U.S. school and college basketball, women's basketball, international basketball, and professional basketball.

For listings of U.S. college tournament, international (world men's and world women's), and professional champions, see RELATED ENTRIES under SPORTING RECORD in the *Ready Reference and Index*.

U.S. school and college basketball. The growth and development of the school and college game over the



(Left) An outdoor game in 1892. Players shoot into a closed-bottom peach basket. (Centre) A modern court in a U.S. professional game, 1965, showing free-throw lane and rectangular backboard. (Right) Boys playing in a public field house in Chicago. Note fan-shaped backboard.

By courtesy of (left) the Basketball Hall of Fame, Springfield, Mass., (right) Chicago Park District; photograph, (centre) *Sports Illustrated*/ Sheedy and Long © Time Inc.

years is reflected in the rules changes made by the NBC and its predecessors. Intended to improve, refine, and equalize competition, many of them were designed to simplify procedures and speed up the game.

In the early years of the game scores were usually low (winning scores of less than 30 points were not unusual) and action, at least from the spectators' viewpoint, was slow. A team with a small lead usually would protect it by stalling—retaining possession of the ball indefinitely by dribbling and passing it back and forth. To reduce opportunities for this manoeuvre a rule change in 1932–33 required that a team must advance the ball beyond midcourt within 10 sec or lose possession of the ball, and no player could stand within the foul lane with the ball for more than 3 sec. Another step to speed the game was the elimination, in 1937–38, of the centre jump after a field goal or foul shot, the ball being put in play from out-of-bounds on the end line where the score was made. In line with these efforts, Frank W. Keaney, coach at Rhode Island University for 28 years, is credited with introducing the now-popular "racehorse," or fast breaking, high scoring style of play. Another 1930s innovation was the one-hand shot, popularized by Stanford University's Angelo (Hank) Luisetti who scored as many as 50 points a game.

Other rule changes over the years have attempted to offset some of the advantages enjoyed by a team with one or more extremely tall players. Joe Lapchick at 6 ft 5 in. was regarded as a big man when he played for the Original Celtics in the 1920s, and young men much over that height generally were thought to be too clumsy and fragile to be really good players. But some did manage to get on teams, and in 1936–37 the 3-sec foul lane rule was changed to limit their effectiveness under the basket by requiring that no member of the team having the ball could stand within the foul lane more than 3 sec—with or without the ball. And in 1937–38 defensive players were prohibited from interfering with the ball or the basket while the ball was on or above the basket—a rule obviously directed against those who were able to reach that high. Later, in 1944–45, goal tending (*i.e.*, interfering with a ball above the level of the hoop on its downward flight) became illegal.

Meanwhile the tall men began to make their mark, led by George Mikan, 6 ft 10 in. who in the 1940s scored over 550 points in each of his last two years at De Paul University, and more than 11,000 points in nine seasons as a professional. Ed Macauley, 6 ft 8 in., of St. Louis University was ranked college player of the year in 1948 and in ten professional years was named to all-pro honours three times. Coaches began to recruit and develop tall boys—from grade school up—and to build their teams around the tallest. An indication of the prominent role of tall men in the game may be seen in the 20-year All America team named by the Associated Press in 1967: Bill Bradley, 6 ft 5 in., Princeton; Ed Macauley, 6 ft 8 in., St. Louis; Bill Russell, 6 ft 10 in., San Francisco; Wilt Chamberlain, 7 ft 1$\frac{1}{16}$ in., Kansas; and Oscar Robertson, 6 ft 4 in., Cincinnati. That same year, when 7 ft 1$\frac{3}{8}$ in. Lew Alcindor (later Kareem Abdul Jabbar) began to play at UCLA, the colleges banned the dunk shot, in which a man who can reach higher than the basket stuffs the ball through the hoop.

The big men did not completely take over the game. Bob Cousy, 6 ft 1 in., for example, was generally regarded as one of the greatest playmakers and best ball handlers of all time and, with Bill Russell in the 1950s and 1960s, led the Boston Celtics to six world professional titles in seven seasons. Some developments over the years have helped players whatever their size. Thus the slightly smaller and more perfect basketball, coupled with the incessant practice of youngsters to better themselves, have led to remarkable performances, such as that of Pete Maravich, 6 ft 5 in., of Louisiana State University who in 1970 set an all-time college scoring record with successive yearly totals of 741, 1,138, 1,148, and 1,381 points, the latter representing a 31-game average of 44.5 points per game.

Basketball's growing popularity has been largely due to new strategy, shooting excellence, larger playing arenas, the desire of youngsters to secure college scholarships on the strength of their playing abilities, and the salaries paid outstanding players by the professional teams. Because of the nationwide appeal of the game, an increasing number of colleges, especially smaller institutions, have either abolished or de-emphasized football and have relied upon basketball to gain national athletic prominence. Between December and March, seldom does a week go by when some relatively small or lightly regarded college does not upset a larger or formidable foe. Coast-to-coast scheduling and an increasing number of national or sectional televising of games permits the average fan to see and evaluate the best teams and players.

Intersectional basketball contests

Intersectional college contests were a rarity before Dec. 29, 1934, when arrangements were made for two out-of-state teams to play two local teams at Madison Square Garden arena in New York City. On that date, the University of Notre Dame (Indiana) lost to New York University 25–18 and St. John's lost to Westminster College (Pennsylvania) 37–33 before an audience of 16,000. The success of that season's program induced principal cities to follow suit as the growth of air travel permitted teams to play anywhere in the country with a minimum of classes missed by the student players.

Such intersectional contests brought about a more uniform interpretation of officiating, helped introduce new playing styles and offensive and defensive strategies, and focussed greater public attention upon the more successful teams and finer players. Following World War II, many colleges built larger indoor field houses to accommodate their undergraduates and alumni, and many municipalities put up new and larger auditoriums, thus enabling college teams to make long trips with intermittent stops for games. This not only enriched their athletic as-

sociation coffers but also brought them greater national athletic fame.

By 1970 at least 30 colleges travelled more than 10,000 miles in a season or played before more than 200,000 spectators. During the 1968–69 campaign, the national champion UCLA covered 22,523 miles and appeared before 393,390 fans. College basketball as a whole drew 22,000,000 spectators.

Basketball tournaments

The first national tournament was staged by the AAU in 1897 and was won by the 23rd-Street YMCA, of New York City, which subsequently became a professional travelling team, the New York Wanderers. With most of the best collegians turning professional upon graduation, the AAU tournament today is at its lowest ebb. Although the YMCA was prominently identified with the game in its early years, it did not hold its first national tournament until 1923, and it was discontinued after 1962.

Naismith was of considerable help to Emil S. Liston in organizing the first national college tournament in 1937, conducted by the National Association of Intercollegiate Athletics, in Kansas City, Missouri. This group now numbers 550, largely small colleges. This colourful tournament consumes an entire week of eliminations from a field of 32 teams, each determined via district play-offs.

In 1938 New York City basketball writers organized the first National Invitational Tournament (NIT); the local colleges took it over the following year, and all games are played at Madison Square Garden. Today, the field consists of 16 teams, largely independents and runners-up in conference play.

In 1939 the NCAA started its tournament, now called the University Division. The field consists of the champions of the nations' principal conferences (*e.g.*, the Eastern College Athletic Conference, Western [Big Ten] Conference, Pacific Coast Intercollegiate Athletic Conference), plus the teams without conference affiliations that have the best competitive records each year. There are sectional eliminations, and the survivors of four area playoffs qualify for the finals, the winner being acclaimed as the national champion. To accommodate its minor college members, the NCAA inaugurated the College Division tournament in 1957, with the winners of the four-team play-of district championships providing an eight-team field for the finals, which are played in Evansville, Indiana, each March. Since 1945, the National Junior College Tournament has been held at Hutchinson, Kansas. In addition to the above, there are about 200 preconference tournaments usually during the Christmas recess, other national small college tournaments, and two women's national tourneys.

High school basketball

More than 20,000 high schools in the United States had basketball teams in the 1970s. Forty-eight of the 50 states annually conduct statewide tournaments in one or as many as five divisions. The other two states, New York and California, which formerly conducted state tournaments, now have district or area championships. Ten states—Arkansas, Georgia, Louisiana, Iowa, Oklahoma, Pa., Rhode Island, South Carolina, Tennessee, and Texas—conduct annual tournaments for girls. Between 1917 and 1930, the University of Chicago, under Stagg, sponsored a national schoolboy tournament that helped disseminate the finer points of various systems of play and standardize scholastic eligibility rules, but the National Federation of State High School Athletic Associations, founded in 1920, opposed its continuance. The Federation consists of all 50 state high school athletic and/or activities associations, the District of Columbia, and similar groups from Guam, the Philippines, Virgin Islands, and seven Canadian provinces.

The most famous high school team of all time was the Passaic, New Jersey, "Wonder Team" between 1919 and 1925, which won 159 consecutive games, until defeated by Hackensack, Feb. 6, 1925, by 39–35. Another colourful team was Carr Creek, Kentucky, of 1927–28, consisting of eight players, all related, who owned no uniforms, lost the state tourney finals in four overtime periods, to Ashland High (13–11), then won three contests in the national tournament in Chicago before losing to Vienna, Georgia, the national champions.

The first high schools to play basketball are believed to be Holyoke, Massachusetts, and Central High School, of Philadelphia, occasional victors over local college and town teams.

By 1970 the seasonal interscholastic basketball attendance, was approximately 125,000,000. In some communities, especially in Indiana, where the state's 580 schools draw 13,000,000 in any season and 1,500,000 to the state tournament eliminations, some communities have high school gymnasiums that seat more than the total population of the town.

Schoolboy basketball

While basketball is largely regarded as a winter sport, it is played on a 12-month basis—upon summer playgrounds, in church, municipal and industrial halls, schoolyards, family driveways, in summer camps—often on an informal basis between two or more contestants. Youngsters often fashion their own games with an undersized ball and hoops below the regulation height nailed to the side of a building or utility pole. Many grade, or grammar, schools, youth groups, municipal recreation programs, churches, and other organizations conduct basketball programs for youngsters of less than high school age. Jay Archer, of Scranton, Pennsylvania, introduced "biddy" basketball in 1950 for boys and girls under 12 years of age. The ball is smaller (28 in. in circumference) than the official ball and the baskets are lower, 8½ ft instead of 10 ft above the court, which measures 60 × 40 ft, foul line 12 ft from backboard, and there are 6-min quarters. An international tournament has been held since 1952, and the success of this program has encouraged many European countries to foster "mini" basketball in hopes of improving the quality of their games when the young future players and fans grow up. Some junior high schools and grammar schools encourage intercity and intracity games and tournaments.

Women's basketball (U.S.). Clara Baer, who introduced basketball at The H. Sophie Newcomb College for Women in New Orleans, Louisiana, had a hand in fashioning the women's style of play with her set of women's rules, published in 1895. On receiving a diagram of the court from Naismith, Miss Baer mistook dotted lines, indicating the areas wherein the players might best execute team play, to be restraining lines, with the result the forwards, centres, and guards confined their activity to specified areas. This seemed appropriate because many felt that the men's game was too strenuous for women. Women's rules over the years frequently have been modified; there are six players on a team and the court is so divided that the three forwards play in the forecourt and do all the scoring, while the three guards cover the backcourt. Senda Berenson staged the first women's college basketball game in 1893 when her freshmen and sophomore Smith College girls played against one another. In April 1895 the girls of the University of California (Berkeley) played Stanford University.

Basketball is included in many high school and college girls' athletic programs as an intramural sport. More women's colleges participate on an intercollegiate basis than ever before.

Most high school girls' interscholastic games are played preliminary to the boys' varsity games. In some states such contests are prohibited or discouraged. In Iowa the girls' tournament frequently outdraws the boys' tournament in attendance, the girls drawing 450,000 for sectional, district, and final rounds.

Outstanding teams and records

Baskin (Louisiana) High School won 218 successive games between 1947 and 1953. The famous girls high school coach, Bertha Frank Teague, of Byng High, Ada, Oklahoma, retired after the 1968–69 season with a 43-year record of 1,152 wins and 115 losses. Over one period, her teams won 96 games in a row. Her teams won eight state titles, were runners-up seven times and qualified for the state tourney on 22 occasions.

The Hanes Hosiery (North Carolina) team streak of 102 victories was terminated in 1953–54, the Wayland (Texas) Flyers rolled up a series of 131 victories in 1957–58, and Nashville Business College won the national AAU women's title in 1969 for the eighth successive year, but—after sponsoring a team for 30 years—did not

have a team the following season. An earlier famous women's club was the Edmonton Grads of Alberta, Canada, former students at the McDougall Commercial School of that city who, between 1915 and 1940, won 502 of 522 games—78 successively.

The U.S. six-player, divided forecourt and backcourt game has come to be known as the "rover" game. New five-player rules as used in the men's game and in women's international competition were adopted experimentally by the AAU in the late 1960s (see also below *International Basketball*).

International basketball. Basketball's popularity has increased manyfold since World War II through clinics sponsored by the U.S. Department of State and international exchange of teams, coaches, etc. Also, U.S. servicemen stationed abroad have aided materially.

Greatest credit for the success of the game in other lands goes to Forrest C. Allen, coach emeritus at the University of Kansas, a Naismith disciple who, as a committee of one, struggled for six years to place the game on the Olympic program. Allen was rebuffed by the U.S. Olympic Committee in his efforts to have the game demonstrated at the 1932 Olympics in Los Angeles. Undaunted, he attended the games where his efforts were encouraged by Sohaku Ri, of Waseda University, Tokyo, who raved about the success of basketball at the Far Eastern Olympics, and by Count Soyejima, president of the Japanese Basketball Association. On Oct. 19, 1933, the Berlin Organization Committee voted to have basketball introduced at the games of the XI Olympiad, scheduled for Berlin in 1936. A total of 21 nations competed in the 1936 Olympics and, most appropriately, U.S. college coaches raised sufficient funds to permit Naismith to attend the games and to toss up the first ball. Basketball has been one of the official sports in every subsequent Olympics. Prominent contenders in the Olympics have included the U.S. (undefeated from 1936 through 1968), Canada, Mexico, Poland, France, Brazil, U.S.S.R., Uruguay, Argentina, Italy, Spain, and Puerto Rico (see further ATHLETIC GAMES AND CONTESTS: *Olympic record*). Basketball has also been a fixture of the quadrennial Pan-American Games since their inauguration in 1951.

Basketball in the Olympics

The popularity of basketball upon the Olympic schedule has had a profound effect upon the growth of the game. In most nations, it ranks third in numbers both of fans and participants, behind only soccer and cycling. The international game is governed by the Fédération Internationale de Basketball Amateur (more commonly called FIBA) representing 129 nations, 31 of which are European.

In 1932 a lack of uniformity in rules caused Czechoslovakia, Portugal, and Switzerland to call a meeting at Geneva, and seven other nations—Latvia, Italy, Argentina, Greece, Hungary, Bulgaria, and Rumania—sent representatives. The body adopted the U.S. college rules of that time but since has amended them. Under international rules the court differs in that there is no front or back court, and the free-throw lanes form a modified wedge shape. There are some differences in rules governing substitutions, technical and personal fouls and free throws, intermissions and time-outs, and out-of-bounds throw-ins; a team must try for a goal within 30 sec; and dunking is legal.

Since 1935, European men's and women's championships have been held periodically (except for war years). In 1969 Czechoslovakia met Greece in the Olympic outdoor stadium in Athens before 65,000 fans. In 1953, in Moscow, the European men's championships drew 50,000 spectators. It is conservatively estimated that more than 5,000,000 men and women play basketball in Europe, at all levels. In all countries except the U.S., women play under the same rules as men.

European and World championships

In western Europe, the top division teams are sponsored by industrial firms or sports-social clubs, while in the Communist nations the game is state operated, although since 1963 government finances have been gradually decreased. Spain, Italy, France, and Belgium have a rule that permits any top division team to import one player who usually is a fine rebounder—and an American. In each nation the leagues consist of a dozen teams each, as a rule, and, at the conclusion of every season, the two lowest in standing drop into the next division, and are replaced by the top teams of the lower division. The basketball season starts in mid-October and extends into April, when the National Cup Tournament is held, with each national winner then competing in the "Cup of Cups" championship.

In the VI World championships held in Ljubljana in May 1970, the host nation Yugoslavia won, with the others finishing as follows: Brazil, U.S.S.R., Italy, United States, Czechoslovakia, Uruguay, Cuba, Panama, Canada, South Korea, Australia, United Arab Republic. Three games were played each day, each with separate admissions, and over 150,000 attended the seven-day program. Marshal Tito, watching his first basketball, inaugurated the tournament, which was televised to nine nations with an estimated audience of 50,000,000.

European nations and several members of FIBA are going more energetically into mini basketball, especially Italy, France, and Spain, with a view of encouraging the game among youngsters for future development. Such a program not only will create more and superior players, but also more fans, an advantage the United States has long enjoyed.

Professional basketball. Professional basketball appeared a few years after the game was introduced in 1891. After the YMCA's banned the sport in the mid-1890s, the various teams sought out halls they could hire and charged admission. A Trenton, New Jersey, team became the first professional club when in 1898 it hired the auditorium of the local Masonic order for $25 and was amazed to learn that there was a profit.

The play-for-pay game prospered largely in the Middle Atlantic and New England states. Trenton and the New York Wanderers were the first great professional clubs, followed by the Buffalo Germans, who started out in 1895 as 14-year-old members of the Buffalo (New York) YMCA and, with occasional new members, continued for 44 years, winning 792 out of 878 games.

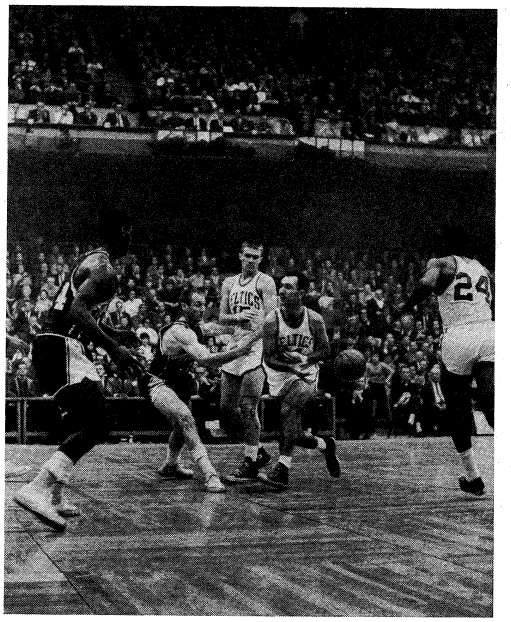
A group of basketball stylists who never received the acclaim they deserved because they wore the colours of three different towns in their heyday consisted of Edward and Lew Wachter, Jimmy Williamson, Jack Inglis (a behind-the-back dribbler), and Bill Hardman. They introduced the bounce pass and long pass as offensive weapons and championed the rule that made each player, when fouled, shoot his own free throw. Performing as Co. E of Schenectady, New York, they defeated the Kansas City Blue Diamonds in three straight games in 1905 in what was billed as the "world championship." While playing for Co. G, of Gloversville, New York, in 1907–08, this same contingent trounced the Buffalo Germans twice on the same day, and, in 1915, when they proved too strong for the Hudson River League, they engineered the first transcontinental tour in the history of the game, meeting all comers, under any set of rules, going as far west as Montana and winning all 29 games played. Upon their return, they won 10 more.

Early stylists

Before World War II, the most widely heralded professional team was the Original Celtics, which started out in 1915 as a group of youngsters from New York City, kept adding better players in the early 1920s, and became so invincible that the team disbanded in 1928, only to regroup in the early 1930s as the "New York Celtics." They finally retired in 1936. The Celtics played every night of the week, twice on Sundays, and largely on the road. During the 1922–23 season, they won 204 of 215 games.

The Original Celtics, Rens, and others

Another formidable aggregation was the New York Renaissance (the "Rens") organized by Robert Douglas in 1923, and regarded as the strongest all-Negro team of all time. During the 1925–26 campaign, they split a six-game series with the Original Celtics. During the 1932–33 season, the Rens won 88 consecutive games. In 1939 they trimmed the Harlem Globe-trotters and the Oshkosh All Stars for the world championship pro tournament in Chicago. On Feb. 16, 1935 in Kansas City, the Rens defeated the New York Celtics by one point by controlling the ball for the last six minutes.



Bob Cousy passes ball to teammate on his left as he continues to drive toward the opposition.
Sports Illustrated/Rich Clarkson © Time Inc.

The Celtics were elected to the Naismith Hall of Fame in 1959, and the Rens in 1964.

The two most successful professional coaches of all time were the Rens' Douglas, who won 2,318 games and lost 381 over a 22-year period; and Hall of Famer Frank Morgenweck, who coached for 32 years with the same team, operating in two or three different cities within the same season.

Among some of the great professional clubs were the New York Nationals; Fond du Lac, Wisconsin; East Liverpool, Ohio; the Paterson New Jersey Crescents; and the South Philadelphia Hebrew All Stars—better known as the Sphas.

The pro leagues

The first professional league was the National Basketball League (NBL) formed in 1898 to protect the players against unscrupulous promoters and to save the sport from extinction because of rough play. Their game differed from the college rules in that a netting usually separated the players from the spectators and the players bounced off the ropes like prizefighters in a ring; the ball could not go out of bounds; and the players were permitted to resume dribbling after halting. Because of the inequality of the teams or inability of the weaker clubs to meet their expenses, leagues did not survive for more than three or four seasons and invariably new circuits came into existence. As a rule, the same players performed in all these leagues, some playing for several cities or clubs within the same seasons, the better players often being teammates one night and opponents the next day.

The Depression of the 1930s hurt professional basketball and a new NBL was organized in 1938 in and around the upper Midwest. The play-for-pay ranks started to assume major league status with the organization of the new Basketball Association of America, on June 6, 1946, under the guidance of Walter A. Brown, president of the Boston Garden. Brown's contention was that professional basketball would succeed only if there was sufficient financial support to nurse the league over the early lean years; the game was conducted upon a high standard; and all players were restricted to exclusive contracts such as used in baseball, with a reserve rule protecting each team from raiding another club. Following a costly two-year war, the BAA and the NBL combined on Aug. 11, 1949, to form a 17-team National Basketball Association (NBA), which thereafter continued to attract the leading college players.

To help equalize the strength of the teams, the NBA established an annual college draft permitting each club to select a college senior in inverse order to the final standings in the previous year's competition, thus enabling the lower standing clubs to select the more talented collegians.

The game was saved from public apathy through three radical rules changes written in for the 1954–55 season: (1) a team must shoot for a basket within 24 sec after acquiring possession of the ball; (2) a bonus free throw is awarded a player anytime the opposing team commits more than six (later five, now four) personal fouls in a quarter or more than two personal fouls in an overtime period; and (3) two free throws are granted for any backcourt foul.

The leading teams in each division participate in a post-season play-off series with the eventual winner declared the world champion.

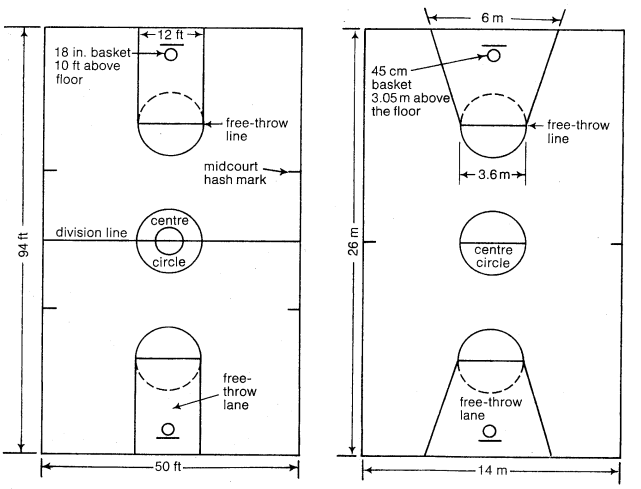
A rival 11-team American Basketball Association, with George Mikan as commissioner, was launched in the 1967–68 season and a bitter feud developed between these circuits for the top collegian talent each season.

PLAYING THE GAME

Synopsis of rules. The rules governing play of the game are based on Naismith's five principles (cited under *History* above) requiring a large, light ball, handled with the hands; no running with the ball; no player restricted from getting the ball when it is in play; no personal contact; and a horizontal, elevated goal. The rules are spelled out in specific detail by the governing bodies of the several branches of the sport and cover the playing court and equipment, officials, players, scoring and timing, fouls, violations, and other matters.

Equipment, officials, players, and scoring

The general dimensions and plans of the U.S. college court (which are essentially the same as those of U.S. high school and professional basketball courts) and the international court are shown in the accompanying diagram.



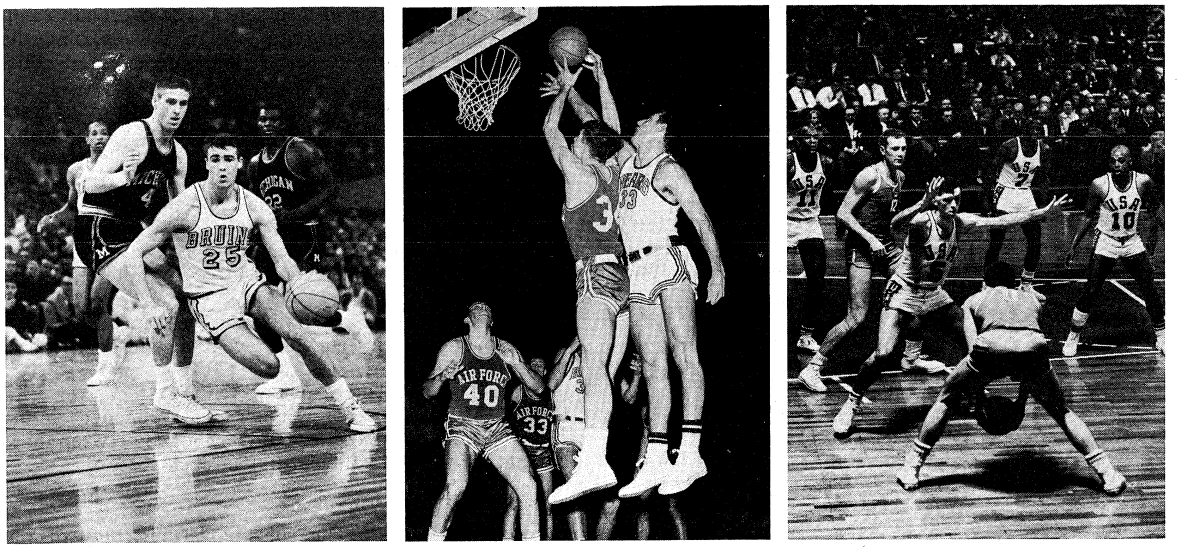
Basketball Courts.
(Left) U.S. college and high school court. (Right) International court.

The officials include a referee, an umpire, two timers, and two scorers. The International Association of Approved Basketball Officials, founded 1921, trains and certifies officials.

One player on each team acts as captain and speaks for his team on all matters involving the officials, as, for example, interpretation of rules. Specific rules cover such details as the numbering of players, the sizes of the numbers the players must wear on their jerseys, and colour of uniforms (white or light for home teams, dark for visitors). Other rules cover eligibility and related matters.

Since 1895–96 a field goal has scored 2 points, a free throw 1 point. Changes have been proposed from time to time: since 1967–68 the ABA awarded 3 points for a field goal scored from 25 ft or more from the basket.

Basketball is a noncontact game. A player may pass or



(Left) Player drives around his guard in college game. (Centre) Guard attempts to impede shot made by opponent in college game. (Right) Player guards opponent in U.S.–U.S.S.R. Olympic game, 1964.

(Left, right) *Sports Illustrated* © Time Inc., (left) Rich Clarkson, (right) Richard Meek, (centre) Ed Kirwan—Graphic Arts

Fouls and violations

bounce (dribble) the ball to secure a position whereby he or a teammate may make a try for a basket from the floor. A foul is committed whenever a player makes such contact with an opponent so as to put him at a distinct disadvantage. It provides the offended player with a free throw, an unhindered throw for a goal from behind the foul line. Under National Basketball Committee (NBC) rules, if the offended player was fouled while in the act of shooting, if his toss was successful the basket counts and he is awarded one free throw; if he missed his shot, he is awarded two free throws. Unsportsmanlike conduct by a player or coach incurs a technical foul and possible banishment.

Under NBC rules violation occurs when a player (with the ball) takes an excessive number of steps, or slides; causes the ball to go out-of-bounds; steps over the foul line while tossing for a free throw; steps over the end line or sideline while tossing the ball in to a teammate, or fails to pass the ball in within 5 sec; runs with, kicks, or strikes the ball with his fist; dribbles a second time after having once concluded his dribble; remains more than 3 sec in his free-throw lane while he or his team has the ball; causes the ball to go into the backcourt or retains the ball in the backcourt more than 10 sec. Penalty is loss of the ball—opponents throw in the ball from the side.

Basketball terms and definitions. *Blocking.* Any illegal personal contact that impedes the progress of an opponent who does not have the ball.

Dribble. Ball movement by a player who taps the ball in the air or on the floor and then touches (bounces) it one or more times. A dribble ends when a player touches the ball with both hands simultaneously or loses contact with it.

Held ball. Called when two opponents have one or two hands so firmly upon the ball that neither can gain possession without undue roughness. It also is called when a player in the front court is so closely guarded that he cannot pass or try for a goal or is obviously withholding the ball from play.

Jump ball. A method of putting the ball into play. The referee tosses it up between two opponents who try to tap it to a teammate.

Pass. Throwing, batting, or rolling the ball to another player. The main types are: (1) the chest pass—ball is released from a position in front of the chest; (2) the bounce pass—the ball is bounced on the floor to get it past a defensive opponent; (3) the roll pass on the floor; (4) the hook pass (side or overhead); and (5) the baseball pass—the ball is thrown a longer distance with one hand in a manner similar to a baseball throw.

Pick. See *Screen.*

Pivot. A movement in which a player with the ball steps once or more in any direction with the same foot while the other foot (pivot foot) is kept at its point of contact with the floor.

Rebounding. Both teams attempting to gain possession of the ball after any try for a basket is unsuccessful but the ball does not go out-of-bounds and remains in play.

Screen. Legal action of a player who, without causing contact, delays or prevents an opponent from reaching his desired position. A pick.

Shots from the field. The main field shots are: the lay-up (the shooter, close to the basket, jumps and lays the ball against the backboard so it will rebound into the basket) and the over-the-rim shot (the player jumps and shoots or lays the ball over the rim). From a distance of up to 24 ft from the basket many players use a one-hand push shot from a stride, jump, or standing position, and a hook shot, which is overhead. From longer range, so-called set shots may be the two-hand shot starting from a chest position or the one-hand shot.

Travelling (running with the ball). Progressing in any direction in excess of the prescribed limits while holding the ball.

Turnover. Loss of possession of the ball by a team before any member has been able to try for a basket.

Principles of play. Each team of five players consists of two forwards, two guards, and a centre, usually the tallest man on the team. At the beginning of each period the ball is put into play by a jump ball at centre court; *i.e.*, the referee tosses the ball up between the opposing centres, higher than either can jump, and when it descends each tries to tap it to one of his teammates, who must remain outside the centre circle until the ball is tapped. Jump balls are also held when opposing players share possession of the ball (held ball), or cause it to go out-of-bounds at the same time. After each successful basket (goal or field goal) the ball is put back in play by the team that is scored on by one player passing the ball in from behind the end line where the score was made. The ball is put in play in the same manner after a successful free throw; if two have been awarded, after the second if it is successful. After violations (see above) the ball is awarded to the opposing team to be passed in from out-of-bounds from a point designated by an official.

A player in possession of the ball must pass or shoot before taking two steps or must start dribbling before taking his second step. The dribble stops if the player permits the ball to come to rest or when he touches it with both hands at once, when he must stop or pass the ball. The ball may be tapped or batted with the hands, passed, bounced, or rolled in any direction.

Systems of offense

As basketball progressed, various coaches and players devised intricate plays and offensive manoeuvres. Some systems emphasize speed, deft ball handling, and high scores; others stress ball control, slower patterned movement, and low scores. A strategy based on speed is called fast break. When fast-break players recover possession of the ball in their backcourt, as by getting the rebound from an opponent's missed shot, they race to beat the defense to its backcourt by a combination of speed and passing and try to make a field goal before the opponents have time to set up a defense. The fast break and its many variants are sometimes described as the racehorse style of play.

Some teams, following an overall game plan or when they do not have the opportunity for a fast break, employ a more deliberate style of offense. The guards bring the ball down the court toward the basket carefully and maintain possession of the ball in the front court, passing, dribbling, and screening opponents in an effort to set up a play that will free a player for an open shot. Generally, set patterns of offense use one or two pivot men who play near the free-throw area at the low post positions (between the free-throw line and the end line) or at high post positions (between the free-throw line and the basket). The pivot, or post, men are usually the taller men on the team and are in position to receive passes, pass to teammates, shoot, screen for teammates, and tip-in or rebound (recover) missed shots. All of the players are constantly on the move executing the patterns designed to give one player a favourable shot—and at the same time place one or more teammates in good position to tip-in or rebound if he misses.

Systems of defense

Systems of defense also have developed over the years. One of the major strategies is known as man-to-man, or man-for-man. Each player has a specific opponent whom he guards whenever he moves, except when he "switches" with a teammate when he is screened, or in order to guard another player in a more threatening scoring position. Another major strategy is the zone or five-man defense. In this system each player has a specific area to guard irrespective of which opponent plays in that area. The zone is designed to keep the offense from driving in to the basket and force them to take long shots, while preventing tip-ins and controlling rebounds from the backboard. A great many variations and combinations have been devised to employ various aspects of both man-to-man or zone defense strategies. The press, which can be man-to-man or zone, exists when a team guards an opponent so thoroughly that it forces the opposition to hurry its movements, especially to commit errors. Well-coached teams can modify both their offensive and defensive strategies according to the shifting circumstances of the game and in response to their opponents' particular strengths and weaknesses and styles of play.

BIBLIOGRAPHY

General references: JAMES NAISMITH, *Basketball: Its Origin and Development* (1941), an autobiography and explanation of how the game's inventor came to prepare the original 13 rules; W.G. MOKRAY (ed.), *Encyclopedia of Basketball* (1963), complete lists of college and professional records, All America teams, Hall of Fame electees, with brief documented resumes of basketball upon all levels; ROBERT BRUCE, *Annotated Bibliography of Basketball Literature* (1947), the best reference available for books and outstanding magazine articles published before 1947.

Texts on coaching and how to play: BOB COUSY and FRANK POWER, *Basketball: Concepts and Techniques* (1970), a well-written text by an outstanding player and his former assistant coach; JOHN WOODEN, *Practical Modern Basketball* (1966), a sound book by the most successful college coach in history; CLAIR BEE and KENNETH NORTON, *Science of Coaching* (*Basketball Series*), 2nd ed. (1959), on every phase of basketball; F.C. ALLEN, *Better Basketball: Techniques, Tactics and Tales* (1937), on a famous coach's knowledge and psychological handling of players and opponents; A.M. WEYAND, *Cavalcade of Basketball* (1960), a well-written, well-researched reference that relates many phases of the game, items of which are not found elsewhere; ADOLPH RUPP, *Championship Basketball* (1948), an explanation of how Rupp won more college games than any other basketball coach.

Women's basketball: HELEN M. LAWRENCE and G.I. FOX, *Basketball for Girls and Women* (1954), a history of women's basketball, showing how the various rule changes came about and how to teach the sport; BERTHA FRANK TEAGUE, *Basketball for Girls* (1962), by an outstanding women's high school coach.

Yearbooks, guides, and handbooks: Converse Basketball Yearbook (annual), the season's review, vital statistics on all teams, players, and coaches, and special stories not found elsewhere; *NCAA Basketball Guide* (annual), *NBA Official Guide* (annual), *American Basketball Association Guide* (annual), *Fédération Internationale de Basketball Amateur Handbook*, *American Association for Health and Physical Education* (*AAHPER*) *Women's Basketball Guide* (annual), *Amateur Athletic Federation Handbook*, good reference books for those who wish to keep tabs on the records of the teams and players.

(W.G.M.)

Basketry

Though it would appear that basketry might best be defined as the art or craft of making baskets, the fact is that the term is one of those the limits of which seem increasingly imprecise the more one tries to grasp it. The category basket may include receptacles made of interwoven, rather rigid material, but it may also include pliant sacks made of a mesh indistinguishable from netting—or garments or pieces of furniture made of the same materials and using the same processes as classical basketmaking. In fact, neither function nor appearance nor material nor mode of construction are of themselves sufficient to delimit the field of what common sense nevertheless recognizes as basketry.

In this article the term will be taken to mean a handmade assemblage of vegetable fibers that are relatively large and rigid, so as to make a continuous surface, usually (but not exclusively) a receptacle. The consistency of the materials used distinguishes basketry, which is handmade, from weaving, in which the flexibility of the threads requires the use of an apparatus to put tension on the warp, the lengthwise threads. What basketry has in common with weaving is that both are means of assembling separate fibres by twisting them together in various ways.

MATERIALS

There is no region in the world, except in the northernmost and southernmost parts, where man does not have at his disposal materials, such as twigs, roots, canes, and grasses, that lend themselves to basketry.

The variety and quality of materials available in a particular region has a bearing on the relative importance of basketry in a culture and on the types of basketry produced. Rainy, tropical zones, for example, have palms and large leaves that require plaiting techniques different from those required for the grass stalks that predominate in the dry, subtropical savanna regions or for the roots and stalks found in cold temperate zones. The interrelationship between materials and methods of construction might in part explain why the principal types of basketry are distributed in large areas that perhaps correspond to climatic zones as much as to cultural groups: the predominance of sewed coiling, for example, in the African savannas, in the arid zones of southern Eurasia and of North America; of spiral coiling and twining in temperate regions; and of various forms of plaiting in hot regions. There is also a connection between the materials used and the function of the basket, which determines whether rigid or soft materials—either as found in nature or specially prepared—are used. In the Far East, for example, twined basketry made of thin, narrow strips (laths) of bamboo is effective for such objects as cages and fish traps that require solid partitions with openings at regular intervals. Often soft and rigid fibres are used together: rigid to provide the shape of the object and soft to act as a binder to hold the shape.

Aesthetic goals and choice of material

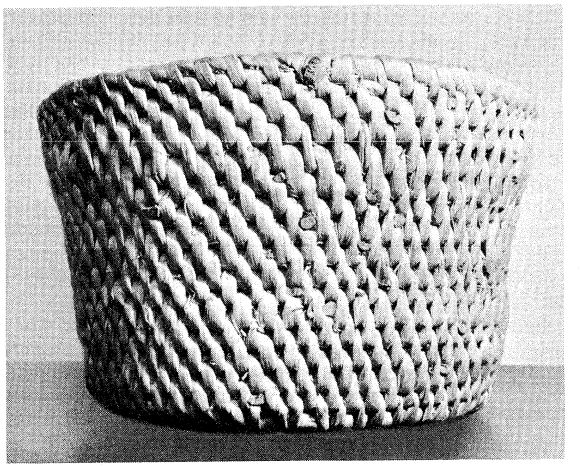
Finally, materials are chosen with a view toward achieving certain aesthetic goals; conversely, aesthetic goals are limited by the materials available. The effects most commonly sought are delicacy and regularity of the

threads; a smooth, glossy surface or a dull, rough surface; and colour, whether natural or dyed. Striking effects can be achieved from the contrast between threads that are light and dark, broad and narrow, dull and shiny —contrasts that complement either the regularity or the decorative motifs obtained by the intricate work of plaiting.

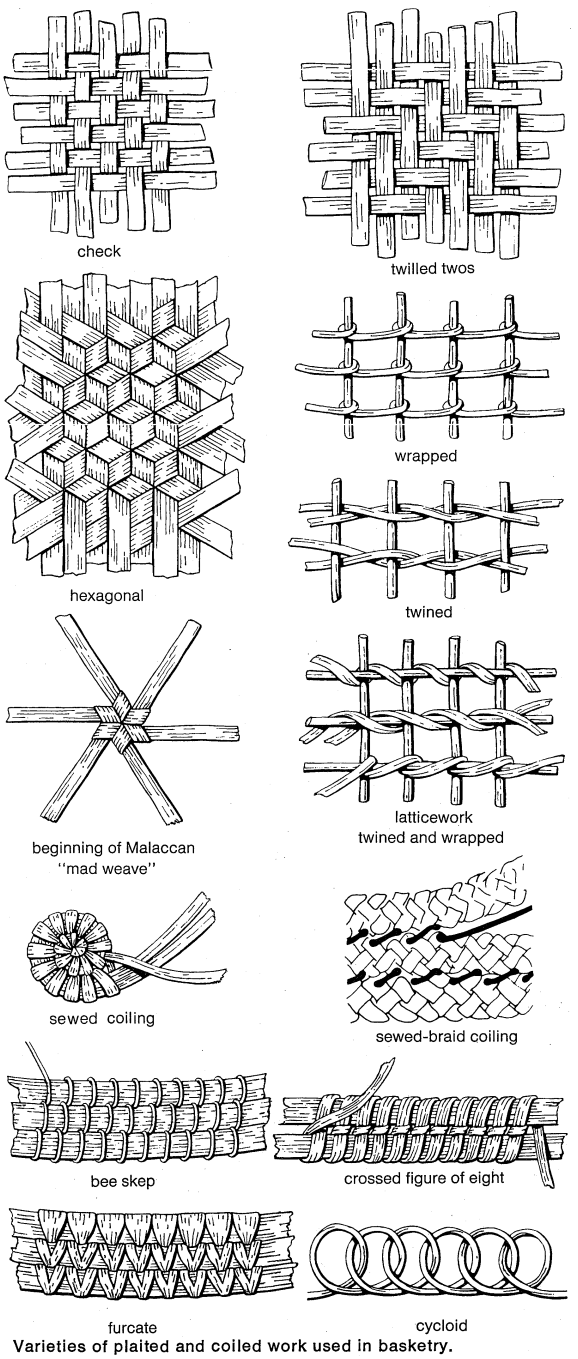
TECHNIQUES

Despite an appearance of almost infinite variety, the techniques of basketry can be grouped into several general types according to how the elements making up the foundation (the standards, which are analogous to the warp of cloth) are arranged and how the moving element (the thread) holds the standards by intertwining among them.

Coiled construction. The distinctive feature of this type is its foundation, which is made up of a single element, or standard, wound in a spiral around itself. The coils are kept in place by the thread, the work being done stitch by stitch and coil by coil. Variations within this type are defined by the method of sewing, as well as by the nature of the coil, which largely determines the type of stitch.



Spiral-coiled basket with twill effect, from Bialystok region, Poland. In the Musée de l'Homme, Paris.
By courtesy of H. Balfet



Varieties of plaited and coiled work used in basketry.

Spiral coiling. The most common form is spiral coiling, in which the nature of the standard introduces two main subvariations: when it is solid, made up of a single whole stem, the thread must squeeze the two coils together binding each to the preceding one (giving a diagonal, or twilled, effect); with a double or triple standard the thread catches in each stitch one of the standards of the preceding coil. Many other variations of spiral coiling are possible. Distribution of this type of basketry construction extends in a band across northern Eurasia and into northwest North America; it is also found in the southern Pacific region (China and Melanesia) and, infrequently, in Africa (Rhodesia).

Sewed coiling. Sewed coiling has a foundation of multiple elements—a bundle of fine fibres. Sewing is done with a needle or an awl, which binds each coil to the preceding one by piercing it through with the thread. The appearance varies according to whether the thread conceals the foundation or not (bee-skep variety) or goes through the centre of the corresponding stitch on the preceding coil (split stitch, or furcate). This sewed type of coiled ware has a very wide distribution: it is almost the exclusive form in many regions of North and West Africa; it existed in ancient Egypt and occurs today in Arabia and throughout the Mediterranean basin as far as western Europe; it also occurs in North America, in India, and sporadically in the Asiatic Pacific. A variety of sewed coiling, made from a long braid sewed in a spiral, has been found throughout North Africa since ancient Egyptian times.

Distribution of sewed coiling

Half-hitch and knotted coiling. In half-hitch coiling, the thread forms half hitches (simple knots) holding the coils in place, the standard serving only as a support. There is a relationship between half-hitch coiling and the half-hitch net (without a foundation), the distribution of which is much more extensive. The half-hitch type of basketry appears to be limited to Australia, Tasmania, Tierra del Fuego in South America, and Pygmy territory in Africa. In knotted coiling, the thread forms knots around two successive rows of standards; many varieties can be noted in the Congo, in Indonesia, and among the Basket Makers, an ancient culture of the plateau area of southwestern United States, centred in parts of Arizona, New Mexico, Colorado, and Utah.

The half-hitch and knotted-coiling types of basketry each have a single element variety in which there is no foundation, the thread forming a spiral by itself analogous to the movement of the foundation in the usual type. An openwork variety of the single element half hitch (called cycloid coiling) comes from the Malay area; and knotted single-element basketry, from Tierra del Fuego and New Guinea.



Sewed-coiling basket in the tarantula design, Papago culture, Arizona. In the collection of A.E. Robinson.
By courtesy of Mrs. A.E. Robinson

Noncoiled construction. Compared to the coiled techniques, all other types of basketry have a certain unity of construction: the standards form a foundation that is set up when the work is begun and that predetermines the shape and dimensions of the finished article. Nevertheless, if one considers the part played by the standards and the threads, respectively, most noncoiled basketry can be divided into three main groups.

Wattle construction. A single layer of rigid, passive, parallel standards is held together by flexible threads in one of three ways, each representing a different subtype. (1) The bound, or wrapped, type, which is not very elaborate, has a widespread distribution, being used for burden baskets in the Andaman Islands in the Bay of Bengal, for poultry cages in different parts of Africa and the Near East, and for small crude baskets in Tierra del Fuego. (2) In the twined type, the threads are twisted in two or threes, two or three strands twining around the standards and enclosing them. The twining may be close or openwork or may combine tight standards and spaced threads. Close twining mainly occurs in three zones: Central Africa, Australia, and western North America, where there are a number of variations such as twilled and braided twining and zigzag or honeycomb twining. The openwork subtype is found almost universally because it provides a perfect solution to the problem of maintaining rigid standards with even spacing for fish traps and hurdles (portable panels used for enclosing land or livestock). Using spaced threads, this subtype is also used for flexible basketry among the Ainu of northern Japan and the Kuril Islands and sporadically throughout the northern Pacific. (3) The woven type, sometimes termed wickerwork, is made of stiff standards interwoven with flexible threads. It is the type most commonly found in European and African basketry and is found sporadically in North and South America and in Near and Far Eastern Asia.

Lattice construction. In lattice construction a frame made of two or three layers of passive standards is bound together by wrapping the intersections with a thread. The ways of intertwining hardly vary at all and the commonest is also the simplest: the threads are wrapped in a spiral around two layers of standards. This method is widely used throughout the world in making strong, fairly rigid objects for daily use: partitions for dwellings, baskets to be carried on the back, cages, and fish traps (with a Mediterranean variety composed of three layers of standards and a knotted thread). The same method, moreover, can be adapted for decorative purposes, with threads—often of different colours—to form a variety of motifs similar to embroidery. This kind of lattice construction appears mainly among the Makah Indians of the U.S. Pacific Northwest and in Central and East Africa.

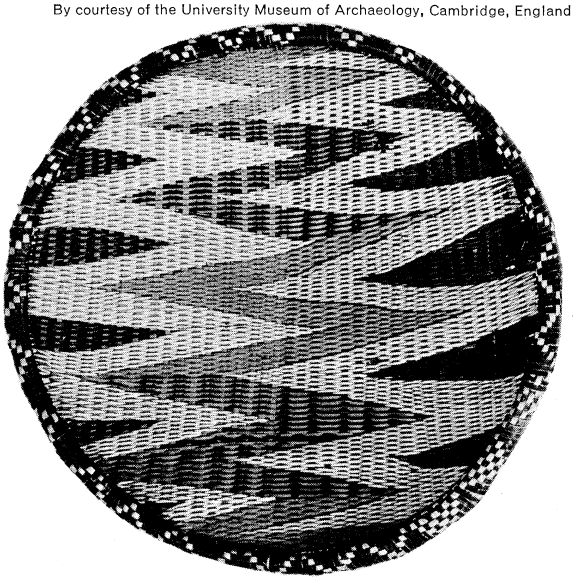
Uses of lattice construction

Matting or plaited construction. Standards and threads are indistinguishable in matting or plaited construction; they are either parallel and perpendicular to the edge (straight basketry) or oblique (diagonal basketry). Such basketry is closest to textile weaving. The materials used are almost always woven, using the whole gamut of weaving techniques (check, twill, satin, and innumerable decorative combinations). Depending on the material and on the technique used, this type of construction lends itself to a wide variety of forms, in particular to the finest tiny boxes and to the most artistic large plane surfaces. It is widely distributed but seems particularly well adapted to the natural resources and to the kind of life found in intertropical areas. The regions where it is most common are different from, and complementary with, those specializing in coiled and twined ware; that is, eastern and southeastern Asia (from Japan to Malaysia and Indonesia), tropical America, and the island of Madagascar off the east coast of Africa.

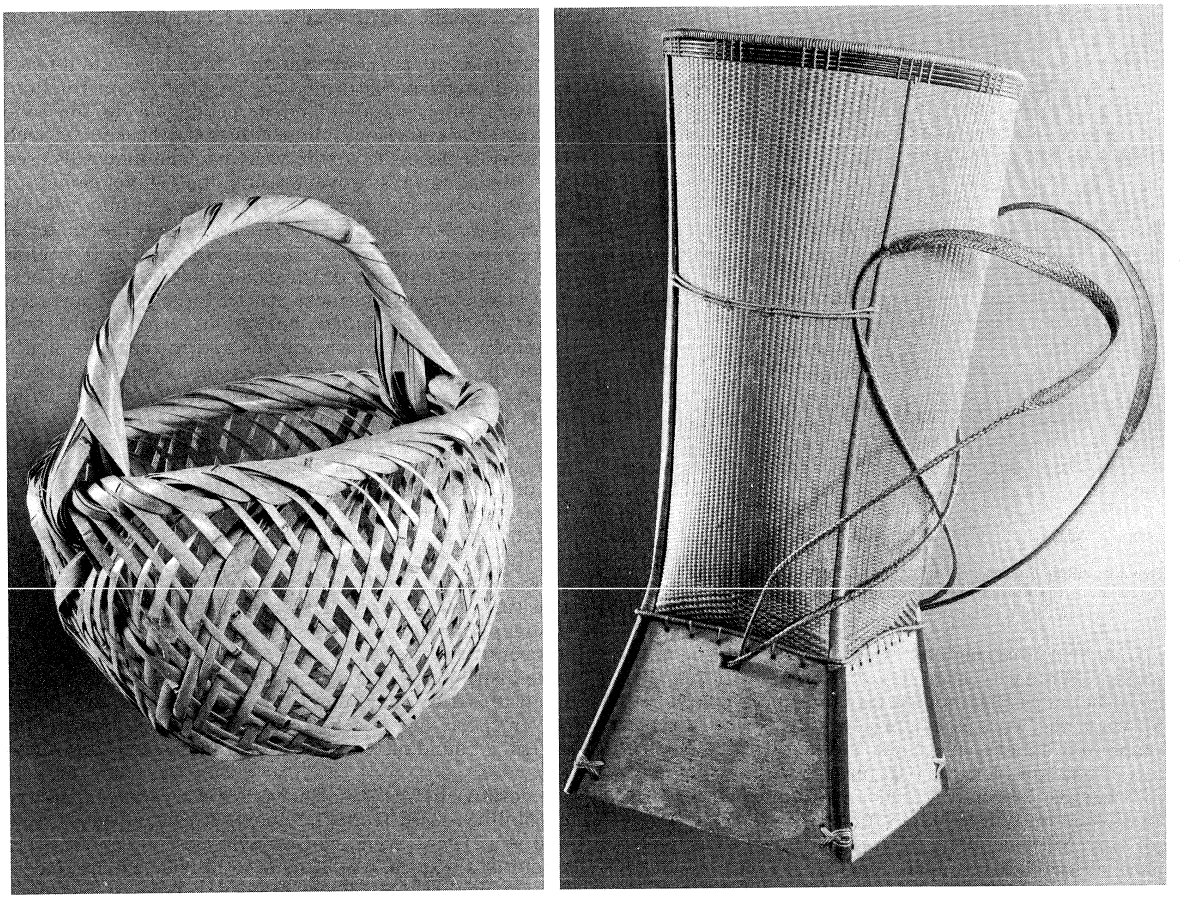
One variety of matting or plaited work consists of three or four layers of elements, which are in some cases completely woven and in others form an intermediate stage between woven and lattice basketry. The intermediate type (with two layered elements, one woven) is known as hexagonal openwork and is the technique most common in openwork basketry using flat elements. It has a very wide distribution: from Europe to Japan, southern Asia, Central Africa, and the tropical Americas. A closely woven fabric in three layers, forming a six-pointed star design, is found on a small scale in Indonesia and Malaysia.

Decorative devices. Clearly, a variety of decorative possibilities arise from the actual work of constructing basketry. These, combined with the possible contrasts of colour and texture, would seem to provide extensive decorative possibilities. Each particular type of basketry, however, imposes certain limitations, which may lead to convergent effects: hexagonal openwork, for example, forms the same pattern the world over, just as twilled weaving forms the same chevrons (vertical or horizontal). Each type, also, allows a certain range of freedom in the decoration within the basic restrictions imposed by the rigidity of the interlaced threads, which tends to impose geometric designs or at least to geometrize the motifs. In general, the two main types of basketry—plaited and coiled—lend themselves to two different kinds of decoration. Coiled basketry lends itself to radiating designs, generally star- or flower-shaped compositions or whirling designs sweeping from the centre to the outer edge. Plaited basketry, whether diagonal or straight,

By courtesy of the University Museum of Archaeology, Cambridge, England



Double-thick wickerwoven tray, from Ruanda-Urundi, Africa. In the University Museum of Archaeology, Cambridge, England.



Plaiting.
(Left) Bamboo flower basket of diagonal openwork plaiting, from Japan. (Right) Burden basket of straight-woven plaiting, from Vietnam. In the Musée de l'Homme, Paris.
By courtesy of the Musée de l'Homme, Paris

Typical coiled and plaited basketry decorations

lends itself to over-all compositions of horizontal stripes and, in the detail, to intertwined shapes that result from the way two series of threads, usually in contrasting colours, appear alternately on the surface of the basket.

Other art forms have been influenced ornamentally by basketry's plaited shapes and characteristic motifs. Because of their intrinsic decorative value—and not because the medium dictates it—these shapes and motifs have been reproduced in such materials as wood, metal, and clay. Some notable examples are the interlacing decorations carved on wood in the Central African Congo; basketry motifs engraved into metalwork and set off with inlayed silver by Frankish artisans in the Merovingian period (6th to 8th centuries); and osier patterns (molded basketwork designs) developed in 18th-century Europe to decorate porcelain.

SYMBOLISM

The Babylonian god Marduk "plaited a wicker hurdle on the surface of the waters. He created dust and spread it on the hurdle." Thus ancient Mesopotamian myth describes the creation of the earth using a reed mat. Many other creation myths place basketry among the first of the arts given to man. The Dogon of West Africa tell how their first ancestor received a square-bottomed basket with a round mouth like those still used there in the 20th century. This basket, upended, served him as a model on which to erect a world system with a circular base representing the sun and a square terrace representing the sky.

Like the decorative motifs of any other art form, the geometric, stylized shapes may represent natural or supernatural objects, such as the snakes and pigeon eyes of Borneo, and the kachina (deified ancestral spirit), clouds, and rainbows of the Hopi Indians of Arizona. However, the fact that these motifs are given a name does not always mean they have symbolic significance or express religious ideas.

Sometimes symbolism is associated with the basket itself. Among the Guayaki Indians of eastern Paraguay, for example, it is identified with the female: the men are hunters, the women are bearers as they wander through the forest; when a woman dies, her last burden basket is ritually burned and thus dies with her.

USES

As domestic utensils. Household basketry objects consist primarily of receptacles for preparing and serving food and vary widely in dimension, shape, and watertightness. Baskets are used the world over for serving dry food, such as fruit and bread, and they are also used as plates and bowls. Sometimes—if made waterproof by a special coating or by particularly close plaiting—they are used as containers for liquids. Such receptacles are found in various parts of Europe and Africa (Chad, Rwanda, Ethiopia) and among several groups of North American Indians. By dropping hot stones into the liquid, the Hupa Indians of northwestern California even boil water or food in baskets.

Waterproofing

Openwork, which is permeable and can be made with mesh of various sizes, is used for such utensils as sieves, strainers, and filters. Such basketry objects are used in the most primitive cultures as well as in the most modern (the tea strainers used in Japan, for example). The flexibility of work done on the diagonal is put to particularly ingenious use by the Africans in beer making and, above all, by Amazonian Indians in extracting the toxic juices from manioc pulp (a long basketwork cylinder is pulled down at the bottom by ballasting and, as it gets longer, compresses the pulp with which it had previously been filled).

Finally, basketry plays an important part as storage containers. For personal possessions, there are baskets, boxes, and cases of all kinds—nested boxes from Madagascar, for example, which are made in a graduated

series so that they fit snugly one within another, or caskets with multiple compartments from Indonesia. For provisions, there are baskets in various sizes that can be hung up out of the reach of predators, and there are baskets so large that they are used as granaries. In The Sudan in Africa, as in southern Europe, these are usually raised off the ground on a platform and sheltered by a large roof or stored in the house, particularly in Mediterranean regions; for preserving cereals they are sometimes caulked with clay.

In house and furniture construction. Some of these granaries are not far from being houses. Basketry used in house construction, however, usually consists of separately made elements that are later assembled; partitions of varying degrees of rigidity used as walls or to fence in an enclosure; roofs made of great basketry cones (in Chad, for example); and, above all, mats, which have numerous uses in the actual construction as well as in the equipping of a house. Probably the oldest evidence of basketry is the mud impressions of woven mats that covered the floors of houses in the Neolithic (*c.* 7000 BC) village of Jarmo in northern Iraq. Mats were used in ancient Egypt to cover floors and walls and were also rolled up and unrolled in front of doorways, as is shown by stone replicas decorating the doorways of tombs dating from the Old Kingdom, *c.* 2686–2160 BC. It is known from paintings that they were made of palm leaves and were decorated with polychrome (multicoloured) stripes, much like the mats found in Africa and the Near East. Two notable examples of modern mats are the pliant ones, made of pandanus leaves, found in southern Asia and Oceania and the tatami, which provide the unit of measurement of the surface area of Japanese dwellings.

Woven mats

Just as basketry has been used for making containers and mats, so from ancient times to modern it has been used for making such pieces of furniture as cradles, beds, tables, and various kinds of seats and cabinets.

In dress and ornament. In addition to the use of basketry for skirts and loincloths (particularly common in Oceania), supple diagonal plaiting has even been used to make dresses (Madagascar). Plaited raincoats exist throughout eastern Asia as well as Portugal. Basketry most frequently is used for shoes (particularly sandals, some of which come close to covering the foot and are plaited in various materials), and, of course, for hats—the conical hat particularly common in eastern Asia, for example, and the skullcaps and brimmed hats found in Africa, the Americas, and much of Europe.

To protect head and body against weapons, thick, strong basketry has been used in the form of helmets (Africa, the Assam region in India, and Hawaii); armour (for example, armour of coconut palm fibre for protection against weapons made of sharks' teeth by the Micronesia inhabitants of the Gilbert Islands); and shields, for which basketry is eminently suitable because of its lightness.

In addition to clothes themselves, there are numerous basketry accessories: small purses, combs, headdresses, necklaces, bracelets, and anklets. In West Africa there are even chains made of fine links and pendants plaited in a beautiful, bright yellow straw in imitation of gold jewelry. Many objects are plaited just for decoration or amusement such as ornaments like those used for Christmas trees or for harvest festivals and scale models and little animal or human figurines that sometimes serve as children's toys.

As ritual objects. There is often no very clear distinction between accessories and ritual ornaments, as in the ephemeral headdresses made for initiation rites by the young Masa people in the Cameroon; dance accessories; ornaments for masks, such as the leaf masks that the Bobo of Upper Volta make with materials from the bush.

More clearly ritual in nature are the palms (woven into elaborate geometric shapes and liturgical symbols) carried in processions on Palm Sunday by Christians in various Mediterranean regions; some, like those from Elche in Spain, are over six feet (nearly two metres) high and take days to make. In Bali an infinite variety of plaiting techniques are involved in the preparation of ritual offerings, which is a permanent occupation for the women, a hundred of whom may work for a month or two preparing for certain great festivals.

In hunting and fishing. Baskets are used throughout the world as snares and fish traps, which allow the catch to enter but not to leave. They are often used in conjunction with a corral (on land) or a weir (an enclosure set in the water), which are themselves made either of pliable nets or panels of basketry. In Africa as well as in eastern Asia a basketry object is used for fishing in shallow water; open at top and bottom, this object is deposited sharply on the bottom of shallow rivers or ponds, and, when a fish is trapped, it is retrieved by putting a hand in through the opening at the top.

Baskets as snares

In harvesting. Basketry is also used in harvesting foodstuffs; for example, in the form of winnowing trays (from whose French name, *van*, the French word for basketry, *vannerie*, is derived). A curious basket, found in the Sahel region south of the Sahara, is swung among wild grasses and in knocking against the stalks collects the grain.

As a means of transport. Baskets are used as transport receptacles; they are made easier to carry by the addition of one or two handles or straps depending on whether the basket is carried by hand, on a yoke, or on the back. The supple two-handled palm-leaf basket, common in North Africa and the Near East, existed long ago in ancient Mesopotamia; in Europe and eastern Asia, the one-handled basket, which comes in a variety of shapes, sizes, and types of plaiting, is common; in Africa, however, where burdens are generally carried on the head, there is no difference between baskets used for transporting goods and those used for storing or preserving.

Burden baskets are large, deep baskets in which heavy loads can be carried on the back; they are provided either with a headband that goes across the forehead (especially American Indian, southern Asia), or with two straps that go over the shoulders (especially in Southeast Asia and Indonesia).

There are three fairly spectacular types of small basketry craft found in regions as far apart as Peru, Ireland, and Mesopotamia: the balsa (boats) of Lake Titicaca, made of reeds and sometimes fitted out with a sail also made of matting; the British coracle, the basketry framework of which is covered with a skin sewn onto the edge; and the gufa of the Tigris, which is round like the coracle and made of plaited reeds caulked with bitumen.

ORIGINS AND CENTRES OF DEVELOPMENT

Prehistoric basketry. Something about the prehistoric origins of basketry can be assumed from archaeological evidence. The evidence that does exist from Neolithic times onward has been preserved because of conditions of extreme dryness (Egypt, Peru, southern Spain) or extreme humidity (peat bogs in northern Europe, lake dwellings in Switzerland); because it had been buried in volcanic ash (Oregon); or because, like the mats at Jarmo, it left impressions in the mud or on a pottery base that had originally been molded onto a basketry foundation. More recently, when written and pictorial documentation is available, an activity as humble and banal as basketry is not systematically described but appears only by chance in narratives, inventories, or pictures in which basketry objects figure as accessories.

On the evidence available, researchers have concluded that the salient characteristics of basketry are the same today as they were before the 3rd millenium BC. Then, as now, there was a wide variety of types (and a wide distribution of most types): coiled basketry either spiral or sewed, including furcate and sewed braid (mainly in Europe and the Near East as far as the Indus valley); wattlework with twined threads (America, Europe, Egypt) and with woven threads (Jarmo, Peru, Egypt); and plaited construction with twilled weaving (Palestine, Europe).

To list the centres of production would almost be to list all human cultural groups. Some regions, however, stand out for the emphasis their inhabitants place on basketry or for the excellence of workmanship there.

American Indian basketry. In western North America the art of basketry has attained one of its highest peaks of

perfection and has occupied a pre-eminent place in the equipment of all the groups who practice it. The American Indians are particularly noted for their twined and coiled work. The Chilkat and the Tlingit of the Pacific Northwest are known for the extreme delicacy of their twined basketry; the California Indians, for the excellence of their work with both types; and the Apache, Moki, and other Pueblo Indians of the southwestern interior of the U.S., for coiled basketry remarkable for the bold composition of its decoration and for the delicacy with which it is carried out.

Central and South American basketry is similar in materials and plaiting processes. The notable difference lies in the finishes used, and here the Guyana Indians of northeastern South America excel, being among the best basketry makers in the world and using a technique of fine plaiting with a twill pattern.

Importance of basketry in Oceania

Oceanic basketry. Various plaiting processes have been highly developed in Oceania, not just for making utilitarian articles but also for ceremonial and prestige items, such as very finely twined cloaks in New Zealand, statues in Polynesia, masks in New Guinea, and decorated shields in the Solomon Islands. In Oceania, as in southern Asia, one can speak of a vegetal civilization, in which basketry predominates over such arts as metalwork and pottery. Particular mention should be made of the Sakai of the Malay Peninsula and of the Australian aborigines, whose meagre equipment includes very delicate basketry done by the women. The Sakai use various plaiting techniques and the Australians tight twining.

By courtesy of the Musée de l'Homme, Paris



Basketry mask, from the Sepik region of New Guinea. In the Musée de l'Homme, Paris.

African basketry. Africa presents an almost infinite variety of basket types and uses. In such regions as Chad and the Cameroon basketry is everywhere in evidence—edging the roads, roofing the houses, decorating the people, and providing the greater part of domestic equipment. The very delicate twill plaited baskets of Congo are notable for their clever patterning. In the central and eastern Sudanese zone the rich decorative effect of the sewed coiled baskets is derived from the interplay of colours. The Great Rift Valley lake area produces coiled and twined basketry the elegance of which results from restrained decoration and careful finishing.

East Asian basketry. The temperate zone of East Asia produces a variety of work. Bamboo occupies a particularly important place both in ordinary basketry equipment and in objects with a primarily aesthetic function (Japanese flower baskets, for example). The production of decorative objects is one feature that distinguishes East Asian basketry from Near Eastern and African, which is primarily utilitarian. Southeast Asia must be mentioned, together with Indonesia and Madagascar, as among the first places for fine decorative plaiting techniques.

European basketry. In Europe almost the whole range of basketry techniques is used, chiefly in making utilitarian objects (receptacles for domestic and carrying purposes and household furniture) but also in making objects primarily for decorative use.

Modern basketry. Even in the modern hyperindustrial world, there seems to be a future for basketry in two directions. Because of its flexibility, lightness, permeability and solidity, it will probably remain unsurpassed for some utilitarian ends; such articles, however, because they are entirely handmade, will gradually become luxury items. As a folk art, on the other hand, it has the advantage of needing no investment, since the essential requirements have been available to anyone for at least 50,000 years: a simple awl, nimble fingers, and patience.

BIBLIOGRAPHY. H.H. BOBART, *Basketwork Through the Ages* (1936); M.L. LEE, *Basketry and Related Arts* (1948); G.M. CROWFOOT, "Textiles, Basketry and Mats," in *A History of Technology*, vol. 1 (1954); GEOFFREY H.S. BUSHNELL, "Basketry," *Encyclopedia of World Art*, vol. 2, col. 387–400 (1960); H. BALFET, "La Vannerie, essai de classification," *L'Anthropologie* (1952; Eng. trans. and preface by M.A. BAUMHOFF, "Basketry: A Proposed Classification," in *Papers on Californian Archaeology*, no. 47–49, pp. 1–21, 1957), a detailed explanation of the classifications used in this article and the source of portions adapted for use here with permission of the Archaeological Survey, University of California, Berkeley; O.T. MASON, "Aboriginal American Basketry," *Report* of the U.S. National Museum, pp. 171–548 (1904), the classic work on basketry; H. MUNSTERBERG, *The Folk Arts of Japan*, ch. 3 (1958); A.E. ROBINSON, The Basket Weavers of Arizona (1954), traditional and modern basketmaking in Arizona; *La Charpaigne* by C. DE FRANCE (Comité du Film Ethnographique, Musée de l'Homme, Paris), a film on the process of making a traditional type of basket.

(He.Ba.)

Basque Language

Basque, the only remnant of the languages spoken in southwestern Europe before the region was romanized, is currently used in a narrow area of approximately 10,000 square kilometres (3,900 square miles) in Spain and France. The number of Basque-speaking people outside of that territory, in Europe and in the Americas, however, is far from insignificant. In Spain the Basque-speaking region comprises the province of Guipúzcoa, parts of Biscay and Navarra, and a corner of Álava, and in France the western region of the *département* of Pyrénées Altantiques. Although no statistics are available, the number of speakers, who are largely bilingual, might be judiciously estimated at something over 500,000. Most of them live in the highly industrialized Spanish part of the Basque country. The Basques have derived their name, Euskaldunak, from Euskara, the native word for their language. According to the classification of the philologist Prince Louis-Lucien Bonaparte (1813–91), there are eight modern dialects of Basque. Dialectal division is not strong enough to mask the common origin or to preclude mutual understanding. Basque attained official status only for a short period (1936–37) during the Spanish Civil War, under the Basque autonomous government.

Origins and classification. Basque remains an isolated language with no known linguistic relatives. The hypothesis of the German philologist Hugo Schuchardt (1842–

1927), which once had wide currency, posited an intimate genetic connection between Basque and Iberian (see below) and the Hamito-Semitic (Afro-Asiatic) language group. This theory was superseded by attempts to establish a more or less close link between Basque and Caucasian, the language group indigenous to the Caucasus region. A lack of common linguistic characteristics between the Basque and Hamito-Semitic languages makes Schuchardt's hypothesis extremely dubious. There are, however, some common features that favour the relationship between Basque and Caucasian. Still, proof of a genetic relationship beyond reasonable doubt appears remote. Perhaps the most promising theory current in the 1970s involves the comparison of Basque with the long-extinct Iberian, the language of the ancient inscriptions of eastern Spain and of the Mediterranean coast of France. But, despite amazing phonological coincidences, Basque has so far contributed next to nothing to the understanding of the now-readable Iberian texts. Therefore, it is possible that the similarity may have resulted from close contact between Basques and Iberians and not from a genetic linguistic relationship.

Link between Basque and Iberian

History of the language. At the beginning of the Christian Era, dialects of Euskarian (Basque) stock were probably spoken north and south of the Pyrenees and as far east as the Valle de Arán in northeastern Spain. It is likely that only the disruption of Roman administration in these regions saved the Basque dialects from being completely overcome by Latín. It is also likely that the Basque tongue, which had a firm foothold in the country that then began to be called Vasconia, experienced a substantial expansion toward the southwest, which carried it to the Rioja Alta (High Rioja) region in Old Castile and near Burgos. The more eastern Basque dialects, separated from the main area by Romance-speaking populations, were doomed. During the Middle Ages, Basque, the language of a population more peasant than urban, could not possibly hold the field as a written language against Latin and its successors, Navarrese Romance and, to a certain extent, Occitan (the *langue d'Oc*, also called Provençal) in the kingdom of Navarre. Since the 10th century, Basque has slowly but steadily lost ground to Castilian Spanish; in the north, however, where French is a more modern rival, the Basque-speaking area is practically the same as it was in the 16th century. In the last two centuries, above all in industrial centres, Basque has had to fight for survival in the heart of the Basque-speaking country, as well as on the frontier of the Basque-speaking area.

Basque records and writing

Latin inscriptions from the Roman period, found mostly in southwestern France, record a handful of proper names of unmistakable Basque etymology. From AD 1000 on, records consisting chiefly of proper names but also of Basque phrases and sentences grew more numerous and reliable. The first printed Basque book, dating from 1545, began an uninterrupted written tradition. Scholarly Basque literature, with its prevailing religious interests, has been neither abundant nor varied until recent times. Intense efforts are now being made to introduce Basque as a vehicle of private primary education. In addition, a model of a unified, standard written language also seems to be gaining increasing acceptance.

Phonology. The sound pattern of Basque is, on the whole, similar to that of Spanish. The number of distinctive sounds is relatively low compared with other languages. Combinations of sound (*e.g.*, consonant clusters) are subject to severe constraints. It can confidently be asserted that certain types of consonant clusters, such as *tr*, *pl*, *dr*, and *bl*, were all but unknown about two millennia ago. The common sound system underlying the systems of the present Basque dialects has five (pure) vowels and two series of stopped consonants—one voiced (without complete stoppage in many contexts), represented by *b*, *d*, *g*, and the other voiceless, represented by *p*, *t*, *k*. Nasal sounds include *m*, *n*, and palatal *ñ*, similar to the sound indicated by *ny* in "canyon." In this respect, as in others, Basque orthography coincides with the Spanish norm. There are two varieties of *l*, the common lateral *l* and a palatal variety, *ll*, as in Spanish, that sounds similar to the *lli* in "million" (as *l* + *y*). The Basque *r*, made by a single tap of the tongue against the roof of the mouth, contrasts with a rolled or trilled *r*, written *rr*. Two phonological features are worthy of special attention. Sibilants (both fricatives and affricates) made with the area of the tongue directly before the dorsum (the back of the tongue) are distinct from the apical sibilants, produced with the tip of the tongue. The letter *z* in Basque symbolizes the predorsal fricative, and *tz*, the predorsal affricate sound; *s* and *ts* represent the apical fricative (similar to Castillian Spanish *s*) and affricate, respectively. (A fricative is a sound, such as English *f* or *s*, produced with friction and, hence, without complete stoppage in the vocal tract; an affricate is a sound, such as *ch* in "church" or *j* in "jam," that begins as a stop and ends as a fricative, with incomplete stoppage.) In addition to these hissing sibilants, Basque also includes the hushing ones, written as *x* and *tx*; they are like the English *sh* and *ch*. The *x* and *tx* sounds, along with the palatal sounds written as *ll* and *ñ*, often have an expressive value (diminutive, endearing) in comparison with their non-palatal counterparts; *e.g.*, *hezur* means "bone" and *hexur* "little bone" (fish bone, for example); *sagu* is "mouse" and *xagu* "little mouse."

Varieties of sibilants

The phonology of some Basque dialects may be more complex than that presented in the preceding paragraph. In the easternmost Souletin region, for example, the dialect has acquired, by internal development or by contact with other languages, a sixth oral vowel—rounded *e* or *i*—and nasal vowels, voiced sibilants, and voiceless aspirated stops. The aspiration accompanying stop consonants consists of a small puff of air. There is also, word-initially and between vowels, an aspirated *h*, once common but now peculiar to the northern dialects. It has also been retained in the proposed standard form of Basque.

Grammar. The mention of two features is unavoidable in describing Basque syntax. Basque is, in the first place, a language of the so-called ergative type. That is, it has a case denoting the agent of an action. Hence, what in English would stand for the subject of a transitive verb is expressed in Basque by means of a suffix *-k*; for example, in the sentence "the foot serves the hand, and the hand serves the foot," *oinak zerbitzatzen du eskua, eta eskuak oina*, the first word, meaning "the foot," is composed of three elements, *oin* "foot," *-a*, "the," and *-k*, which marks the Basque equivalent of the subject of the verb. The fourth word, meaning "the hand," does not have the *-k* ending. In the second clause, *eta eskuak oina*, the word for hand, *eskuak*, now has the ergative *-k* ending to indicate that the hand is the agent of the clause "the hand serves the foot." The subject of an intransitive verb, which is not distinguished from the object of a transitive verb, has no overt mark—*e.g.*, in "if the belly does not eat, the belly itself will fail," *sabelak jaten ez ba du, sabela bera ihartuko da*, the first term, *sabelak* "the belly," has the *-k* marker because it is the agent of a transitive verb "eat"; but, in the second clause, *sabela* is the subject of the intransitive verb "fail" and, therefore, has no overt grammatical mark.

The second characteristic feature of Basque concerns the finite verb, which acts as a summary of all the noun phrases in the sentence. It has markers for all three persons—the 1st, 2nd, and 3rd—and may contain as many as three personal references (for subject, direct object, and indirect object). *Da*, for example, means "is," *du* means "he has it," and *dio* means "he has it for him" in the sentence *oinari ez dio eskuak kolperik emaiten* "the hand does not give a blow [*kolpe*] to the foot [*oin-a-ri*]." In certain situations the interlocutor can also be referred to within the verb. Further, most Basque verbs have only a compound conjugation; *e.g.*, *erori da* "he has fallen," literally, "he is fallen," and *jaten du* "he eats [is eating] it."

Although some ancient prefixes are still apparent in modern Basque, they are no longer productive, so that Basque can be characterized as an over-all suffixing language; that is, it appends suffixes to the ends of words. There is one declension with suffixes or postpositions to indicate number and case; *e.g.*, *etxe-a* "the house," but

etxe berri-a "the new house," and *etxe berri-a-ri* "to [for] the new house." Suffixes, under certain restrictions, may be heaped upon one another. Theoretically, genitival endings indicating possession may be added to one another without limit. This is similar to the case in English of "the button of the coat of the son of the Major of York"; in Basque, however, the phrase "of the" is indicated by an ending, *-(r)en*, added to the noun. Noun suffixes can also be attached to verb forms in order to express subordination of the clauses in which the verb forms appear; *e.g.*, *da* "is," *den* "which is," *dena* "that (*-a*) which is," *denean* "when there is," literally, "in that which is." Prefixes are also used for that purpose; *e.g.*, *ez du jaten* "he does not eat" with the particle *ba* "if" becomes "if [the belly] does not eat," *jaten ez ba du*.

Vocabulary. Basque has preserved a peculiar and distinctive appearance, despite the overwhelming pressure it has been subjected to during at least 2,000 years. Nevertheless, its borrowings from the neighbouring languages, especially words and idioms, can hardly be underrated. Loanwords from the Romance languages are numerous. Some of them bear the stamp of their archaic Latin ancestry; *e.g.*, *bake* "peace" from Latin *pax, pacis*, *bike* "pitch" from Latin *pix, picis*, and *errege* "king" from Latin *rex, regis*. Contrary to a widely held opinion, Indo-European loanwords of non-Latin origin are extremely scarce. Derivation, the formation of new words by the use of suffixes, is accomplished partly through the use of borrowed suffixes. This practice, as well as the compounding of nouns to form new words, as in *bizkar-hezur* "backbone," has been very much alive throughout the history of the language. Basque itself has contributed but little to the Spanish, Occitan, French, and English languages. But family and place names of Basque coinage are frequent in Spain and in Latin America, where they can be found in such proper names as Aramburu, Bolívar, Echeverría, and Guevara.

Latin loanwords

BIBLIOGRAPHY. RENE LAFON, "La lengua vasca," *Enciclopedia lingüística hispánica*, vol. 1 (1960), perhaps the best short introduction to Basque, both descriptive and historical; HUGO SCHUCHARDT, *Primitiae linguae Vasconum*, 2nd ed. (1968), detailed commentary of an Old Basque text, with an introduction and up-to-date bibliography by A. TOVAR; P. LAFITTE, *Grammaire basque*, 2nd ed. (1962), a standard normative grammar; J. COROMINES, *Estudis de toponímia catalana*, 2 vol. (1965-70), presents new data on the survival of Basque dialects in the Middle Ages; J.M. LACARRA, *Vasconia medieval* (1957), authoritative review by an historian of the linguistic situation in and around the Basque country; LUIS MICHELENA, *Fonética histórica vasca* (1961), essay on the reconstruction of the phonological system of Proto-Basque; LUIS MICHELENA (ed.), *Textos arcaicos vascos* (1964), an annotated collection of documents from the Antiquity to 1700; RENE LAFON, *Le système du verbe basque au XVIe siècle*, 2 vol. (1943), the best account of form and function of the Basque verb; in *Le Langage*, directed by ANDRE MARTINET, pp. 1414–1437 (1968), may be seen as a rather skeptical survey of old and recent theories about genetic relationship; A. TOVAR, *La lengua vasca*, 2nd ed. (1954; abridged Eng. trans., *The Basque Language*, 1957), and *The Ancient Languages of Spain and Portugal* (1961), a discussion of the problem of the position of Basque among these now extinct languages.

(L.M.)

Batteries and Fuel Cells

In strict usage, the term battery designates an assembly of two or more cells that convert chemical energy directly to electrical energy. The term, however, has long been applied equally to a single cell. In a voltaic cell, for example, chemical reactions take place so that electrons are released on one part (the anode, or negative electrode) of the cell and caused to flow through an external circuit to the other part (the cathode, or positive electrode). The process continues until the circuit is interrupted or one of the reactants is exhausted (see Figure 1).

At rest, a cell exhibits a potential difference, or voltage, between the two electrodes, which is determined by the amount of chemical energy available when an electron is transferred from one electrode to the other and is therefore subject to the chemical nature of the materials

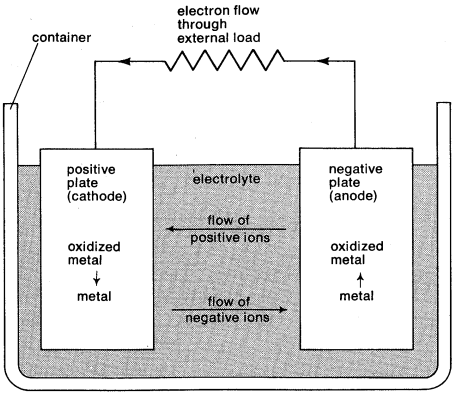


Figure 1: Basic components of an electrochemical cell.
From G. Vinal, *Primary Batteries* (©1950); John Wiley & Sons, Inc.

used in the electrode rather than the size of the cell. The current which flows from a cell is determined by the resistance of the total circuit, including that of the cell itself. If very large currents are required, a low-resistance cell is needed. This can be achieved by the use of electrodes with large areas. The maximum current that can be drawn from a cell is, thus, determined by the area of the electrodes. When a current flows, the cell voltage decreases because of the internal resistance of the cell and the slowness of the chemical processes at the electrodes. The drop in voltage, called polarization, is recovered when the current is stopped.

The cell has a limited energy content, called its "capacity," usually given in ampere-hours, and determined by the quantity of electrons that can be released at the anode and accepted at the cathode. When all of the chemical energy of the cell has been used up—usually because one electrode has become completely exhausted—the voltage falls to zero and will not recover. The capacity of the cell is determined by the amount of active material in the electrode, or, in other words, by the electrode volume (or the thickness for an electrode of a given area).

In a fuel cell, electricity is produced directly by the reaction of a gas or liquid fuel supplied to one electrode, and oxygen or air supplied to the other. To be continuously useful, the electrodes and the electrolyte between them should be unchanged by the reaction. Cells and batteries that do not depend upon a chemical reaction have been developed to convert solar and nuclear energy directly to electrical energy. These are commonly called solar batteries, thermal batteries, and nuclear batteries.

Fuel cells

Battery applications include such things as flashlights and radios, starting systems for internal-combustion engines, the propulsion of special-purpose vehicles, submarines, standby power sources for telephone systems, emergency lighting, and communications devices.

HISTORY

In 1791 Luigi Galvani, an Italian professor of anatomy at the University of Bologna, accidentally discovered that it was possible to cause a direct or continuous flow of current along an electrical conductor by bringing two dissimilar metals into contact with a moist substance. About 1800, acting upon Galvani's discovery, Alessandro Volta, professor of natural philosophy at the neighbouring University of Pavia, created the first battery, the epoch-making "voltaic pile" (see Figure 2). Volta assembled a series of silver and zinc disks in pairs, separating each pair with a sheet of pasteboard soaked in a conducting liquid (salt water). When the top disk of silver was connected by an external wire to the bottom disk of zinc, a current was produced. Volta believed that the current was generated by the contact of silver with zinc, whereas it was actually developed across the salt-soaked pasteboard, which was an electrolyte. Volta constructed piles with as many as 60 pairs of disks, but in use the pile always dried out and ceased to operate, forcing him finally to make a modification that he called a crown of cups. This device consisted of a series of cups filled with a salt solution, each cup

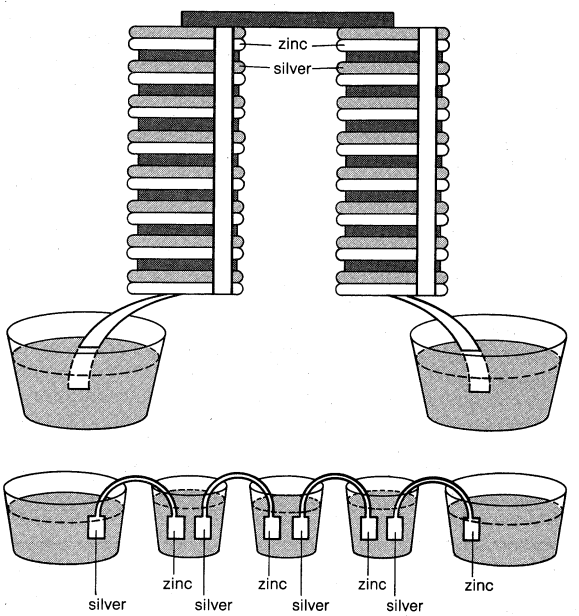


Figure 2: Alessandro Volta's (top) pile and (bottom) crown of cups.

From G. Vinal, *Primary Batteries* (©1950), John Wiley & Sons, Inc., after the *Philosophical Transactions of the Royal Society* (1800)

containing a piece of zinc and silver. The zinc in each cup was connected electrically to the silver of an adjacent cup or cell and the cups arranged in a circle. He measured the strength of his battery simply by placing his fingers across the end terminals.

In 1836 a professor of chemistry in London developed the classic form of the primary cell; that is, a battery that is nonchargeable, and once used up cannot be used again. In this cell, known as the Daniell cell, the positive electrode, or anode, was a rod of pure zinc, immersed in sulfuric acid (the electrolyte). A coating of mercury protected the zinc from attack by the acid. The negative electrode, or cathode, consisted of a copper canister, containing sulfuric acid saturated with copper sulfate (see Figure 3).

From G. Vinal, *Primary Batteries* (©1950); John Wiley & Sons, Inc.

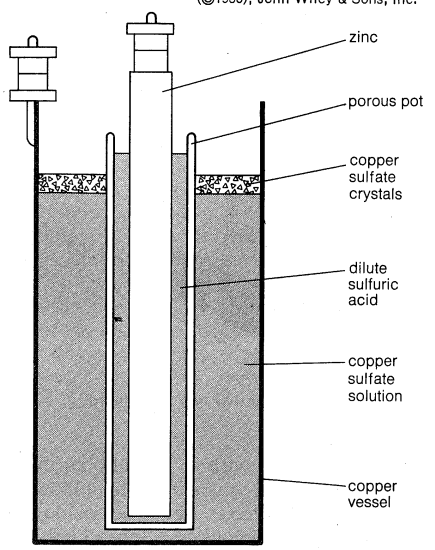


Figure 3: A Daniell cell.

Three years later, in 1839, Sir William Grove, a British jurist who made notable contributions to science, published a description of a battery of cells using platinum electrodes housed in inverted test tubes in a bath of sulfuric acid and water. When an electric current was passed through this battery, hydrogen and oxygen collected in the tubes because of the decomposition of the water. From this "charged" battery, an appreciable current could be drawn. It was the first fuel cell in which the reactants, hydrogen and oxygen, were not included in the electrodes themselves, and it was also the first secondary or storage battery, so-called because the action can be reversed by a direct current so that the cell is made to appear to "store" electricity.

The lead-acid cell, the first practical storage battery and probably the most widely used battery today, was invented by Gaston Planté, a French physicist, in 1859. Planté's cell consisted of two sheets of lead, separated by strips of rubber and rolled into a spiral. When immersed in a 10 percent solution of sulfuric acid and charged, the cell was capable of storing electrical energy. Though Planté's cell could deliver its stored energy very rapidly (that is, it could produce a large current) it remained a laboratory curiosity for 20 years.

The first "dry" cell was invented about 1865 by Georges Leclanché, a French chemist, and remains one of the most widely used primary cells. In its original form (see Figure 4), it consisted of a glass jar filled with the elec-

From G. Vinal, *Storage Batteries* (©1951); John Wiley & Sons, Inc.

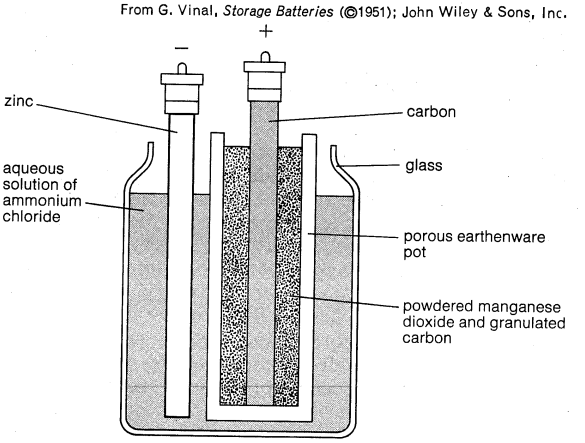


Figure 4: Georges Leclanché's cell.

trolyte, an ammonium chloride solution. The anode was a zinc rod, protected from corrosion by mercury. When a current flowed, zinc ions were formed and passed into solution. In the centre of the jar was a porous pot, filled with a mixture of manganese dioxide and carbon powder that served as an electrical conductor to connect the manganese dioxide powder particles with the cathode, a carbon rod in the centre of the cell. When a current flowed, manganese dioxide was reduced to manganous oxide.

PRIMARY OR VOLTAIC BATTERIES

Dry cells. Primary batteries or cells are those that, once discharged, are discarded. Early primary cells used liquid electrolytes and were not very portable. Today, most primary cells use electrolytes in jelly form, absorbed in a sheet of a porous material, and are known as dry cells.

Acidic dry cells. The modern acidic dry cell is simply an improved version of the 1865 cell described above, one in which the electrolyte is made into a jelly, the porous pot is replaced with a muslin bag, and the glass jar is replaced by a thin zinc cylinder that is the anode. To inhibit leakage, more zinc than necessary is used, and some modern cells also have a steel outer jacket. Dry cells

One version of this cell, which has been turned "inside out," uses a central zinc rod, a fabric separator, and annular rings of a manganese dioxide and carbon mixture packed inside a steel cylinder. Dry cells are almost universally used today in flashlights, portable radios, and toys.

Alkaline dry cells. Another improved version of the Leclanché cell is the manganese-alkali cell, in which the ammonium chloride has been replaced by potassium hydroxide. This electrolyte has a lower resistance, allowing larger currents to be drawn from it. Constructed on the The manganese alkali cell

"inside-out" principle, with a porous zinc anode in the centre, the cell provides higher capacity than that obtainable in a conventional Leclanché cell of the same size. Manganese-alkali cells are used in portable devices, which require large amounts of current (have high current drain), such as devices containing motors, rather than with electronic equipment. Battery-operated razors, cameras, and tape recorders are in this category.

Another modern alkaline dry cell is the mercury cell, using mercuric oxide (mixed with graphite) instead of manganese dioxide at the cathode, and an alkaline electrolyte (potassium hydroxide). This alternative cathode is capable of sustaining far higher currents than manganese dioxide, thus providing a greater capacity for a given battery weight and size. Mercury batteries are expensive, however, compared to other dry-cell types. They are limited to devices that require high current pulses or extremely long service. Photographic flashguns and hearing aids are typical uses.

Other dry cells. The "air–depolarized" cell or "metal–air" cell uses an air electrode from a fuel cell as the cathode and a conventional metal anode, usually zinc. Since an air electrode uses atmospheric oxygen as its active material, it has an infinite capacity and is light in weight. In recently developed zinc-air cells, the space that is normally occupied by the manganese dioxide is filled instead with extra zinc. Ten times more capacity can thus be obtained. In another form of zinc-air cell, the spent zinc anodes can be removed and replaced with new anodes. This mechanically rechargeable battery is useful for portable communications devices. Less successful, but significant, attempts have been made to develop magnesium-air, aluminum-air, and iron-air battery systems.

Wet cells. Another potentially important class of primary cell is the so-called reserve battery. It was found in the early 1970s that such high-energy compounds as magnesium, silver chloride, and cuprous chloride, could be used to obtain a high amount of energy from small or light batteries for special applications, primarily military. Since these materials are attacked by electrolytes, the battery is assembled with the electrolyte stored in a separate container. The electrolyte is inserted into the battery immediately before it is required for use, and the battery consequently enjoys a short but active life. In one form, seawater is the electrolyte; in another, liquid ammonia.

Standard cell. Since the voltage of a cell is determined solely by the chemical nature of the electrodes and electrolyte, a carefully made cell can be used as a standard reference voltage source. Identical cells, constructed from the same materials anywhere in the world and operated under the same conditions, will always produce the same voltage.

Weston cadmium cell

The best known standard cell is the Weston cadmium cell (see Figure 5), which consists of a glass tube in the form of the letter H. A pool of mercury at the bottom of each limb makes contact with a metal electrode inserted through the glass. At the anode, the mercury is amalgamated with 10 percent cadmium, in contact with crystals of cadmium sulfate. At the cathode, the mercury pool is in contact with a layer of mercurous sulfate, in turn covered with cadmium sulfate crystals. The electrolyte is a saturated solution of cadmium sulfate. If the materials of this cell are prepared in a carefully controlled and specified way, the voltage is 1.01864 volts at 20° C, with a well-defined temperature coefficient (that is, a close and regular relationship exists between changes in temperature and changes in voltage). Used as a definition standard for the volt, it forms one of the basic standards for all other electrical units. Many modern electrical devices use a Weston cadmium cell in their circuitry as a standard reference voltage, but the construction is miniaturized and somewhat modified from the classical laboratory cell described. One commercial Weston cell is in the form of a cylinder, four centimetres long and one centimetre in diameter, which can operate in any orientation.

From G. Vinal, *Storage Batteries* (©1951); John Wiley & Sons, Inc.

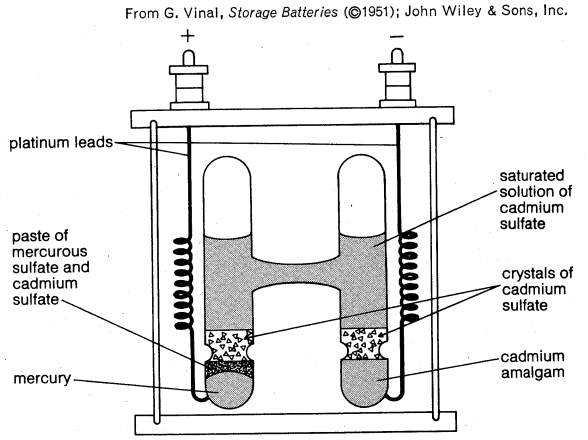


Figure 5: Weston normal or saturated standard cell.

SECONDARY OR STORAGE BATTERIES

Some cells can be recharged by passing a current through them in the reverse direction. The chemical processes that occurred at the electrodes during discharge are then reversed, and the cell recovers its original state, except that some energy is lost during the charge-discharge cycle.

Lead acid. When dynamogenerators became common, Planté's cell was studied more intently and is now widely used in vehicles of all kinds. This cell produced active lead oxide and spongy lead deposits on the surfaces of lead sheets, however, and, thus, had shortcomings. It was clear that better performance could be obtained if these materials were plastered into an open mesh or grid, made of lead. The electrodes of modern lead-acid batteries employ this form of construction (see Figure 6), using grids

From G. Vinal, *Storage Batteries* (©1951); John Wiley & Sons, Inc.

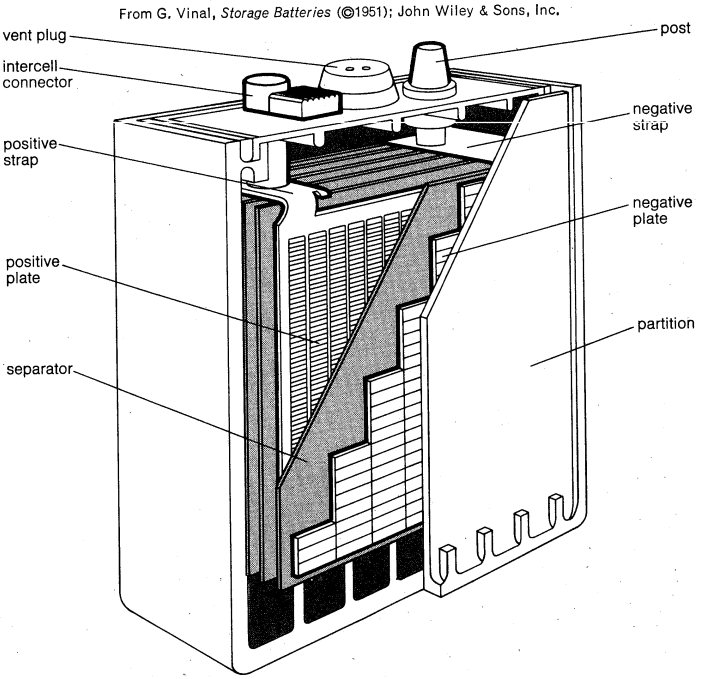


Figure 6: Construction of the automotive type lead-acid battery (cut-away view).

made from an alloy of lead with 5 percent antimony added to improve mechanical qualities. The plates are plastered with a mixture of litharge (lead monoxide) and dilute sulfuric acid. Some red lead is often added. The negative plate also contains small amounts of lampblack, barium sulfate, and organic extracts of wood to act as "expanders" to improve the performance. The plates are cured or dried in air, and then processed to provide positive and negative electrodes. The negative electrode is reduced to spongy lead, while the positive electrode is oxidized to lead dioxide. The plates, separated from each other by a porous insulating separator that may be wood, rubber, plastic, or glass fibre, are immersed in an elec-

trolyte of dilute sulfuric acid contained in a glass or rubber composition tank.

Lead-acid batteries

Normally several cells, each with several plate pairs connected in parallel, each cell providing 2.05 volts, are connected in series to make 6- or 12-volt batteries. The electrolyte in each cell is housed separately in its own compartment.

During discharge each plate is converted to lead sulfate, and the sulfuric acid is used up, producing water. The amount of charge remaining in the cell can be determined by measuring the density of the electrolyte compared to water (specific gravity). The specific gravity of the electrolyte varies from about 1.28 (that is, the electrolyte is 1.28 times as dense as water) when fully charged, to 1.12 when discharged.

When a direct current is passed through the battery, the electrode reactions are reversed and the cell is recharged. If the cell is overcharged, electrolysis of water occurs, producing hydrogen and oxygen. Overcharging, thus, results in a loss of water that must be replaced. It is never necessary to add more sulfuric acid.

Nickel-cadmium. The nickel-cadmium cell consists of a nickel hydroxide cathode and a cadmium anode in an electrolyte that is a solution of potassium hydroxide. The plates are highly porous nickel sheets, impregnated with cadmium and nickel salts and chemically converted to the appropriate oxides. In the pasted-plate type, the plates are made from nickel-plated steel structures containing many pockets filled with active materials, cadmium oxide in one plate and nickel hydroxide in the other.

The plates are assembled in a steel container filled with a solution of potassium hydroxide. No porous separators are used, the electrodes being held in place by plastic or rubber moldings that prevent them from touching each other. The potassium hydroxide is not used up, so that the specific gravity of the electrolyte cannot be used to indicate the amount of charge remaining in the battery. Because it is lighter, it is more often used in portable equipment, such as radios. Modern developments include a sealed nickel-cadmium battery for use in cordless appliances.

Nickel-iron. Once important as a long-life, rugged storage device, the nickel-iron battery, sometimes called the Edison cell, is being displaced by the nickel-cadmium battery. In the nickel-iron battery, the positive plates are nickel-plated steel "pocket" units, packed with a mixture of nickel hydroxide and graphite, the latter added as an aid to conductivity. The negative plates are similar units, packed with a mixture of iron powder, ferrous oxide, and mercurous oxide. On discharge, the nickelous hydroxide is converted to nickelic hydroxide, as in the nickel-cadmium cell, while the iron is converted to ferrous oxide; on charge, the action is reversed. The electrolyte is potassium hydroxide, to which a small amount of lithium hydroxide is added to improve the cell's capacity.

Silver-zinc. Another type of storage battery, increasingly significant in recent years, is the silver-zinc battery, in which the electrolyte is a solution of potassium hydroxide saturated with zinc hydroxide. The negative electrode, a porous plate of pure zinc, to which some mercury is added to suppress the corrosion of zinc by the electrolyte, is formed around a silver mesh screen. The positive electrode is a silver screen, pasted with silver oxide, usually electrochemically oxidized to silver peroxide. The plates are prevented from touching each other by separators made from cellulose sheets. The cells have a high energy-to-weight ratio, and are useful in applications in which lightness is more important than cost. These batteries have a limited-cycle life. That is, the number of times they can be charged and discharged is limited by the stability of the separator, which ultimately allows short circuiting. Lives of from 30 to 300 cycles have been reported for different designs.

FUEL CELLS

Like any other electrochemical cell, a fuel cell consists of two separate electrodes and an electrolyte. It differs from others in that the chemicals supplying the electrode processes are stored separately, and are supplied to the electrodes on demand. The reactants can be any pair of materials that are oxidizing (combining with oxygen molecules) and reducing (removing oxygen molecules), but the simplest fuel cell to describe, construct, and understand is one that operates on hydrogen and oxygen. Hydrogen reacts on the surface of a conducting electrode to produce hydrogen ions and electrons. The ions are transmitted through the electrolyte to the other electrode, where they absorb electrons from the electrode and react with oxygen to produce water. The reaction proceeds for as long as there is a supply of reactants and an external circuit for electrons to flow to balance the reactions. In reverse, this process is the electrolysis of water to produce hydrogen and oxygen. The reactants usually are gaseous, and the electrode must therefore provide a stable interface (point of contact) between the gaseous reactant phase, the electrolyte phase (usually liquid), and the electronic-conducting phase (usually solid).

This so-called three-phase interface problem is the key to successful fuel-cell electrode design. Many different techniques have been tested to achieve it; they fall into three main categories—capillary electrodes, in which the gas-liquid interface is stabilized within the pores of a conductor by surface-tension effects; hydrophobic electrodes, in which the interface is stabilized within the pores by wet-proofing part of the structure; and redox electrodes, in which the reactants and electrolyte exist together in one phase, in contact with a solid electrical conductor.

Although the electrode-electrolyte combination is the focal point of the fuel cell, the cell must also be equipped with a means of supplying fuel and oxidant in a suitably pure form at an appropriate rate and pressure. This may involve generating hydrogen from liquid fuel and/or scrubbing or filtration or both to remove impurities. Waste heat and water from the chemical reaction must also be removed at a controlled rate. Such an association of components, known as a fuel-cell system, is a self-balancing and self-contained device.

Because the fuel cell is not a heat engine, but operates at a uniform temperature, it is not subject to thermodynamic losses of heat engines, whose efficiency is about 40 to 50 percent. Theoretically, fuel cells can operate at a thermodynamic efficiency close to 100 percent. Moreover, since they contain few or no moving parts, fuel-cell systems are more reliable, need less maintenance, and are inherently quieter than conventional dynamo electrical generators. They produce a less noxious exhaust because electrochemical reactions are clear and complete.

Fuel-cell applications

For these reasons, fuel cells have possible applications in space vehicles, where efficiency and reliability are important; in remote areas where fuel transport costs and maintenance are at a premium; in crowded cities, where their quietness and clean exhaust are desirable; and in domestic "local generating stations" where quietness, cleanliness, and low operating costs are attractive. In the early 1970s, however, the high initial cost of fuel cells severely restricted their use.

Aqueous-acidic electrolyte type. Conventional, low-cost fuels consist basically of hydrocarbon compounds such as coal and oil that, when oxidized, produce carbon dioxide and water. If such a fuel is used in a fuel cell, the electrolyte must be compatible with carbon dioxide, and must therefore be acidic. Since the carbon dioxide and water are removed by evaporation, the electrolyte itself must not be volatile. This restricts the choice of common acidic media to sulfuric acid and phosphoric acid, both of which have been extensively used in experimental fuel cells.

Since acids are notoriously corrosive, especially under the conditions that prevail at the oxygen or air electrode; the choice of conducting material, though somewhat limited, includes gold, platinum, titanium, and carbon. A slightly wider, though still restricted choice exists for the fuel electrode. Similar corrosion restrictions apply to the choice of catalysts (material that increase reaction rates without themselves taking part in the reaction); platinum metals have commonly been used.

Evidence has been obtained to show that hydrocarbons may be directly oxidized at temperatures of 100° C and above, but at rates and efficiencies that are disappointingly low. Alcohols, which are more reactive, may be dissolved in the electrolyte, and higher reaction rates have been achieved.

A special type of the aqueous-acidic electrolyte cell is provided by the use of an ion-exchange membrane as electrolyte. A sulfonated polystyrene-type membrane, equilibrated with hydrogen ions, can conduct H^+ ions very readily. Platinum-black catalyst is applied to each face of the membrane and current is picked up by a grooved- or corrugated-metal collector, in the manner of the capillary cell. These collectors are usually titanium or tantalum to resist corrosion.

It is necessary that water formed in the reaction not be allowed to dilute the electrolyte. It is, therefore, either evaporated, or picked up by a system of wicks and carried outside the cell. This type of cell operates at ambient temperature and pressure on hydrogen or impure hydrogen fuel, and oxygen or air. It is limited to an upper temperature of about 50° to 70° C (120°–160° F) by the thermal stability of the membrane in the presence of oxygen and platinum catalyst. In one system developed for space use, the cells were arranged to give two kilowatts when fed with pure hydrogen and oxygen.

A number of successful demonstrations have been made of acid-electrolyte cells that operate on a mixture of hydrogen and carbon dioxide, cheaply produced from natural gas and steam.

Aqueous-alkaline electrolyte type. Alkaline electrolytes in general are less corrosive than acids, permitting the use of a wider range of materials. For this reason, far more research and development have been carried out on alkaline than on acid fuel-cell systems. Unfortunately, alkaline electrolytes are attacked by carbon dioxide, forming carbonates. Carbon dioxide is a minor constituent of atmospheric air, the product of oxidation of hydrocarbons, and a common impurity in commercial hydrogen supplies, and must be completely removed from the feed to alkaline fuel cells.

Capillary cell

Two different types of aqueous-alkaline electrolyte fuel cells have been developed. One is known as the capillary cell, or trapped electrolyte cell, in which the electrolyte is contained within the pores of an insulating matrix—for example, a sheet of asbestos paper or glass-fibre filter paper. The other is the free electrolyte cell, in which the electrolyte is circulated through the gap between two electrodes.

In the capillary cell, the electrodes are formed by a mixture of a powdered catalyst, such as platinum, and a wetproofing agent, such as the organic compound polytetrafluoroethylene. This mixture is embedded in the pores of a woven metal screen, usually nickel, which acts as an electrical conductor and a structural member. The electrodes are clamped against opposite faces of the sheet of matrix material by means of conducting plates of nickel or nickel-plated steel. The plates are formed so as to leave grooves or passages for the gas supply to flow between the plate and the electrode, while the parts in contact allow the electrons to flow from the electrode, through the plate, and into the electrode of the adjacent cell of the series. The electrolyte is contained within the pores of the matrix, wetting the catalyst particles. The major part of the electrode structure is kept dry and accessible to gas by the waterproof nature of the polytetrafluoroethylene, which is in the form of small agglomerated particles in a porous mass. Water produced by the reaction of hydrogen and oxygen, for example, is carried from the cell by an excess of either hydrogen or oxygen. Water evaporates into this stream and is condensed outside the cell. Excess heat is removed from the cell either by a system of fins mounted on the edge of the nickel plates or by a coolant that flows through the hollow passages within the plates.

In free-electrolyte cells, the electrodes are a dual layer structure, with a fine-pore layer on the side facing the electrolyte. Gas is fed to the coarse face of the electrode, at a pressure insufficient to overcome surface-tension forces and expel the electrolyte from the fine pores. Stabilization of the interface is assisted in some cases by making part of the electrode structure wetproof with polytetrafluoroethylene or paraffin wax. Electrodes of nickel and of carbon have been used.

To obtain a reaction between hydrogen and oxygen at temperatures around 0° to 70° C (32° to 160° F), the normal regime for these cells, highly active catalysts are used. Most frequently, platinum group metals in finely divided form, known as blacks, have been employed. Silver and carbon also have been used as oxygen electrodes, while nickel and nickel boride have been used at the hydrogen electrode.

Although it requires pure gas supplies, this type of cell starts operating at room temperatures or below and has found use in specialized military and space applications. Though demonstrated in a number of commercial applications, it has not yet been produced in commercial quantities.

Fused-salt electrolyte type. To avoid the use of expensive catalysts, much work has been carried out on fuel cells operating at higher temperatures. Experimental work has been carried out at 200° C (390° F) using aqueous electrolytes pressurized to 600 pounds per square inch in which the precious metal catalysts were not used. A modification of such an experimental cell, using 85 percent potassium hydroxide electrolyte at 200° C, has been operated at ambient pressures. This system was developed between 1958 and 1967 and provided electrical power for several space-flights including the first manned moon landings.

Space flight power

An experimental cell that has been developed uses a mixture of sodium-potassium-lithium carbonates and operates at approximately 500° to 700° C (930°–1,290° F). The electrolyte is usually contained in the pores of a ceramic matrix, analogous to the capillary cell previously described. The electrodes are porous, metal layers applied to each face of the matrix, consisting of nickel at the fuel side and silver, copper oxide, or nickel oxide at the oxidant side.

A mixture of air and carbon dioxide is fed to the oxidant electrode, where it reacts to form carbonate ions. These ions are transmitted through the electrolyte to the fuel electrode, where they are discharged by reaction with the carbon monoxide or hydrogen fuel to form carbon dioxide (and water). This type of cell requires the addition of carbon dioxide to the air to allow it to operate, in contrast to the alkaline cell that requires all carbon dioxide to be eliminated. The addition is usually achieved by mixing some of the fuel exhaust with air, and feeding the mixture to the air electrode. All traces of unreacted fuel must first be removed.

At the high temperature of operation, a mixture of steam reacts on a nickel surface (*i.e.*, the fuel electrode) to form hydrogen and carbon monoxide, both of which can react electrochemically. Cells of this type have been operated on such hydrocarbon fuels as methane and kerosene.

Though this type of cell is inherently cheap to construct and simple to operate, problems of corrosion and instability have limited its development, and no commercial applications have yet been found. It is particularly suitable as a combined source of electricity, heat, and hot water for small-scale industrial and domestic units.

Redox type. Bringing together three phases—gas, liquid, and solid—at the electrodes is one of the major problems with the fuel cell. The redox (reduction-oxidation) cell eliminates the problem. In it, a solid catalytic metal electrode is placed in contact with an electrolyte containing a dissolved substance in two different states of oxidation; for example, a mixed solution of ferrous iron and ferric iron. Ferrous iron behaves as the "fuel" that is oxidized to ferric iron on the surface of the electrode. At the other electrode, another reduction-oxidation couple (redox couple), such as a mixture of bromine and bromide ions, acts as an oxidant. When a current is drawn, the two electrolytes become richer in the products of reaction. To avoid a direct chemical reaction between the reactants, a porous separator keeps the two electrolytes

apart, permitting the transfer of common ions but not the reactive ions themselves.

The electrolyte in the fuel compartment is circulated through a regenerator, in which it is reduced from the ferric to the ferrous form by means of conventional fuel, such as hydrogen or carbon monoxide. The other is regenerated by blowing air through the solution. Both of these processes occur on the surface of a catalyst, but no electrochemical process is involved.

This type of cell has been limited to academic interest because of the additional complexity of the system, the inefficiency due to leakage of reactants through the separator, and the relatively slow regeneration processes. Much progress has been made toward better "three phase" electrodes, of the types described previously.

Other fuel cells. Another type of electrolyte is a solid oxide such as zirconia, which conducts oxygen ions above approximately 900° C (1,650° F). Thin sheets or tubes of zirconia, stabilized in the "fluorite" crystal structure by the addition of calcium oxide or yttrium oxide, are faced on either side with porous metal electrodes of nickel or nickel oxide. At the high temperature of operation, hydrocarbon fuels are broken down or react with steam to form hydrogen, which reacts with the oxide ion to form water. Oxide ions are replaced in the electrolyte by the reaction of atmospheric oxygen at the other electrode. Problems of materials stability, particularly at the air electrode, have prevented this type of cell from finding applications that would have to be limited to relatively large units maintained at about 1,000° C (1,830° F).

A special type of alkaline cell uses a dissolved substance as fuel rather than a gas. Hydrazine, dissolved in a potassium-hydroxide electrolyte, is pumped through the pores of a nickel electrolyte. Nitrogen and water, the products, must be removed without the loss of excessive amounts of fuel. The air-electrode catalyst is silver, which is selective to oxygen and does not decompose hydrazine. Cells of this type have been developed and used for military purposes, in which the high cost and toxicity of the fuel can be tolerated. A similar system using methanol requires the replacement of the electrolyte each time the system is refuelled because the carbon dioxide product attacks the alkali. This has been used for remote area communications devices, but has not yet achieved commercial significance.

NUCLEAR AND SOLAR CELLS

There is at present no commercially practical device for converting the energy released in a nuclear-fission process directly into electricity. But the radiation emitted from some radioactive substances, such as strontium-90, can be caught on a collector that develops a negative charge, leaving the source positive. Such a nuclear cell can generate high voltages but only small currents.

The nuclear battery

Radiation from the sun can be converted directly into electrical energy by an array of photoelectric cells, or solar cells. When sunlight falls on these cells, a voltage is generated and small currents can be drawn. Little practical use has been made of such cells in terrestrial operation, owing to the relatively low amount of energy per unit area arriving through the atmosphere, the poor efficiency of the cells, and their high cost. They are widely employed, however, in artificial satellites and unmanned spacecraft. Because they produce a steady supply of electrical energy when illuminated, they are usually employed to maintain secondary batteries in a fully charged condition, thus providing for peak loads, as in the operation of transmitting equipment.

PRESENT PROBLEMS AND FUTURE TRENDS

Efficiency and cost considerations. Batteries and fuel cells convert chemical energy directly into electrical energy by an electrochemical process. Such a process is far more efficient than most other indirect routes for converting chemical to electrical energy, or other forms of energy conversion. The Table shows typical efficiencies possible from various types of processes.

Efficiency of Various Energy Conversion Devices

conversion device	efficiency
Electrochemical	90
Heat engine	40
Thermionic	8
Thermoelectric	7
Solar cell (photoelectric)	12
Nuclear (β + current)	1

The cost of electrical energy obtained from electrochemical processes is rather more expensive than that obtained from heat engines operating on cheap fossil fuel (coal and oil). For example, zinc, which costs about $300 per ton in bulk, even if converted from its raw state to electricity at 100 percent efficiency, would cost over 30 cents per kilowatt-hour. It is quite impossible to operate a primary zinc battery for less than this cost. In practice, the energy costs from practical primary batteries depend enormously on battery size and type, ranging from about $10 per kilowatt-hour and up. A large-scale power station, in contrast, generates electricity for as little as under one cent per kilowatt-hour. Because packaged electrical energy is convenient, users are prepared to pay this high price for energy in some applications.

A secondary battery stores electrical energy in a convenient and moderately portable form. Both the charging and discharging processes are accompanied by losses. Efficiency depends mainly on the type of battery and the rate at which it is charged and discharged, but rarely exceeds 80 percent. The capital costs of secondary batteries are far higher than those of primaries, but as they can be used many times over, this is often not a significant factor. Mass-produced lead-acid batteries for automobile use are designed to be sold at a relatively low price, and, consequently, are not as long-lived as they could be at a higher price. There is a far higher unit cost for the more reliable and rugged batteries of electric trucks or submarines.

Battery efficiency loss

Production. Since commercial fuel cells have not yet been produced in quantity, it is impossible to obtain a true estimate of their cost. Generally speaking, fuel cells employing expensive fuels, such as hydrazine, are cheaper to construct than are more complex systems that use cheap fossil fuels. There will be a trade-off between the capital cost of the cell and the cost of the electrical energy that it produces. It is recognized that the main factor that has thus far prevented a commercial fuel cell from becoming a reality is the very high cost of the cell itself.

The battery industry has had a period of rapid growth since 1960, partly due to the introduction of new systems, but mainly due to the proliferation of new uses and increased traditional applications for conventional batteries.

Several exploratory attempts have been made to introduce fuel cells into commercial use, but by 1970 none had been particularly successful. Fuel cells operating on methanol have been used in small numbers to power TV repeater stations, and navigation buoys and beacons. Hydrogen-oxygen fuel cells have been demonstrated propelling a tractor, forklift truck, and a delivery van. Hydrazine-air fuel cells are believed to have had limited military use for communication systems. Hydrogen-oxygen fuel cells have been successfully used as the primary electrical power supply for spacecraft.

Future developments. One of the common faults with early primary batteries was that they were not portable. The development of the dry cell eased this problem, but these cells were not really leakproof, especially after complete discharge, when the zinc can corrodes. The dry-type Leclanché cells are now being made more resistant to leakage, and the trend is expected to continue.

A new type of dry cell, still under development, uses magnesium rather than zinc in an alkaline–manganese dioxide cell configuration. Magnesium stores far more energy than zinc of similar weight, but suffers from corrosion problems and does not deliver large current pulses.

Miniaturization of electronic circuitry has led to tiny devices such as radios and hearing aids that are dwarfed

by the batteries that drive them. There is a trend toward miniaturizing batteries, especially multicell packs that produce about nine volts. As cells become smaller, more advanced types of cells become attractive in order to maintain a reasonable capacity. The development of cheaper manganese-alkali, mercury, and metal-air cells in these miniature sizes promises to continue.

The shelf life of primary batteries is constantly being improved, but is still limited by internal corrosion of the metal electrodes. Developments of advanced battery systems with solid electrolytes, or improvements in the storage characteristics of conventional batteries, will be directed toward this problem.

It is possible that cheaper reserve-type batteries will become available for one-shot emergency applications. Commercial standby power supplies tend to use trickle-charged secondary cells at present, but reserve-type batteries would require no maintenance and be cheaper in the long run. A reserve-type starter battery for an automobile for emergency use has been introduced.

The biggest use of secondary batteries is in automobiles. Trends in the industry are to reduce the selling cost and to prolong the guarantee period of the battery. Much research is going into the development of lightweight batteries for portable applications, and for the propulsion of electric vehicles. For portable use in cordless appliances, nickel-cadmium appears to be the best choice. Successful sealing of secondary cells, preventing gassing vents, has been accomplished. Further developments with lighter, sealed, lead-acid cells, cheaper nickel-cadmium cells, and longer-lived cells designed to withstand frequent complete discharge, will continue.

Nickel-cadmium battery applications

For vehicle propulsion, no conventional battery appears capable of supplying the right characteristics for passenger-car performance. Research will continue on such high-energy systems as sodium-sulfur, lithium-chlorine, and lithium-tellurium. For larger urban vehicles, lighter batteries also are required, but it is possible that nickel-zinc, nickel-cadmium, or even lead-acid can be successful with suitable development. For special vehicle and aircraft applications, where a higher cost can be justified, the silver-zinc cell appears promising, and development of this type of battery to make it capable of deep discharge and long cycle life is likely to continue.

Another secondary battery, still under development but worthy of mention, is the sodium-sulfur battery that uses liquid sodium as one electrode, a solid aluminum-oxide electrolyte, and a sulfur and sodium sulfide mixture on the opposite side. A carbon current collector completes the cell. Current is produced when liquid sodium forms ions that migrate through the aluminum oxide and react with sulfur. The cell is reversible and has a high-energy capacity for its weight. Although it operates only at temperatures of about 250° C (580° F), its lightweight characteristics make it a contender for the battery of an electric car of the future. There are other high-energy batteries of this type also under development, one of which uses lithium and chlorine as the reactants.

After the initial surge of interest in fuel cells in the early 1960s, the only real applications to emerge have been for powering manned spacecraft. For longer missions to more distant planets, solar or nuclear generators are more appropriate, but regenerable or reversible fuel cells may be developed to store energy produced by these devices. For earth-orbit missions, and for lunar laboratory applications, hydrogen-oxygen cells are likely to remain in use.

The earlier promise of the fuel cell as a power source for automobiles is now recognized as a long-term prospect, and although research will continue toward this, it is likely to be at a low level. Increasing interest is being shown in static electrical generation from natural gas and liquid petroleum fuels for relatively small, local power stations, and the development of these for small communities will continue.

BIBLIOGRAPHY. "Batteries and Electric Cells, Primary," and "Batteries and Electric Cells, Secondary," in the *Kirk-Othmer Encyclopaedia of Chemical Technology*, 2nd ed., vol. 3, pp. 99–271 (1964).

Batteries: G.W. VINAL, *Primary Batteries* (1950), *Storage Batteries*, 4th ed. (1955); S.U. FALK and A.J. SALKIND, *Alkaline Storage Batteries* (1969); R. JASINSKI, *High Energy Batteries* (1967).

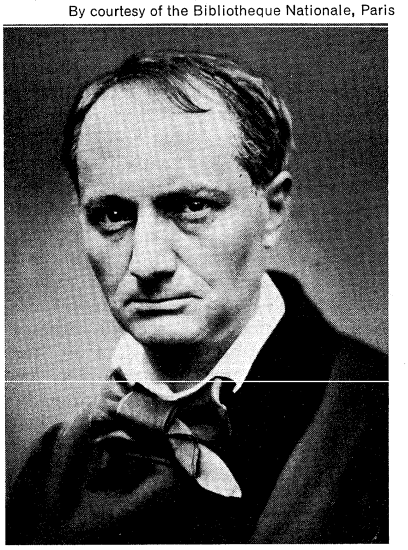
Fuel cells: K.R. WILLIAMS, *An Introduction to Fuel Cells* (1966); H.A. LIEBHAFSKY and E.J. CAIRNS, *Fuel Cells, and Fuel Batteries: A Guide to their Research and Development* (1968); D.P. GREGORY, *Fuel Cells* (1970).

(D.P.G.)

Baudelaire, Charles

Prosecuted for obscenity and blasphemy, and long after his death still identified in the public mind with depravity and vice, Baudelaire has become above all others of his age the poet of modern civilization, seeming to speak directly to the 20th century. Rejecting the posing of the Romantics, he revealed himself in his often introspective poetry as a seeker of God without religious beliefs, searching in every manifestation of life—the colour of a flower, the frown of a prostitute—for the true significance. Both as poet and critic he appeals to man's condition in the modern world; and modern, too, are his refusal to admit restriction in the poet's choice of theme and his assertion of the poetic power of symbols.

By courtesy of the Bibliotheque Nationale, Paris



Baudelaire, photograph by Étienne Carjat, 1863.

Early years

Charles-Pierre Baudelaire was born in Paris on April 9, 1821. His father, François Baudelaire, an elderly widower, in 1819 had married a young woman, without dowry, who had despaired of acquiring through marriage the luxury and security for which she longed. Baudelaire was their only child, and on him she lavished all the devotion of her ardent nature. His father, who had retired from his position in the civil service on a substantial pension, was a man of culture and an amateur painter of some merit. He taught his son, when only four or five, to appreciate the beauty of form and line, thus laying the foundation of the sureness of taste that was to make him one of the most interesting art critics of the 19th century.

François Baudelaire died in February 1827. In November 1828 his widow married Jacques Aupick, a soldier who already had risen in the ranks and was to become a general, an ambassador, and a senator. Anxious that his stepson should learn discipline, in 1832 he sent Charles as a boarder to the Collège Royal at Lyons. There, in spite of the strict military routine of the school, he seems to have been happy; and he won several prizes. He also began to show a feeling for language and to develop a literary style.

In 1836, when his stepfather was transferred to Paris, he was sent to the Lycée Louis-le-Grand. There, instead of fulfilling Aupick's claim that he would "bring honour to the establishment," he proved troublesome and undisciplined. To his masters he seemed an example of precocious depravity, adopting what they called "affectations unsuited to his age" and cultivating his gifts for outra-

geous paradox. He developed a tendency to moods of intense melancholy, and he also became aware that he was by nature solitary.

After passing his *baccalauréat* examinations in 1839, he rejected his stepfather's offer of a post in the diplomatic service and, to his mother's alarm, announced that he meant to live by writing. His chief wish was for freedom, leisure to read what he liked and to enjoy the student life of the Latin Quarter. Like many future writers, he enrolled as a law student, remaining at the École de Droit, nominally at least, until 1840. It was probably at this time that he became addicted to opium and hashish and contracted the venereal disease from which he was to die.

Visit to the East

In 1841, hoping to wean him from his friends in whose company he was leading a life of debauchery, his stepfather sent him on a voyage to India, intending him to stay there for at least two years. Baudelaire set sail on June 9 but, becoming bored, amused himself by scandalizing the other passengers by his unconventional behaviour, and at Mauritius, where the boat put in for repairs after a storm (in which Baudelaire had behaved with great courage), he declared that he would go no farther. Persuaded to go on to Réunion, he there insisted on taking the next boat home and arrived back in France in February 1842. The voyage, however, and his three weeks in Mauritius, had deepened and enriched his imagination and had given him a store of images on which he was to draw in his poetry. He never forgot this, his only experience of the East, but kept for it a nostalgic, mystical yearning that gives his poetry its characteristic quality. He had gone away a boy, still uncertain of himself and his future; he returned a man, his imagination on fire, determined as never before to become a poet.

On attaining his majority in April 1842, he gained control of the capital left him by his father and, leaving home, determined to satisfy his inherited taste for luxury. He spent his money recklessly on fine clothes and on rich furnishings for his apartment at the Hôtel Lauzun, in the Île Saint-Louis, and lived the life of a typical "dandy" of the period. Knowing nothing of business or finance, he regarded his inheritance as a fortune and soon fell prey to cheats and moneylenders, thus laying the foundation for the pile of debts that were to cripple him for the rest of his life. It was while living at the Hôtel Lauzun that his reputation for eccentricity, affectation, and immorality was confirmed; in his desire to shock he did not, however, differ from most of the poets and artists of the Paris of his time.

Jeanne Duval: the "Black Venus"

By 1844 Baudelaire had formed an association with the mulatto woman Jeanne Duval, who was to bring him much unhappiness. For a time he loved her passionately, and even at the end, when her cruelty, treachery, and stupidity had driven him to attempted suicide, he still felt in some ways attached to her. She inspired his first cycle of love poems, the "Black Venus"; and these are among the finest erotic poems in the French language.

During those early years of leisure and freedom from anxiety, Baudelaire was composing many—perhaps most—of the poems that were to form part of *Les Fleurs du mal,* his one collection, which comprised the Lesbian poems, the poems of revolt and decay, and the great erotic poems. At this time, too, he became acquainted with many artists, among them Delacroix and Courbet, and so acquired the knowledge of painting that was to give his art criticism much of its distinction and originality.

When within two years he had run through half his inheritance, his family early in 1844 obtained a decree by which the remainder of his capital was placed in trust, and he received the income in monthly installments. Baudelaire was wounded that his mother should have consented to a step that put an end to his freedom. In attempting to secure his future, his family had misguidedly prevented him from recovering his independence; still heavily in debt, he was unable, out of the £75 a year allowed him, to clear off his debts without borrowing.

This sudden change in Baudelaire's circumstances ended his life of luxury and carefree leisure; in the future he was to know only straitened means and eventually real poverty. He began to be uncertain of his own gifts, and his bitterness against his family was deepened by doubts whether they were perhaps right to try to prevent him from following a literary career. The melancholy he had known in adolescence returned, and what he called his mood of "spleen" became more frequent. It was at this time that the first of his great poems of spleen were written. Among his friends were many more unfortunate than himself, and he developed a sympathy for suffering humanity. Attracted by the revolutionary idealism of many of his friends, he took part in the February revolution of 1848 that resulted in the establishment of a republic.

The Salons of 1845 and 1846

Determined to prove that he could live by his pen, he had meanwhile become a professional writer. His first published work was a piece of art criticism, a review of the Salon of 1845 that reveals a perceptive and farsighted judgment and shows that he had already formulated a conception of what modern art should be. His "Salon de 1846" is a landmark in aesthetic criticism: no longer content to give an account of the exhibition, he puts forward independent and original theories and gives the first hint of his later concept of the *correspondances* between nature and art, claiming that painting, like music, has its own harmony, of light and shade, and that in nature, colour is melody. In 1845 and 1846 some of his poems were printed in avant-garde journals, to which he also contributed articles and reviews.

In 1847 he published his only novel, the autobiographical *La Fanfarlo.* Begun much earlier, it is of interest mainly for its analysis of his personality during the period when he was living in luxury at the Hôtel Lauzun.

What Baudelaire did between the June revolution of 1848, in which thousands of workmen were killed in Paris, and December 1849 is not known, nor is it certain why he was then in Dijon or how long he stayed there. By 1850 he was back in Paris, as destitute and unhappy as ever. His mother had refused to write to him until he showed signs of reformation; and although she intended to goad him into working regularly, his brief burst of activity had ended without conspicuous success, and he was further discouraged by her sternness. Many articles were planned but never written; many begun but never finished. But in these years of experience and suffering he had prepared himself for his great creative period. Spiritually, his nature had been enriched, and after Pres. Louis-Napoleon Bonaparte's coup d'etat of December 1851, which ended his active interest in politics, he was ready for the opening of his mature period.

Translation of Edgar Allan Poe

This began with his discovery early in 1852 of the writings of Edgar Allan Poe, and he set to work at once to translate them. His first article on Poe—the first in any foreign language—appeared in March and April in the *Revue de Paris* and was followed by several translations published in reviews, among them his only attempt at translating a poem, "The Raven." From 1852 to 1865 he was occupied in translating Poe and in writing critical articles on him. *Histoires extraordinaires* appeared in 1856; *Nouvelles Histoires extraordinaires* in 1857; *Aventures d'Arthur Gordon Pym* in 1858; *Eureka* in 1864; and *Histoires grotesques et sérieuses* in 1865. The first two had long critical introductions.

As translations these are, at their best, classics of French prose: Baudelaire's mother had been born in England the daughter of an émigré, and he had spoken English as a child. In Poe he found for the first time someone who belonged to his own spiritual family and who had already reached, independently, conclusions toward which he had been groping. Poe thus gave him confidence in his own aesthetic theories and ideals of poetry.

In April 1852 Baudelaire had left Jeanne Duval—though by no means, as it turned out, for good. He could not, however, live without the company of women; and seeking someone to love, he turned first to the actress Marie Daubrun, and when she rejected him, to Mme Apollonie-Aglaé Sabatier, a well-known beauty and former artist's model and friend of many artists and writers whom he had known for many years. She was the inspiration of his cycle of the "White Venus." In 1854 he renewed his association with Marie Daubrun, who in-

spired the cycle of the "Green-Eyed Venus." In many of the poems in these two cycles he reaches the highest peak of his art.

Baudelaire's growing reputation as Poe's translator and as art critic at last enabled him to publish some of his poems; and in June 1855 the *Revue des Deux Mondes*, the bastion of conservative Romanticism, ventured to print a selection of 18, submitted by Baudelaire as representative and chosen because they were original and startling in expression and themes. Their publication brought him notoriety, and he was widely accused of obscenity. In the spring of 1857, however, nine more poems appeared in *La Revue Française* and three in *L'Artiste;* and in June *Les Fleurs du mal* was published. But as a result, Baudelaire, the publisher, his friend Poulet-Malassis, and the printers were prosecuted; and in a famous trial for obscenity and blasphemy, they were found guilty of an offense against public morality and fined. Six poems were banned—a ban lifted only in 1949. Although a few readers understood and appreciated Baudelaire's intention and consummate artistry, for several generations *Les Fleurs du mal* remained a byword for depravity, morbidity, and obscenity. Baudelaire published a second edition in 1861, greatly enlarged and enhanced but omitting the banned poems, which were first republished in Belgium in 1866 in the collection *Les Épaves*. A third edition, further enlarged, was being prepared in 1866 when Baudelaire became paralyzed; it was published posthumously by his friend Charles Asselineau, although probably not as Baudelaire had planned it. It contains, however, six "Nouvelles Fleurs du mal," first published in 1866 in *Le Parnasse Contemporain*, as well as some poems that do not belong to the plan of the collection.

Publication of *Les Fleurs du mal*

The failure of *Les Fleurs du mal*, from which he had expected so much, was a bitter blow to Baudelaire, and the remaining years of his life were darkened by a growing sense of failure, disillusionment, and despair. His platonic relationship with Mme Sabatier had ended sadly, and Jeanne Duval, from whom he had finally parted in 1861, remained a constant burden and anxiety. Although some of his finest works were written in these years, few were published in book form. Some appeared in periodicals—his "Salon de 1859" in *La Revue Francaise*; "Richard Wagner et Tannhaüser à Paris" in *La Revue Européene* (1861); "Le Peintre de la vie moderne" (the draftsman Constantin Guys) in *Le Figaro* (1863); and his prose poems, intended to form part of the collection *Le Spleen de Paris*, in various papers. This last was a work of which Baudelaire was particularly fond and in which he had been engaged for many years—he was still working on it just before his final collapse. He had taken the idea from Aloysius Bertrand's *Gaspard de la nuit*, but the subject is that of his poems in verse of the same period, and in mood the work reflects the settled pessimism of the aging and deeply saddened Baudelaire. These poems in prose express even more poignantly than does *Les Fleurs du mal* his feeling for Paris, for the teeming modern city, and his compassion for the failures and outcasts in its streets.

In 1860 Poulet-Malassis published two studies of the effects of hashish and opium as *Les Paradis artificiels*, and in 1861, the second version of *Les Fleurs du mal*. In 1862 he was declared bankrupt; Baudelaire was involved in his publisher's failure and his financial difficulties became desperate. To escape his creditors, and in an attempt to dispose of the copyright of the works he had ready for publication, in 1864 he went on a lecture tour in Belgium. It proved a failure, and he was unsuccessful in negotiating a contract for his books. This was a bitter disappointment, for he had wished particularly to publish his critical works, in which he had defined his theory of aesthetics; he regarded his work as an organic whole and his critical prose therefore as important as his poetry. To appreciate his poetry fully it is necessary to understand his ideas of the nature of art. Each of his poems is a crystallization of his vision, and his criticism is a meditation on the nature of a work of art and on the principles that underlie it. Baudelaire believed that every great creative artist must in the end become also a critic; his criticism explains his poetry, and his poetry is an extension of his aesthetic theory.

Theories of art and poetry

In February 1866, while still in Belgium, at Namur, Baudelaire became seriously ill. Taken back to Paris, he died there in his mother's arms on August 31, 1867. Of the many invited to give orations at his funeral, only Asselineau and the poet Théodore de Banville accepted; but they were among his oldest friends.

Baudelaire died unrecognized, with many of his writings still unpublished and those that had been published out of print. Among poets, however, opinion soon began to change: the future leaders of the Symbolist movement who were at his funeral were already describing themselves as his followers. By the 20th century he had become widely recognized as one of the great French poets of the 19th century. His admirers even claimed that he revolutionized the sensibility and way of thinking and writing throughout western Europe, and that the formulation of his aesthetic theory marks a turning-point in the history of poetry and, indeed, in the history of art. For it was in this theory that the Symbolist movement found its source.

MAJOR WORKS

VERSE: *Les Fleurs du mal* (1857; 2nd enl. ed. 1861; definitive ed., 1868), notable among many translations are those by Roy Campbell, 1952; *Les Épaves* (1866), 23 poems, including six banned poems from *Les Fleurs du mal.*

PROSE: *La Fanfarlo* (1847), autobiographical *nouvelle*, published in the *Bulletin de la Société des Gens de Lettres*; *Les Paradis artificiels* (1860), essays, including a study of and translation from Thomas De Quincey's *Confessions of an English Opium-Eater; Petits Poèmes en prose* (1869), later entitled, as intended by Baudelaire, *Le Spleen de Paris;* among translations are those by Arthur Symons (1905), Michael Hamburger (1946), and Louise Varèse (1951); *Curiosités esthétiques and L'Art romantique* (1868), critical essays, including "Salon de 1845," "Salon de 1846," "Salon de 1859," "Richard Wagner et Tannhäuser à Paris," and "Eugène Delacroix," translations by Jonathan Mayne in *The Mirror of Art* (1955), and *The Painter of Modern Life, and Other Essays* (1964).

TRANSLATIONS (FROM EDGAR ALLAN POE): *Histoires extraordinaires* (1856); *Nouvelles Histoires extraordinaires* (1857); *Adventures d'Arthur Gordon Pym* (1858); *Eureka* (1864); *Histoires grotesques et sérieuses* (1865).

BIBLIOGRAPHY. The first edition of Baudelaire's *Oeuvres complètes*, 7 vol. (1868–70), was edited by C. ASSELINEAU, with a preface by THEOPHILE GAUTIER. It does not contain the *Journaux intimes*, other posthumous works, or the correspondence. Fragments of the *Journaux intimes* were published in *Le Livre* in September 1884, and the first complete edition in 1909. *Lettres 1841–1866* were published in 1905, *Lettres inédites à sa mère* in 1918, and *Dernières lettres inédites à sa mère* 1926. Some *Juvenilia* were published by JULES MOUQUET in 1932. A complete edition of his entire works, including *Juvenilia, Oeuvres posthumes* and *Correspondance générale*, was edited by JACQUES CREPET (1922–53); a revised version of this, under the editorship of GEORGES BLIN and CLAUDE PICHOIS, began appearing in 1968. The best one-volume edition is that by YVES LE DANTEC for the "Pléiade Series" (1950); a further volume contains translations from Poe. ROBERT KOPP'S 1969 edition of *Petits poèmes en prose* may be considered definitive. The best edition of *L'Art romantique* is by L.J. AUSTIN (1968). ASSELINEAU was the first to attempt a vindication of Baudelaire's character in *Charles Baudelaire, sa vie et son oeuvre* (1869); and in 1872 a group of friends published a collection entitled *Souvenirs, correspondances, bibliographie*, ed. by CHARLES COUSIN. The first complete biography was included in *Oeuvres posthumes et correspondances inédites*, which was published in 1887 and revised and enlarged in 1907 by EUGENE and JACQUES CREPET. For systematic studies of Baudelaire's vocabulary, see R.T. CARGO (ed.), *A Concordance to Baudelaire's "Les Fleurs du mal"* (1965); and *Concordances, index et relevés statistiques des "Fleurs du mal,"* issued by the Centre d'étude du vocabulaire français de la Faculté des Lettres de Besançon (1965). A short selection of the many works on Baudelaire and his writings follows: L.J. AUSTIN, *L'Univers poétique de Baudelaire* (1956), evaluates the significance of Baudelaire in the development of French poetry in the 19th century; the title of W.T. BANDY (comp.), *Baudelaire Judged by His Contemporaries (1845–1867)* (1933), is self-explanatory, as is that of *Baudelaire en 1848* (1946), which he wrote with JULES MOUQUET; GEORGES BLIN, *Le Sadisme de Baudelaire* (1948), is a collection of essays on several aspects of Baudelaire's

work, not just a study of sadism; LEON BOPP, *Psychologie des "Fleurs du mal,"* 4 vol. (1964–69), is an extremely detailed investigation, based upon close study of the texts of *Les Fleurs du mal*, of the workings of Baudelaire's mind; G.T. CLAPTON, *Baudelaire et De Quincey* (1931), considers Baudelaire in relation to an author who influenced him deeply; P. EMMANUEL, *Baudelaire* (1967), is a valuable study (in French) of Baudelaire's religious attitudes; ALISON FAIRLIE, *Baudelaire: les Fleurs du mal* (1960), is a short study intended for undergraduates: there is no better introduction to Baudelaire in English; ANDRE FERRAN, *L'Esthétique de Baudelaire* (1933), contains a comprehensive account of Baudelaire's philosophy of art; P. MANSELL JONES, *Baudelaire* (1952), is a sensitive and scholarly introductory work; F.W. LEAKEY, *Baudelaire and Nature* (1969), is an exhaustive study of an important aspect of Baudelaire, a work of synthesis as well as of analysis, with a full and up-to-date bibliography; C. MAURON, *Le Dernier Baudelaire* (1966), is a valuable contribution to the understanding of the *Petits poèmes en prose;* D.J. MOSSOP, *Baudelaire's Tragic Hero* (1964), is a study of the design of *Les Fleurs du mal* that is interpreted as a record of the inner struggles of the "poet-hero"; HENRI PEYRE, *Connaissance de Baudelaire* (1951), is a first-rate introduction to the problems of Baudelaire criticism; JEAN POMMIER, *La Mystique de Baudelaire* (1932), has become a standard general critical survey: his *Dans les chemins de Baudelaire* (1945) contains useful essays on Baudelaire's earlier works; FRANCOIS PORCHE, *Baudelaire, histoire d'une âme* (1945), adopts a biographical approach to Baudelaire's poetry; JEAN PREVOST, *Baudelaire* (1953), contains a fine account (in French) of the forces that influenced Baudelaire and discussions of themes and versification; PETER QUENNELL, *Baudelaire and the Symbolists*, 2nd rev. ed. (1954), is a study of Baudelaire and of his influence; MARCEL RUFF, *L'Esprit du mal et l'esthétique baudelairienne* (1955), is a wide-ranging study of all aspects of Baudelaire; JEAN-PAUL SARTRE, *Baudelaire*, (1947; Eng. trans., 1950), is a thought-provoking, though controversial, work of criticism that takes biography as its starting point; ENID STARKIE, *Baudelaire* (1957), is a critical biography that places Baudelaire's work in its cultural context (1957); MARTIN TURNELL, *Baudelaire* (1953), contains a sensitive and detailed commentary on *Les Fleurs du mal;* PAUL VALERY, *Situation de Baudelaire* (1924), is a penetrating essay on the originality of Baudelaire and of his influence on later 19th-century French poets; ROBERT VIVIER, *L'Originalité de Baudelaire*, 2nd ed. (1952), contains a comprehensive and systematic study of Baudelaire's style and poetic technique. Additional titles may be found in R.T. CARGO, *Baudelaire Criticism, 1950–1967: A Bibliography with Critical Commentary* (1968).

(En.S.)

Baybars I

Baybars I, ruler of Egypt and Syria from 1260 to 1277, was the most eminent of the Mamlūk, or "slave," sultans of Egypt and Syria. His enduring fame rests not only on the prowess he showed in his military campaigns, especially those against the crusaders and Mongols, but also on his internal administrative reforms, which stabilized the nascent Mamlūk state and allowed it to grow and flourish. He became a legend in his own lifetime, and an extensive folk literature, *Sīrat Baybars*, purporting to be his life story, is still popular in the Arabic-speaking world.

Al-Malik aẓ-Ẓāhir Rukn ad-Dīn Baybars was born probably in 1223 in the country of the Kipchak Turks on the northern shores of the Black Sea. After the Mongol invasion of their country about 1242, Baybars was one of a number of Kipchak Turks sold as slaves. Turkish-speaking slaves, who had become the military backbone of most Islāmic states, were highly prized, and eventually Baybars came into the possession of Sultan aṣ-Ṣāliḥ Najm ad-Dīn Ayyūb of the Ayyūbid dynasty of Egypt. Sent, like all the Sultan's newly acquired slaves, for military training to an island in the Nile, Baybars demonstrated outstanding military abilities. Upon his graduation and emancipation, he was appointed commander of a group of the Sultan's bodyguard.

Early military successes

Baybars gained his first major military victory as commander of the Ayyūbid army at the city of al-Manṣūrah in February 1250 against the crusaders' army led by Louis IX of France, who was captured and later released for a large ransom. Filled with a sense of their military strength and growing importance in Egypt, a group of Mamlūk officers, led by Baybars, in the same year murdered the new sultan, Tūrān Shāh. The death of the last Ayyūbid sultan was followed by a period of confusion that continued throughout the first years of the Mamlūk sultanate.

Having angered the first Mamlūk sultan, Aybak, Baybars fled with other Mamlūk leaders to Syria, and stayed there until 1260, when they were welcomed back to Egypt by the third sultan, al-Muẓaffar Sayf ad-Dīn Quṭuz. He restored them to their place in the army and conferred a village upon Baybars.

Within a few months of Baybars' arrival, in September 1260, the Mamlūk troops defeated a Mongol army near Nābulus in Palestine. Baybars distinguished himself as the leader of the vanguard, and many Mongol leaders were slain on the field.

For his military achievement, Baybars expected to be rewarded with the town of Aleppo; but Sultan Quṭuz disappointed him. On the way home through Syria, Baybars approached Quṭuz and asked him for the gift of a captive Mongol girl. The Sultan agreed, and Baybars kissed his hand. On this prearranged signal the Mamlūks fell upon Quṭuz, while Baybars stabbed him in the neck with a sword. Baybars seized the throne to become the fourth Mamlūk sultan.

Military career as sultan

Baybars' ambition was to emulate Saladin, the founder of the Ayyūbid dynasty, in the holy war against the crusaders in Syria. As soon as he was acknowledged as sultan, Baybars set about consolidating and strengthening his military position. He rebuilt all the Syrian citadels and fortresses that had been destroyed by the Mongols and built new arsenals, warships, and cargo vessels. To achieve unity of command against the crusaders, Baybars united Muslim Syria and Egypt into a single state. He seized three important towns from the Ayyūbid princes, thus ending their rule in Syria. From 1265 to 1271 Baybars conducted almost annual raids against the crusaders. In 1265 he received the surrender of Arsūf from the Knights Hospitallers. He occupied ʿAtlit and Haifa, and in July 1266 he received the town of Safed from the Knights Templar garrison after a heavy siege. Two years later, Baybars turned toward Jaffa, which he captured without resistance. The most important town taken by Baybars was Antioch (May 1268), which was followed by a number of minor Frankish fortresses. His seizure of additional strongholds in 1271 sealed the crusaders' fate; they were never able to recover from their territorial losses. Baybars' campaigns made possible the final victories won by his successors during the next decades.

Baybars' permanent goal was to contain the continued Mongol attacks on Syria from both north and east that threatened the very heart of the Islāmic east. During the 17 years of his reign, he engaged the Mongols of Persia in nine battles. Within Syria, Baybars dealt with the Assassins, a fanatical Islāmic sect. After seizing their major strongholds between 1271 and 1273, he wiped out the Syrian members of the group.

Baybars also took the offensive against the Christian Armenians, who were allies of the Mongols, devastating their lands and plundering their major cities. In 1276, having defeated the Seljuq troops and their Mongol allies, he personally seized Caesarea (modern Kayseri in Turkey) in Cappadocia. To secure Egypt on the south and west, Baybars sent military expeditions into Nubia and Libya, taking personal command in 15 campaigns and often endangering his life.

Foreign relations

In the interest of good diplomatic relations with the Byzantine Empire, Baybars sent envoys to the court of Michael VIII Palaeologus in Constantinople. The Byzantine sovereign thereupon ordered the restoration of the ancient mosque and permitted the Egyptian merchants and ambassadors to sail through the Hellespont and Bosporus. One of Baybars' principal goals during his reign was to acquire more Turkish slaves to be used in the Mamlūk army; another was to contract an alliance with the Mongols of the Golden Horde in South Russia against the Mongols of Persia. In 1261 Baybars sent an ambassador to the Sicilian king Manfred. Other embassies to Italy followed, and in 1264 Charles of An-

jou, later king of Naples and Sicily, sent an embassy with letters and gifts to Cairo, a remarkable testimony to Baybars' strength and influence. Baybars was also able to sign commercial treaties with such distant sovereigns as James I of Aragon and Alfonso X of León and Castile.

In a brilliant political move Baybars invited a fugitive descendant of the ʿAbbāsid dynasty of Baghdad to Cairo and established him as caliph—head of the Muslim community—in 1261. Baybars wished to legitimize his sultanate and to give pre-eminence to his rule in the Muslim world. The ʿAbbāsid caliphs in Cairo had no practical power in the Mamlūk state, however.

Internal policy

Baybars was, moreover, more than a military leader or a diplomatic politician. He built canals, improved harbours, and established a regular and fast postal service between Cairo and Damascus, one that required only four days. He built the great mosque and the school bearing his name in Cairo. He was also the first ruler in Egypt to appoint chief justices representing the four main schools of Islāmic law.

A sportsman as well as a warrior, Baybars was fond of hunting, polo, jousting, and archery. He was also a strict Muslim, a generous almsgiver, and watchful of the morals of his subjects—he issued a prohibition against the use of wine in 1271.

He died in Damascus on July 1, 1277, after drinking a cup of poison intended for someone else, and was buried in Damascus under the dome of the present aẓ-Ẓāhirīyah Library, which he had established.

BIBLIOGRAPHY. Two important primary sources for the career of Baybars I are the biographies of MUHYI AL-DIN IBN ʿABD AL-ZAHIR (d. 1292) and MUHAMMAD IBN SHADDAD (d. 1285). Unfortunately neither of these is fully extant. Part of the life of Baybars by IBN ʿABD AL-ZAHIR has been published by S.F. SADEQUE, *Baybars I Of Egypt* (1956). See also E.M. QUATREMERE (trans. and annot.), *Histoire des sultans mamlouks de l'Égypte*, vol. 1 (1837); S. LANE-POOLE, *A History of Egypt in the Middle Ages*, pp. 242–275 (1901, reprinted 1968); and G. WIET, "Baybars I," in the *Encyclopaedia of Islam*, new ed., vol. 1 (1960).

(H.Ra.)

Bayern

Bayern (Bavaria in English conventional usage) is the largest *Land*, or state, of the Federal Republic of Germany, with an area of 27,238 square miles (70,547 square kilometres). While it ranks second to Nordrhein-Westfalen in population, with about 10,500,000 by the early 1970s, its density of 385 persons per square mile is rather low in comparison to the rest of Germany. Bayern comprises the eastern part of south Germany and is bounded on the west by the *Länder* of Baden-Württemberg and Hessen, on the north by the German Democratic Republic, on the east by Czechoslovakia, and on the south and southeast by Austria. Munich (München; *q.v.*) is the capital.

History. *The Celts and Romans.* The earliest known inhabitants in the area of present-day Bayern were the Celts. They developed a considerable culture, as evidenced in their aristocratic governing structure and their carefully laid-out cities. In the last decade before Christ's birth they were pressed between Teutonic tribes in the north and the Romans in the south. The Romans divided the southern part into Raetia and Noricum and built fortifications along the northern boundary to keep out the Teutons. Flourishing Roman colonies arose in the south—Augsburg, Kempten, Regensburg, and Passau—that retained much of their Roman character in later times.

German invasions

The Middle Ages. The Romans were overcome in the 5th century by repeated Germanic attacks. The lands were eventually settled by tribes from the Elbe area and from Bohemia, Moravia, and Hungary, who mixed with the remaining Celts and Romans. The tribe that gave the territory its name was the Baiuvarii (Bavarians), who settled in the south between AD 500 and 800. The southwest belonged to the Alemanni, while northern Bavaria was in the hands of the Franks and Thuringians.

In the 7th and 8th centuries Bavaria was Christianized by Irish and Scottish monks (St. Boniface, St. Korbinian, St. Emmeram, and St. Rupert). From about 555 to 758 the Bavarians were ruled by Frankish dukes of the Agilolfing family. The last of the family, Tassilo III, was deposed by Charlemagne, who finally incorporated Bavaria into the Carolingian Empire.

After the partitioning of the empire in 817, the Duchy of Bavaria became a central part of the territory of the East Franks, with its capital at Regensburg. At this time Bavaria experienced a cultural upsurge, stemming primarily from the monasteries. In 1180 the Holy Roman emperor Frederick I Barbarossa gave Bavaria to the count palatine Otto of Wittelsbach, whose family ruled it until 1918. At first the Wittelsbachs possessed only the southeastern part of present-day Bavaria, the rest being fragmented into numerous imperial cities, monastic holdings, and family domains. Territorial changes were frequent. In 1214 the Palatinate was added (which remained Bavarian until 1945). In the 14th and 15th centuries, the power of the dukes was notably weakened by political divisions, until finally Albert IV (1467–1508) reunited Bavaria, making Munich the capital. The first Bavarian university was founded in 1472 in Ingolstadt.

William IV to Bismarck. William IV (1508–50) opposed the Protestant Reformation; under his successor, Albert V (1550–79), Bavaria became a strictly Catholic country. Maximilian I (1597–1651) fought on the side of the Habsburgs in the Thirty Years' War (1618–48) and by his leadership increased Bavaria's prestige, gaining territorial accessions and attaining for himself the title of elector. Bavaria suffered greatly in the later phases of the war from the incursions of Swedish and French armies. Territorial conflict continued throughout the 18th century as Bavaria was ravaged by the wars of the Spanish Succession, the Austrian Succession, and the Bavarian Succession. Despite this, the era was a glorious one culturally.

French occupation

In 1800, French Revolutionary armies occupied Munich. In the following year Bavaria became an ally of France and was able to expand its territories at the expense of Austria, acquiring by 1806 approximately the boundaries it now has. In 1813, shortly before the Battle of Leipzig, Bavaria rejected Napoleon and in 1815 joined the Germanic Confederation against him. The reigns of kings Maximilian I and Louis I saw the consolidation of the country and the achievement of many reforms: the first constitution, a parliament, municipal autonomy, and tax reform. Munich became a flourishing centre of learning and the arts. In 1848, however, Louis was forced to abdicate.

His successor, Maximilian II (ruled 1848–64), proceeded with domestic reform programs while also promoting the arts and learning. Politically, he sought to avoid being dominated by Prussia in a Germany that excluded Austria; he favoured union of the many small German states in order to withstand the force of Prussia's demands. His successor, Louis II (ruled 1864–86), is remembered for his romantic castles and his enthusiasm for the music of Richard Wagner. He refused Bismarck's proposal to incorporate Bavaria into a German domain under Prussian leadership, siding with Austria in the Prussian–Austrian War of 1866. The quick victory of the Prussians and the moderation of their policies led Bavaria to join Prussia in the Franco-German War of 1870 and afterward to share in the establishment of a German empire under Wilhelm I, King of Prussia.

After 1871. In the German constitution of 1871 Bavaria retained a few elements of sovereignty in its diplomatic service, military administration, postal service, and railways but otherwise cast its lot with the German empire. At the end of World War I, King Louis III had to abdicate, and Bavaria became a republic. The five years thereafter were filled with constant unrest: 1919 saw the murder of Kurt Eisner, the Socialist leader, and the establishment of a short-lived soviet republic; in 1920 and 1921 there were right-wing coups; and in 1923 Adolf Hitler and Gen. Erich Ludendorff attempted their unsuccessful *Putsch* in Munich.

After World War II Bavaria became part of the American zone of occupation. The Palatinate was detached and joined to the new Rhineland-Palatinate state. Under the Basic Law of Germany of 1948, Bavaria became a *Land* (state) of the Federal Republic. It retains, however, a certain uniqueness based upon its size, its long history as an independent state, and its great culture.

Geography. Bayern is a country of high plateaus and medium-sized mountains. In the northwest are the wooded sandstone hills of the Spessart; in the north are basalt knolls and high plateaus. The northwest is drained by the Main River, which flows into the Rhine. Moving southeast, the topography varies from the stratified land formations of Swabia-Franconia to shell limestone and red marl, the hill country of the Franconian–Rednitz Basin, and the limestone mountains of the Franconian Jura along the Danube, which divides Bayern north and south. On the eastern edge of Bayern, adjoining Czechoslovakia, is the Böhmerwald (Bohemian Woods) and, in the north, the Frankenwald (Franconian Forest). South of the Danube is a plateau upon which lies the capital, Munich, and beyond it the Bavarian Alps. Bayern's share of the Alps consists of wooded heights of several thousand feet, behind which rise steep ridges and high plateaus (in the west, the Allgäuer Alps; in the east, the Alps of Berchtesgaden). They reach their highest peak in the 9,718-foot (2,962-metre) Zugspitze in Germany's Wettersteingebirge (Wetterstein Range).

Mountain peaks

Bayern's climate is determined by its southern position in Europe along with its relatively high altitude and its distance from the sea. These give it a continental climate that is harsh for middle Europe, although there are some exceptions, such as the Lower Main Valley. Average temperatures in January vary from around freezing in Würzburg and Nürnberg, to 28° F (−2° C) in Regensburg and Munich, and 12° F (−11° C) on the Zugspitze. The July averages are in the middle 60s F except on the Zugspitze, where temperatures are close to freezing.

Bayern's ethnic characteristics are a reflection of its traditions and political history. The southeast is inhabited by an old Bavarian stock, the southwest by people of Bavarian–Swabian descent, and the north by descendants of the Franks. Traditional differences are still visible in their villages. The Franks built large village clusters and laid out their farms in narrow strips. The houses are partly sandstone, partly half-timbered. Row houses with paved floors appear in some areas. In old Bayern and Swabia there are both village clusters and one-street villages; most of the houses have wooden floors. The cities show even more marked differences. In the Swabian and, particularly, the Frankish areas, religious and secular landholders established a large number of towns, most of which remained small and were referred to as dwarf towns. These medieval towns were built compactly within protective walls. The churches, public buildings, and homes were lavishly decorated; the examples that remain are a constant delight to the tourist, notably in Rothenburg, Nördlingen, Dinkelsbühl, and sections of Nürnberg and Regensburg.

Recent industrialization

Industrialization and urbanization, which came to Bayern later than to the rest of Germany, gradually obliterated most of these traditional elements. While nearly half of Bayern's inhabitants still live in places of less than 5,000 population, more than 20 percent live in towns of 100,000 or more. Munich, the largest city, had 1,350,000 inhabitants in 1970; greater Munich, with its suburbs, had 1,800,000. As the third largest city in Germany, Munich is not only the capital of Bayern but a focal point of industry and trade. Its industries include electronics, precision instruments, aircraft, and various kinds of machinery. It plays a leading role in banking, insurance, and gastronomy. It has the largest German university and is also an art centre with excellent museums and galleries. In 1972 the Olympic Games were held there.

Nürnberg is the second largest Bavarian city, with 480,000 inhabitants in 1970; together with its neighbouring cities of Fürth and Erlangen it constitutes a populous industrial complex in Middle Franconia of 800,000 inhabitants. The next three Bavarian cities of considerable size are Augsburg (255,000), Regensburg (130,000), and Würzburg (120,000). Other significant cities include Ingolstadt, Bamberg, Bayreuth, Schweinfurt, Aschaffenburg, Hof, and Landshut.

Population. Bayern's population in 1818 was about 3,000,000; in 1925, about 6,000,000; and, in 1970, about 10,500,000. The Bavarian people consisted originally of descendants of the three tribes already mentioned: Baiuvarii, Franks, and Alemanni. After World War II there was an influx of refugees from the Sudetenland and eastern Europe, where many Germans had lived for centuries. About 20 percent of Bayern's population in 1970 was composed of these refugees. Beginning in the 1960s the industrial areas received large numbers of migrant workers from southern Europe, totalling approximately 400,000 in 1970.

Great changes took place in the religious composition of the population after the war, with a heavy incursion of Protestants. In 1970 about 70 percent of the Bavarians were Roman Catholics, having bishoprics in Munich–Freising, Augsburg, Regensburg, Passau, Bamberg, Eichstätt, and Würzburg. About 25 percent were of Evangelical Lutheran faith, with centres in Munich, Augsburg, Regensburg, Nürnberg, Bayreuth, and Ansbach. The proportion of the population engaged in farming declined from 52 percent in 1882 to 42 percent in 1900, 28 percent in 1939, and 15 percent in 1970. About 45 percent work in industry and 40 percent in services.

The economy. More than half of the state's gross output in 1969 consisted of industrial and handicraft products. Trade, transportation, and services accounted for 40 percent, and agriculture and forestry for 5 percent.

Agriculture. Farms have grown larger and employ fewer hands. There has been a trend to specialization in crops and to production for specific markets. Most farms are family operated and run in size from 38 to 75 acres. Smaller units are also common, and there are a few large landholdings manned by foreign workers.

About 33 percent of Bayern is covered by forests—mainly spruce and pine—and efforts are being made to extend forest areas where the soil is poor. A little over half of Bayern's surface is cultivable, although a quarter of it is given over to gardens and meadows. Rye, wheat, and barley take up about 60 percent of the farmland; potatoes, sugar beet, and other vegetables use 20 percent; and the rest is given to hay, hops, and vineyards. The most important arable land is in Lower Bavaria (Dungau) and in the Frankish district. Meadows are found primarily in the Upper Bavarian–Swabian forelands of the Alps, where cattle are pastured. Allgäu is the centre of German butter and cheese production. Pigs and poultry are found scattered throughout Bayern.

Principal crops

Industry. The development of Bavarian industry was at first hampered by a lack of minerals and poor transportation. Locally there were only brown coal, porcelain clay, a small amount of iron ore in the Upper Palatinate, and a deposit of salt near Berchtesgaden. The mining of coal at the edge of the Alps was discontinued in 1971.

The natural disadvantages have been overcome by the development of hydroelectric power and more recently by the use of oil piped in from the Mediterranean ports of Marseilles, Genoa, and Trieste. Improved transportation has encouraged industries that manufacture and finish quality materials.

After World War II the government made great efforts to attract expanding industries, with the result that Bayern attained a higher rate of industrial growth than the rest of Germany. In 1969 the leading industries were electronics, machinery, chemicals, textiles, automobiles, clothing, and foodstuffs; the range of industrial effort was too broad to give any industry a unique predominance. A wide variety of industries are located in Munich, the Nürnberg area, and Augsburg; important but more specialized centres are Ingolstadt, Schweinfurt, and other smaller cities. The textile and ceramics industry is scattered through northeastern Franconia.

Leading industries

Service trades. Trade and commerce resemble that of the rest of Germany, with the exception of the tourist

trade, which has assumed particular importance in the Bavarian Alps. Bayern had about 400,000 hotel beds in 1969, and three-quarters of them were in the Alpine area.

Transportation. Begun in 1835, the Bavarian railroad system was heavily developed in the second half of the 19th century. All main lines are either electrified or use diesel engines. In 1969 there were 4,779 miles of track. There are first-class highways from Munich to Frankfurt, to Berlin, and to Salzburg. Several other highways were under construction in 1971 (Munich–Lindau and Nürnberg–Passau); others were in the planning stages.

The most important waterway is the Main River, which is navigable as far as Bamberg. The Danube carries vessels as far upstream as Kelheim. A canal linking the two rivers was under construction in the early 1970s. When completed it would carry 1,350-ton ships from the North Sea and the Rhine River to the Black Sea. Bayern has airports at Munich–Riem and Nürnberg.

The Parliament

Government. Under its constitution of 1946, Bayern is a free state with democratic parliamentary institutions. The voters are represented directly in a lower house, the Landtag, elected every four years. The Langtag chooses a minister-president and a Cabinet. There is also a Senate composed of representatives of economic, social, cultural, and religious organizations.

From 1957 onward Bayern was controlled by the Christian-Social Union (CSU), which in 1970 won 124 of the 204 seats in the Landtag. The Social Democratic Party and the Free Democratic Party comprised the opposition, the former with 70 seats in 1970 and the latter with 10.

Bayern is divided into seven administrative regions: Upper Bavaria; Lower Bavaria; Upper Palatinate; Upper, Middle, and Lower Franconia; and Swabia. Each consists of independent cities and districts, the latter being further divided into communities with mayors and councillors. Bayern no longer possesses troops, a diplomatic corps, or an independent postal system. It has its own school system, law courts, and health programs, but these are fully integrated into those of the Federal Republic. In all areas, the Bavarian constitution and laws are subordinate to those of the Federal Republic of Germany.

Cultural life. The educational system resembles that of the Federal Republic generally: elementary school for those from 6 to 10 years of age, followed by five-, six-, or nine-year secondary schools (the latter called *Gymnasia*). There were six universities in 1971: two in Munich and one each in Erlangen–Nürnberg, Würzburg, Regensburg, and Augsburg. Plans were under way for new ones in Bayreuth and Passau.

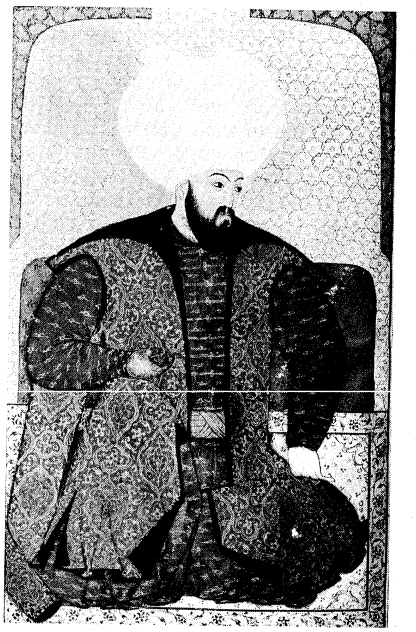
Bayern has a rich heritage in literature, music, and other arts, which have been tenderly nurtured by state, municipal, and private efforts. There are theatres in all the larger cities (including three state theatres in Munich alone), as well as numerous orchestras, museums, and art galleries. The libraries are excellent, particularly the Bavarian State Library in Munich. The Bavarian radio (broadcasting both radio and television programs) is a public state-controlled system with headquarters in Munich. Efforts are being made both locally and nationally to maintain and develop Bayern's strong cultural individuality, in the hope that Munich may become an important cultural centre for all of Germany.

BIBLIOGRAPHY. Statistical information on Bavaria is conveniently obtainable in the *Statistisches Jahrbuch für Bayern* (annual). Recent census material is contained in *Beiträge zur Statistik Bayerns* (1971). The most recent and comprehensive treatment of Bavarian history is MAX SPINDLER, *Handbuch der bayerischen Geschichte*, 4 vol. (1967–). An historical atlas is *Bayerischer Geschichtsatlas* (1969). A classic geographical work on Bavaria is ROBERT GRADMANN, *Süddeutschland*, 2 vol. (1931). Other sources for climate and topography include OSKAR KUHN, *Geologie von Bayern* (1954); K. ROCZNIK, *Wetter und Klima in Bayern* (1960); K. KNOCH (ed.), *Klima-Atlas von Bayern* (1952); and *Topographischer Atlas Bayern* (1968), which has maps of selected landscapes and towns with geographical essays. For the economy, see K. SCHREYER, *Bayern: Ein Industriestaat* (1969), on the postwar development of Bavarian industry.

(K.Ru./R.E.Pa.)

Bayezid II

Bayezid II, Ottoman sultan from 1481 to 1512, extended and consolidated Ottoman rule in the Balkans, in Asia Minor, and in the eastern Mediterranean. He also had to meet the danger from the Ṣafavīd dynasty of Persia (Iran).



Bayezid II, miniature by an unknown Ottoman artist, *c.* 1580. In the Topkapi Museum, Istanbul.

Bayezid II was born, probably in 1447, the elder son of the sultan Mehmed II, the conqueror of Constantinople. On the death of his father in 1481, his brother Cem contested the succession. Bayezid, however, supported by a strong faction of court officials at Constantinople, succeeded in taking the throne. Cem eventually sought refuge with the Knights of Saint John at Rhodes and remained a captive in European hands until his death in 1495.

Under the new reign an immediate reaction set in against some of the policies of Mehmed II. Influenced by the *'ulamā'*, interpreters of the law of Islām, and by the great officials aligned with them, Bayezid restored the Muslim properties dedicated to religious and charitable purposes that Sultan Mehmed had taken over for the state. Bayezid also rejected his father's marked pro-European orientation by such acts as removing from the imperial palace the paintings that Italian artists had executed for Mehmed II.

Expansion of the empire

At the same time, Bayezid II continued the territorial consolidation his father had begun. Hercegovina, in the Balkans, was brought under direct Ottoman control in 1483. The occupation, in 1484, of two fortresses on the estuaries of the Danube and the Dniester strengthened the hold of the Ottomans over the land route to the Crimea, where the Khan of the Krim Tatars had been, in name at least, a vassal of the sultan since 1475. The war of 1499–1503 directed against the Venetian empire in the Levant and in the Balkans carried the process of consolidation still further. It resulted in the Ottoman conquest of Venetian strongholds in Morea (Peloponnesus) and on the Adriatic shore—a triumph amply justifying the program of naval construction that Bayezid had approved in the years before the beginning of the war.

With the expansion of his rule over much of Asia Minor, Bayezid had earlier come into conflict with the Mamlūk sultanate of Egypt and Syria, each side striving to dominate the ill-defined border zones dividing them and to maintain under effective control the small principalities established there. While a Turkish fleet had sufficed to dismantle a large part of Venice's empire, Bayezid, fear-

ing that an alliance of Christian powers using his brother Cem might be formed against him, committed only a modest force against the Mamlūks. The long, drawn-out land war ended in a stalemate.

Suppression of rebellion in Asia Minor

More formidable still was the situation that arose in the lands to the east of Asia Minor. In 1499 the adherents of the Ṣafavīds, a heretical order of Islām, had set out to establish in Persia a powerful regime under their master Esmāʿīl I. The religious teaching of the Ṣafavīds had met with great success among the nomadic Turkmen tribes of Asia Minor, whose warriors formed the main element in the armies of Shah Esmāʿīl (or Ismāʿīl). It was evident that the propaganda of the Ṣafavīds, if allowed to continue without hindrance, might well undermine Ottoman rule within the Asiatic lands. The danger was underlined in 1511, when the adherents of the Shah rose in rebellion against the Ottomans in Asia Minor.

At this same time a dispute over the succession broke out between Bayezid's sons. One of them, Selim, the governor of Trebizond, went to the Crimea in 1511, secured aid there from the Tatar khan, and then crossed the Danube into the Balkans. Defeated in battle against Bayezid, Selim fled to the Crimea. Meanwhile, the Ṣafavīd rebellion had been put down; and Ahmed, another son, who had shared in the victory, marched toward Constantinople. Failing to gain the support of the Janissaries (elite military guards), he turned back to bring most of Asia Minor under his control. Bayezid, fearing that Ahmed might seek assistance from Shah Esmāʿīl and unable to resist pressures from some of his advisers and from the corps of Janissaries, who favoured Selim, he recalled Selim from the Crimea and abdicated (April 1512) in his favour. He died the following month.

Assessment

Bayezid II, known to his people as Adlî ("the Just") was a pious Muslim, strict in his observance of the precepts of the Qur'ān and the Islāmic law. During his reign, much of the state revenue was devoted to the building of mosques, colleges, hospitals, and bridges. He also supported jurists, scholars, and poets, both within and outside the Ottoman Empire. In temperament "molto melancolico, superstizioso e ostinato" ("very melancholic, superstitious, and stubborn"), in the words (1503) of the Venetian ambassador, Bayezid was interested in philosophical and cosmographical studies. He was an affectionate father, kind and forgiving toward his sons as well as toward the men who served him.

BIBLIOGRAPHY. J. VON HAMMER-PURGSTALL, *Histoire de L'Empire Ottoman*, vol. 3–4 (1835–43), a fundamental work based mainly on sources in Ottoman Turkish; S.N. FISHER, *The Foreign Relations of Turkey, 1481–1512* (1948), a most useful monograph based extensively on Venetian source materials; D.M. VAUGHAN, *Europe and the Turk: A Pattern of Alliances, 1350–1700* (1954); the *New Cambridge Modern History*, vol. 1, pp. 395–419 (1957); and the *Cambridge History of Islam*, vol. 1, pp. 308–314 (1970), three works that serve to illustrate the present attitude of historical scholarship towards Bayezid II; the article "Bāyazīd II," in the *Encyclopaedia of Islam*, 2nd ed., vol. 1, pp. 1119–1121 (1960), the standard work of reference for orientalist scholars (with bibliography).

(V.J.P.)

Beaches

Beaches are the sediments that accumulate along the sea or lake shores, the configuration and contours of which depend on the action of marine processes, the kinds of sediment involved, and the rate of delivery of this sediment. There are three different kinds of beaches. The first occurs as a sediment strip bordering a rocky or cliffy coast; the second is the outer margin of a plain of marine or fluvial accumulation (free beaches); and the third, of fairly peculiar character, consists of the narrow sediment barriers stretching for dozens or even hundreds of kilometres parallel to the general direction of the coast. These barriers separate lagoons (*q.v.*) from the open sea and generally are dissected by some tidal inlets. Certain sediment forelands, such as spits, points, and tombolos (connecting an island with a mainland), also are called beaches occasionally.

Beach characteristics. The upper limit of the active beach is the swash line reached by highest sea level during big storms. The lower beach margin is beneath the water surface and can be determined only if there is a definite border present between the sediment layer and the naked surface of the rocky bench. If the sediment cover extends into deep water, the lowest beach margin may be defined as the line where the strongest waves no longer sort and move the sand. It occurs approximately at a depth equal to one-third the wave length or ten times the wave height.

Another definition of the lower beach limit has been suggested by V.P. Zenkovich and V. Longuinov, namely, the transitional zone in which the biggest waves break completely and form the swash flow of reversed directions. It is the zone in which the bottom is briefly exposed prior to occurrence of the next breaker and in which sediment drifts both shoreward and seaward. This definition is more exact, but it is primarily applicable to tideless seas or lake shores.

Broad plains of marine accumulation form in many areas in response to shore aggradation. They are easily identified by the occurrence of a series of well-developed shore ridges termed berms. The maximum width of such a plain is about 30 kilometres (*e.g.*, the southwestern coast of Mexico). These plains are often termed beach plains because their surface was developed by the usual processes of beach formation. Because this surface was subsequently influenced by other relief-forming factors, however, it seems incorrect to call them beach plains.

The total width of exposed beaches varies from several metres along small seas or lake shores to 60 metres (for pebble beaches) or even several kilometres (for sand beaches). There are also the beaches of intermittent character: shingle (pebbles and cobbles) at the upper part and sand at the lower part.

The profile of an active beach varies greatly. Its form and dimensions depend on a number of factors, such as wave parameters, tide height, and sediment composition and distribution. The following, however, constitute some of the profile elements that commonly occur. At the upper part, above high sea level, a beach terrace is located, and there may be a series of beach ridges or berms created by the waves of a previous major storm. This terrace surface is inclined seaward. The next element is a steeper, frontal beach slope or face, and beneath it a low-tide terrace may be developed. If the tides are high enough (over two metres), the frontal slope may be more than one kilometre in width in regions with abundant sand and a shallow bottom. In some areas the low-tide terrace terminates with another inclined shoreface, if the nearshore sea zone is rather deep. Finally, one or several parallel, submarine, long-shore bars with intervening troughs (Figure 1) may exist along sandy shores; if present, these bars constitute the last profile element.

After Fig. 33 (p. 82), "Terminology Used in Describing Beaches," from *Submarine Geology* by Francis P. Shepard (Harper & Row, 1948)

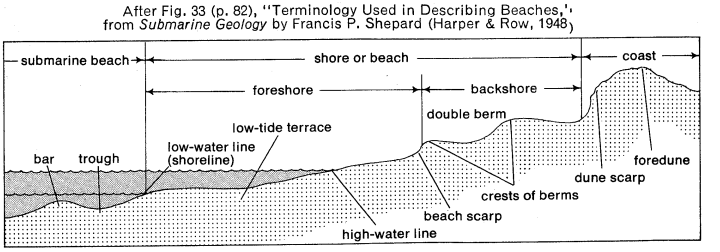


Figure 1: Main elements of the beach profile in a tidal sea.

Some minor relief forms are usually present on the surface of sand beaches. These include oscillation ripples, swash or rill furrows, and the well-known beach cusps (concave seaward) at the beach margin.

Given the established system of strong waves normal to the shoreline, submarine bars are sometimes dismembered and are converted into large crescent elements convex seaward. These relief forms reflect the existence of large water eddies with vertical axes, which form as a result of the ebb and flow of the water. Often, however,

the water outflow proceeds in the form of linear rip currents. These may be so strong that they cause erosion of deep channels in the submarine slopes and sometimes dissect the entire system of submarine bars.

Sediment transport. The declivity of beaches depends on sediment composition as well as on wave action. The coarser the sediment, the steeper and narrower the beach. This fact results from the permeability of the sediment. Part of the swash waters penetrate the sediment, thus decreasing the transporting power of the backwash. Other things being equal, however, wave size determines the height and width of a beach.

Bottom currents take sedimentary particles and move them to and fro in the wave direction; the particles also may shift up or down the slope. The speed and direction of this shifting depends on the hydraulic coarseness of the grains. Thus, sediment movement occurs on the exposed and submarine parts of the beach; it washes off some zones and accumulates elsewhere.

Effect of waves and wave parameters

The formation of the sloping beach partly exposed above sea level results from the unequal speeds and duration of the direct and reverse bottom currents under deformed waves (Cornaglia's principle). If these movements were identical, the beaches would not exist at all; all of the sediment would be carried out to the sea bottom. Measurements of this phenomenon were first carried out by V. Longuinov (1963) on the Black Sea shore.

The beach profile keenly reflects changes in wave parameters and changes in the character of transverse sediment drift. Considerable alteration may be observed in the course of a few hours, and the beach profile may be built anew. Moreover, there are general differences between summer and winter beach profiles. In winter, when there are strong storms, the exposed beach becomes narrower and in some places may completely disappear. In spring the beach begins to aggrade, and in summer it becomes maximally developed. Temporary beach changes may be as great as 3 metres in height and 50 metres in width.

In the opinion of most authorities, the accumulation of material and beach retreat both depend on wave steepness. Destructive waves of low steepness (for which the ratio of wave height to length, H/L, is less than 0.03) wash out the beach and transport part of the sediment to the sea bottom, whereas constructive waves of greater steepness (H/L greater than 0.03) throw the sediment back onto the beach, thus aggrading it.

This regularity, however, has not been corroborated in the case of relatively tideless seas, such as the Black and the Baltic. The main role in these areas is ascribed to wave energy and, consequently, to wave height because the expression for wave energy contains wave height raised to the second power and the length to the first power (see WATER WAVES).

There are some beaches that consist of sediments of marine origin—*e.g.*, oolite (small, rounded, calcium carbonate particle) beaches on the Bahamas and Yucatán, shell beaches along the Black and Caspian seas—and it is quite evident that beach sediment was always delivered to a given area by waves (if boulders and pebbles are present) or by waves and currents (if sand and gravel are present). With waves of equal power, a few hours are sufficient to stop the transverse shifting of beach sediment up or down the slope in the zone of effective action and to attain the equilibrium profile. This means that, for reciprocal wave oscillations, every particle or every unit volume of the beach will remain in a comparatively small zone until the wave parameters change. Then movement up or down the profile slope recommences, and a new equilibrium profile is constructed. This regularity has been repeatedly proved in experimental wave tanks with wave-reproducing machines. The beach profile will attain the same slope after the action of several thousands of waves, if equal parameters are involved. At the same time, submarine bars are formed beneath the water level.

In addition to transverse drifting, there is almost always a longshore sediment displacement in the zone where waves break (swash) and farther offshore. When the beach is formed by coarse sediment, longshore drifting may result from oblique wave approach. Oblique waves make the pebbles roll beachward in the direction of the wave ray and seaward along a line normal to the beach (Figure 2). This motion produces net transport parallel to

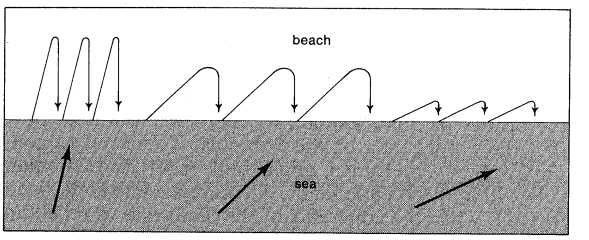


Figure 2: Beach sediment drifting under action of oblique wave incidence.

the shore. In longshore displacement of sand, the main factor is temporary currents that are caused by wave action and, most importantly, by winds that influence water flow. The speed of these currents near the bottom may attain two metres per second, sufficient to permit them to move sand independently. Usually, however, a complex mechanism is involved: waves (diving or spilling breakers) lift the sand grains, and the current carries them in suspension until the moment of settling is reached. The maximum rate of drift of boulders and pebbles occurs at the zone of breaking waves, whereas in the case of sand it occurs in the zone of primary wave surfing, sometimes far from the shoreline. If submarine bars are present, two or even three maximum drifting zones are created.

Wave action will produce excellent sorting of beach sediments. Small particles are washed away because they are in suspension longer than the duration of the wave period. Larger particles are segregated into several zones that accord with their hydraulic coarseness (size in relation to transportability of the currents). Usually coarser particles are located on the exposed part of a beach, close to its rear where the swash is still strong enough to move them. On broad beaches the sediment size decreases in both directions from the crest line. In the remote rear zones, only sand-sized particles will be transported by the weakened swash. Sediment size also decreases at the lower-beach margin. In every zone, however, the sorting is good.

Beach sands

Beach sands in temperate latitudes consist mainly of quartz, some feldspars, and a small percentage of heavy minerals. In the tropics, however, calcareous beaches composed of skeletal remnants of marine organisms and precipitated particles, such as oolites, are widespread.

Sometimes the basement layers of the beach are cemented by calcium carbonate, precipitated from the groundwater. This will commonly result if fresh water penetrates a beach from swamps behind it. If the beach undergoes erosion and thus retreats, the cemented strata become exposed; termed beach rock, they are widespread in the tropics and along the shores of the Mediterranean, Black, and Caspian seas.

Transverse and longshore sediment drift is a very complicated process and many aspects are not yet completely understood or studied. The depth to which sediment drifts has been measured by consecutive soundings (Figure 3). A number of soundings led U.S. scientists to conclude that this process does not occur at depths of more than 10 fathoms (18 metres). Some exact soundings and sediment tracing by Soviet scientists revealed, however, that sands in the Baltic Sea actively move at a depth of more than 20 metres and that pebbles in the Black Sea move at a depth of 30 metres along steep submarine slopes. Sediment movement must be of still greater importance in the nearshore zone of the open ocean. During storms in the Barents Sea, for example, sand is thrown onto the decks of fishing boats from depths of 50 metres.

The application of special indicators, or sediment tracers, has led to increased knowledge of these processes. There are two types of indicators: radioisotopes of a

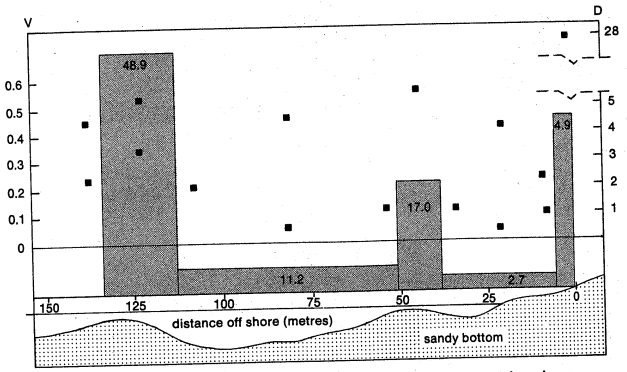


Figure 3: The sediment quantity transported in suspension by waves and currents along the sandy bottom with two bars.

number of metals and luminescent dyes. In using the first, artificial sand is activated in a nuclear reactor and then poured out onto the bottom. A Geiger counter is trailed across this zone at regular intervals. This fixes the frequency of the radioactive impulses at each spot and along the profile in general. The duration of the experiment depends on the half-life of the isotope used. The accuracy of the measurement also is determined by the natural background radioactivity of the bottom sand.

The use of luminescent dyes is more laborious, but there are a number of advantages. The indicator can be observed visually in bottom samples and on the exposed part of the beach. Dyes of different colours make it possible to study sand displacement from points or areas at various depths. By preparing dyed mixtures of various size fractions, it is possible to determine the speed and direction of transport of grains of different size. Highly successful experiments with luminescent tracers were carried out in the U.S.S.R. and in the U.S. before 1966. It was learned that individual sand grains can move at velocities as great as 2.8 kilometres per hour; the total mass of sediments, however, drifts more than ten times slower than this. Pebbles move along the beach at velocities sometimes as great as 260 metres per hour; this fact also was established by the studying of the movement of dyed specimens.

Waves normal to the shore commonly produce a series of beach cusps. The size and shape of these small bights depends on sediment coarseness and wave parameters. Cusp length varies from 2–3 metres to as much as 40–50 metres; cusp width is approximately two to three times less. The mechanism of their formation is not yet entirely clear, but cuspate shorelines are obviously related to potential wave energy and are thought to reflect equilibrium configurations.

Transverse and longshore sediment drifting produce beaches of quite different form. The former creates barrier beaches that consist of long sediment strips (usually sand, in some cases pebbles and shells) above sea level at a considerable distance offshore. This widespread beach type makes up 12–13 percent of the total world coastline. The majority of these beaches occur at the outer margins of alluvial plains, which received much sediment during the low stands of sea level associated with the Pleistocene glaciations. The main factor in the formation of barriers seems, in fact, to have been the late glacial rise of sea level (the rise associated with glacial melting), which ended about 3,000 years ago.

Longshore sediment drift

Longshore sediment drifting is variable; the drift direction changes according to the direction of wave approach. Thus, sediment may migrate in opposite directions, and if the shoreline consists of capes and concavities, the bulk of beach material may be divided into separate masses. Migration in opposite directions also occurs along straight shorelines; but when the prevailing waves tend to impinge along a path that is not perpendicular to the shore, net transport will take place, and this is termed longshore sediment flow. The capacity of the flow depends on the wave energy and frequency. The load and, consequently, the saturation depend on the stock of material on the beach. Sediment flows can transport up to 1,000,000 cubic metres of sand and 100,000 cubic metres of pebbles annually through a given section of shore. Sediment flows may be hundreds and thousands of kilometres in length. They have acted upon seashores during the whole of geological time.

The formation of a number of accumulative bodies (beaches) is connected with wave refraction, namely, the curving of a wave front when waves approach the shore at an acute angle. The wave front lengthens and the energy per unit length of wave is reduced. There is accordingly an optimal angle of wave approach to the shore at which the speed and volume of the transported sediment reach maximum values. Depending on the sediment size and the mean wave parameters, this angle lies between 40° and 50°.

If the shoreline is curved so that the capacity of a saturated sediment flow diminishes from one stretch to the next, part of the load is deposited. This material forms beach forelands, or spits. Another cause for such accumulations is the decrease of wave energy behind an obstacle (an island or a cape) or in a narrow bay. If the sediment flow retains some unused capacity, a more considerable energy drop is required to make the load precipitate and be deposited on submarine and exposed parts of a beach. If beaches are inside narrow bays, wave refraction serves to elaborate their arched outlines. The curved wave always approaches each spot of the beach at a right angle, and this produces much shifting of material offshore until a stable position is reached. Curved, semicircular beaches also occur in broad open bays. The waves approach these beaches from different sides and, consequently, cause permanent migration of sediments. Nevertheless, there exist certain stable outlines corresponding to the sum of all the forces imposed by the waves. The apex of the beach curve corresponds to the predominant wave system.

Beach destruction and preservation. When beach sediment is involved in permanent movement, it undergoes certain attrition. This is especially great for boulders and pebbles, and it reaches 5 percent of the weight as an annual average. Sand grains are only slightly abraded (the lower limit of perceptible abrasion corresponds to a diameter of 0.3 millimetre), but sand losses take place during intense storms if the near-shore sea is deep enough. On the California coast and along the eastern shore of the Black Sea, for example, much sand is carried away through submarine canyons. The beach is a dynamic self-regulating system that has its own input and expenditure balance. Beaches are sensitive to any distortion of this balance. For instance, they become smaller and may even disappear as a result of the intensive use of river waters for irrigation or hydroelectric power. In such cases the sediment discharge as well as the water discharge is reduced. Some construction projects (piers, jetties, breakwaters, etc.) will cause beach sediment accumulation where the longshore flow impinges, whereas erosion will occur along the shore on the opposite side. If there is no net transport, the sediment may be drawn up under the protection of the structure from both sides of the adjoining beaches. This is doubly harmful: protected areas and canals suffer silting, and the adjacent shore is eroded. For many years silting was prevented by means of artificial deepening of harbours and lengthening of inlet jetties. Recently, sediment bypassing to shift the flow

Use of groins and jetties

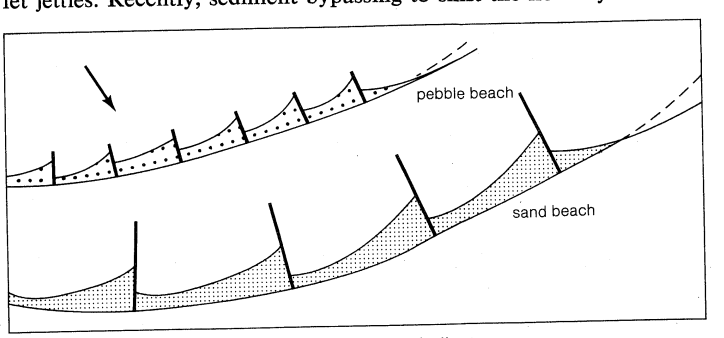


Figure 4: Comparative schemes of groins. Arrow indicates the general direction of incoming waves.

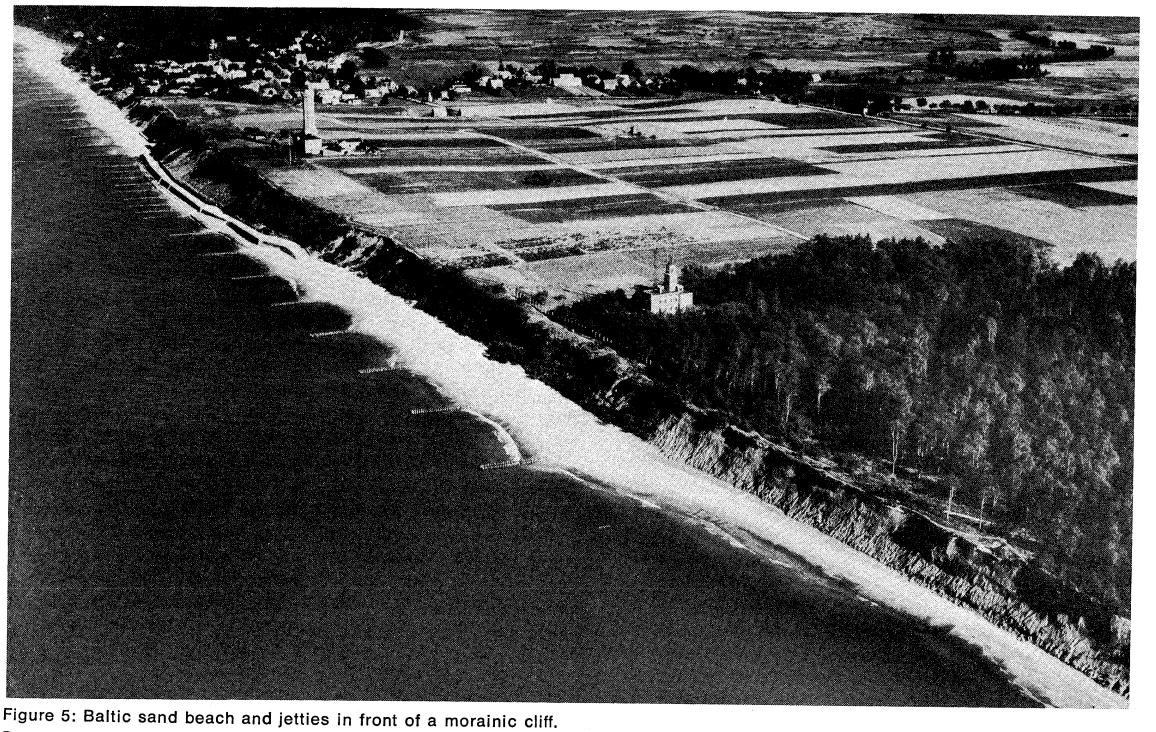


Figure 5: Baltic sand beach and jetties in front of a morainic cliff.
By courtesy of V.P. Zenkovich

from one side of the structure to the other has come into use; it is a most effective method.

Solid rock shores that lack beaches are easily destroyed by the sea. Consequently beaches provide protection for the shore. For this reason, among the several ways of protecting the shores from erosion is the use of groins—short piers that extend out into the sea at an adequate distance (30–40 metres for pebbles, 60–200 metres for sand). Usually there is a series of groins with the spacing interval depending on their length and on the angle of the resultant waves during the strongest storms (Figure 4). As a rule these intervals are equal to one to two groin lengths. The sediment flow deposits its load in the intervals until it takes a new route on the shallowed bottom in front of the groins' heads.

While a system of groins is being filled, no sediment is carried by the downflow, and so a certain part of the shore may undergo considerable erosion. If there is a deficiency of natural sediments, the intervals between the groins are filled up artificially. Gravel and pebbles will remain there for decades, but after that a supplement will be necessary. This kind of shore protection is widely used along the Caucasus and the Crimea coasts of the Black Sea (Figure 5).

Finally, within the last few years, artificial replenishment of the beaches with sand taken from the sea bottom or transported from adjoining dunes has been widely used. The best examples of the application of this method are on the Long Island, Virginia, and Florida beaches of the U.S. and those of the island of Norderney in the North Sea.

From E.C. F. Bird, *Coasts* (1968), p. 139; The Australian National University Press

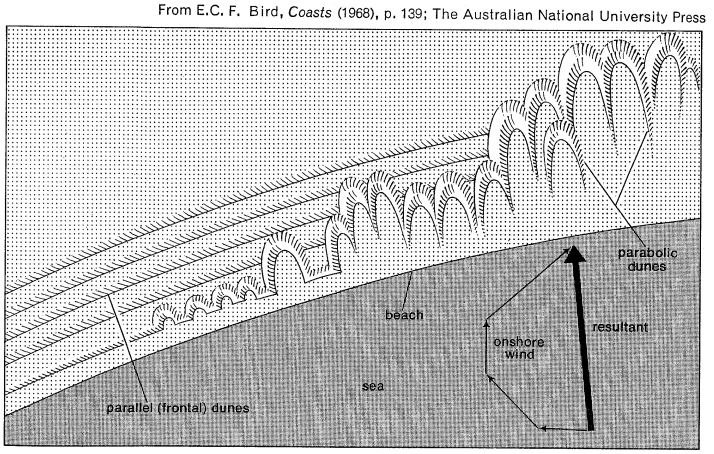


Figure 6: Scheme of frontal and parabolic shore dunes behind the beach and the onshore wind resultant.

In many countries the wind strongly affects the dynamics of the beach. The beach is exposed to the sea wind, and sand is usually blown off to the rear parts of the beach where it forms small hummocks. As these join together, foredunes are being built, and if the beach is well supplied with sand in the right area, several rows of dunes will be formed. When the sand is very abundant, dunes will shift to adjacent low-lying plains and may bury fertile soils, woods, and buildings (Figure 6).

If sand is no longer delivered to the region of developed dunes, gaps will form in the ridges parallel to the shore. In such zones parabolic dunes with their summits coastward are created. After long stabilization, the summits of the parabolas may be broken through by the wind, thus gradually forming a series of ridges parallel to the prevailing winds.

Dunes and vegetation

Vegetation is quite important in dune formation. Primary accumulation of foredunes is facilitated by the growth of marram grass, heather, and other plants. Vegetation may fix dunes that are some distance from the beach, covering them completely and thus protecting them from the winds. The formation of dunes in general will, however, reflect climatic conditions. For example, a humid tropical climate is not favourable for dune formation. Dunes are abundant along the California coast near the Sonora Desert. To the south, however, closer to Matzatlán, precipitation is more than 1,000 millimetres per year, and even embryonic dunes are completely overgrown with bushes. The same conditions pertain along the southwestern coast of the Gulf of Mexico.

Beaches bordered by dunes are widespread on all continents. There is much evidence that the bulk of the sand accumulated during the Holocene transgression. The ocean level was fairly low prior to this latest rise of sea level, and coasts were bordered by broad alluvial plains. In the course of the transgression, sand was transported by waves and wind onto the land in great masses.

The practical significance of beaches is not limited to their function as protectors of the coast or as recreation sites. The sorting mechanism of the offshore waves and

currents determines the accumulation of heavy-mineral (specific weight over 2.7) concentrates. On any sand beach there are thin layers of dark sand that can be seen. Some heavy minerals contain valuable metals, such as titanium, zirconium, germanium, tin, uranium, and gold. In many places the concentrations are so great that they are of industrial significance; placer deposits are worked in India, Brazil, Japan, Australia, the U.S.S.R., and Alaska. Heavy-mineral concentrates also are extracted from the submarine slopes by means of dredging ships.

In view of the public use of the beaches, a special service has been organized in many countries that must watch all the changes in the beaches and maintain them artificially in the best possible state. These are the functions of Nature Conservancy in England and a special Beach Erosion Board in the U.S., which formerly belonged to the U.S. Army Corps of Engineers. Almost every year scientific conferences on coastal engineering are organized by the U.S. Coastal Engineering Research Council.

BIBLIOGRAPHY. Н.А. Айбулатов, *Исследование вдольберегово перемещения песчаных наносов в море* (1966), Soviet study investigating the application of luminescent tracers for the movement of longshore sand drifting in the sea, data collected using the cable way above the sea; W. BASCOM, *Waves and Beaches: The dynamics of the Ocean Surface* (1964), a popular description of the seashore, morphology and dynamics of the beaches and their protection, based on the investigations carried out at the Pacific U.S. Coast; E.C.F. BIRD, *Coastal Land Forms: An Introduction to Coastal Geomorphology with Australian Examples* (1965), a coherent introduction in coastal geomorphology for students in geology and geography, compiled on the basis of worldwide up-to-date sources; P. BRUNN, *Coast Stability* (1954), detailed description (series of 7 pamphlets) of the erosion of western Denmark beaches and their stabilizing with groins; A. GUILCHER, *Morphologie littorale et sous-marine* (1954; Eng. trans., *Coastal and Submarine Morphology*, 1958), a brief explanation of general laws of coastal development and description of coastal features; J.C. INGLE, *The Movement of Beach Sand*, (1966), a comprehensive study using fluorescent tracers and a mathematical treatment of the results; D.W. JOHNSON, *Shore Processes and Shoreline Development* (1919), a first presentation of the subject, emphasizing Johnson's classification schemes; C.A.M. KING, *Beaches and Coasts* (1959), excellent general treatise that incorporates field and model data; Академия наук СССР (ed.), *Динамика береговой зоны бесприливных морей* (1963), treats wave deformation, wave-generated currents, and sediment transport in the Black Sea; R.L. WIEGEL, *Oceanographical Engineering* (1964), chapters on nearshore waves, the dynamics of beaches, and on shore and harbour protection; W.W. WILLIAMS, *Coastal Changes* (1960), popular account of the intermittent character of beaches and their relation to shore stability or changes; V.P. ZENKOVICH, *Processes of Coastal Development* (Eng. trans. 1967; orig. pub. in Russian, 1962), a comprehensive study of the morphology and dynamics of beaches, mainly in tideless seas, with references to much of the Soviet literature on coasts.

(V.P.Z.)

Beaufort Sea

The Beaufort Sea is situated in the Arctic Ocean north of Canada and Alaska. It has no natural boundaries but has been defined by the International Hydrographic Bureau as extending from Point Barrow (Alaska) toward Lands End Point (on Prince Patrick Island, Canadian Arctic Archipelago), and from Banks Island in the east to the Chukchi Sea in the west. The sea owes its name to the British Rear Admiral Sir Francis Beaufort.

Its surface area is 184,000 square miles (476,000 square kilometres) and its volume 115,000 cubic miles. The average depth is 3,293 feet and the greatest depth 15,360 feet. The Beaufort Sea is under ice almost the year round; only in August and September does it break up and then only near the coasts. So far the sea has been most inadequately studied. Serious scientific investigations by icebreakers, drifting stations, atomic submarines, and aircraft began in the 1950s.

Physiography. The shelf of the sea is narrow, especially in the vicinity and east of Point Barrow; it widens somewhat north of the Mackenzie River mouth but nowhere exceeds 90 miles. The usual depth is less than 210 feet, although the continental slope is very steep in the sea's upper part, descending to 5,000 or 6,500 feet. Small gravel islands or shallows are often found on the shelf, the largest of them west of the Mackenzie River mouth. These are the islands of Herschel (seven square miles) and Barter (five square miles). A variety of very small islands and banks are found in the Mackenzie River Delta.

The continental slope of the sea is dissected by numerous submarine valleys. The largest of them begin from Point Barrow, from the Mackenzie River mouth, from 140 miles northeast of the Mackenzie River, and from the Amundsen Gulf, a submerged glacial trough. The Beaufort plateau, with depths from 6,500 to 10,000 feet, protrudes far into the sea west of Banks Island. The geological structure of the bottom is that of a massive platform, fringed with Mesozoic folds. The seismic data, however, negate the submergence hypotheses and indicate a similarity between the crust of the Canadian Basin and of the oceans.

Nature of the coastline

The sea coasts are low lying and covered with tundra over almost their entire extent. Only west of the Mackenzie River mouth do spurs of the Brooks Range approach the coastline. The Banks and Prince Patrick islands are also fairly low, maximum elevations being from about 900 to 2,450 feet. A northwestern passage begins in the eastern portion of the sea through the Amundsen Gulf. The southern coastal regions of the sea are navigable from middle August till late September. A permanent barrier of ice ten feet thick begins near the shores.

Hydrology. Four water masses may be distinguished: surface, subsurface Pacific, deep Atlantic, and bottom waters.

The surface water mass is nearly 330 feet thick and is the coldest of all. Temperature ranges from 29.5° F (−1.4° C) in late summer to 28.8° F (−1.8° C) in winter. Salinity is 28 to 32 parts per thousand.

The subsurface water mass is formed by the waters of the Pacific Ocean and the Bering Sea, which flow through the Bering Strait. This water mass is less noticeable in the northeastern and northern parts of the sea. The Pacific water is much warmer than the surface water and almost reaches the North Pole.

The deep Atlantic water is the warmest of all, its temperature ranging from 32° to 34° F (0° to 1° C) and its salinity from 34.9 to 35.5 parts per thousand. The warmest water is found at a depth of nearly 1,300 feet. The upper boundary of this mass lies at about 650 feet and the lower one at 2,600 to 3,000 feet.

The bottom water begins at depths of more than 3,000 feet. Negative temperature (−0.4° to −0.8° C, or 31.3° to 30.6° F) and almost constant salinity of 34.9 parts per thousand are typical.

Direction of currents

The direction of the surface and subsurface currents is closely related to the general current system of the Arctic Ocean. A clockwise water gyre flows north of the Beaufort Sea: the majority of the sea's currents are thus westward or southwestward. Only in the vicinity of the Mackenzie River mouth is an eastward current recorded. Current velocities in the open sea are not great, ranging from one to two miles per day. Current direction is also strongly affected by local winds.

Bottom sediments. Sedimentary material is supplied to the sea by rivers and is also brought in through the Bering Strait. Deposits from the Mackenzie River are particularly large: the annual liquid discharge of the river is nearly 440 million tons and the sedimentary one about 15 million tons. High concentrations of dolomite and calcium carbonate are distinguishing features of the Mackenzie's deposits, and these minerals are found at great distances from the river delta. Gravel, pebble, and sand deposits, sometimes mixed with mud, are widely distributed on the shelf. The deep-sea bottom is covered with fine clay muds. Dispersed gravel–pebble material carried by ice from the coastal zone into the open sea is a constant component of all types of sediments. Deep-sea muds are brown. Gray muds containing rare microfauna have been found on the continental slope and north of Banks Island. Sediments also contain significant proportions of hornblende and iron oxides.

Marine life. Phytoplankton in the Beaufort Sea is diversified, but its biomass is not large. Representatives of more than 70 phytoplankton species are found. The maximum development is recorded in August and September in the upper 160-foot layer of the open sea, in the open water between ice floes, and in lakelets on the ice floe surfaces. Nearly 80 zooplankton species have been found. The bottom fauna consists of nearly 700 species of polychaetes, bryozoans, crustaceans, and mollusks.

Fishing and sea hunting are for local supply only. The economic development of the coastal regions may in the future be stimulated by the search for marine oil and gas deposits.

BIBLIOGRAPHY. A.J. CARSOLA *et al.*, "Bathymetry of the Beaufort Sea," in G.O. KAASCH (ed.), *Geology of the Arctic*, pp. 678–689 (1961); A.J. CARSOLA, "Recent Marine Sediments from Alaskan and Northwest Canadian Arctic," *Bull. Am. Ass. Petrol. Geol.*, 38:1552–1586 (1954); R.S. DIETZ and G. SHUMWAY, "Arctic Basin Geomorphology," *Bull. Geol. Soc. Am.*, 72:1319–1330 (1961).

(A.P.L.)

Beaumont and Fletcher

Leading poetic dramatists of the early 17th century whose names are associated with a large body of plays, Francis Beaumont and John Fletcher furnish a remarkable example of collaboration in English literature. Their work is melodramatic, with tense situations and startling reversals; the characters are inconsistent and involved in peripheral situations. Although heroic ideals—friendship, honour, duty—are challenged, the issue is often left in balance; their plays often appear to demand a tragic resolution and yet manage to avoid one. The playwrights' talents fruitfully counterpointed each other, Beaumont providing tightness of organization and Fletcher, fertility of invention. The language of their plays is graceful and lucid, an idealized version of aristocratic speech.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Beaumont (left) and Fletcher (right), engravings by George Vertue, 1729.

Known biographical information

Beaumont and Fletcher lived during the reign of King James I (1603–25) and were younger contemporaries of William Shakespeare. They were of a social class somewhat higher than usual for Jacobean professional playwrights. Fletcher's father was an ecclesiastic and a favourite of Queen Elizabeth I, officiating at Mary Stuart's execution and finally becoming bishop of London; Beaumont was the youngest son of a justice of the common pleas of Grace-Dieu priory in Charnwood Forest, Leicestershire. Fletcher was baptized at Rye, in Sussex, on December 20, 1579; Beaumont was born in 1584, presumably at Grace-Dieu. Fletcher apparently entered Bene't (later Corpus Christi) College, Cambridge, in 1591; Beaumont entered Broadgates Hall (later Pembroke College), Oxford, in 1597 and afterward, being intended for a career at law, London's Inner Temple in 1600. Both young men found themselves in uncongenial situations, with insufficient private means.

By 1604 they were part of a circle surrounding Ben Jonson, the dominant figure in literary Bohemia, and they contributed prefatory verses to Jonson's play *Volpone*. Beaumont remains a shadowy figure; Fletcher appears to have been genial and carefree, perhaps as might be expected from the heroes in some of his plays. According to John Aubrey, a 17th-century memorialist, "They lived together on the Banke side, not far from the Play-house, both batchelors; lay together . . .; had one wench in the house between them . . .; the same cloathes and cloake, &c., betweene them" (*Brief Lives*). Their collaboration as playwrights began in about 1607 and was to last for some seven years: in 1613 Beaumont married an heiress, Ursula Isley of Sundridge in Kent, retired from the theatre, died on March 6, 1616, in London, and was buried in Westminster Abbey. Fletcher continued to write plays both alone and in collaboration with others; he may be the Fletcher whose marriage is recorded as taking place in St. Saviour's, Southwark, where he is said (by Sir Aston Cokayne in 1658) to have been buried on August 29, 1625. He had died of the plague in London, where he lingered to be measured for a suit of clothes instead of making his escape to the country. Little or nothing more is known about their lives and characters apart from what can be guessed at from the plays with which they are linked.

Professional career. *Their audience.* From the outset of Beaumont and Fletcher's careers, a change was apparent in the English theatre. The permanent playhouses that had been opened toward the end of Queen Elizabeth's reign had been public theatres, and side by side with them a number of private theatres began to emerge; in 1608 the King's Men, a theatre company for which Shakespeare had written and Beaumont and Fletcher were to write, took over one such private theatre at Blackfriars in London. The same company continued to maintain a public theatre (the Globe) in London, but Blackfriars became more and more their centre. Hitherto the sophisticated and the illiterate had together made up a London audience; Beaumont and Fletcher, however, spent their literary careers writing entirely for the private theatre and its smaller but educated audience, their work being performed on an enclosed, distant stage quite unlike the intimate thrust stage of, for example, the Globe. This ambience suited their adroit, self-conscious stagecraft and their love of casuistry in situation and argument—the last appealing particularly to the legal minds of the Inns of Court men who made up a majority of their audience.

The private theatre and its audience

Their plays. The canon of the Beaumont and Fletcher plays is largely represented by 52 plays contained in the folio *Fifty Comedies and Tragedies. Written by Francis Beaumont and John Fletcher, Gentlemen*, which appeared in 1679. It might properly be termed "The Works of Beaumont, Fletcher, and Company" for Fletcher collaborated with at least six other dramatists after Beaumont's retirement. Fifteen plays can be assigned with some certainty to Fletcher alone; Beaumont's notable sole offering is *The Knight of the Burning Pestle;* joint productions number ten in all, including *The Maides Tragedy*, *Phylaster*, and *A King and No King* (in which plays Beaumont's is assumed to have been the controlling hand, since they manifest a firmer structure than Fletcher's single or collaborative efforts). Three plays were the work of Beaumont, Fletcher, and Philip Massinger (who succeeded Fletcher as chief playwright for the King's Men); 12 plays were by Fletcher and Massinger together; 13 were by Fletcher working with other collaborators, including such outstanding dramatists as Shakespeare, Jonson, Thomas Middleton, John Webster, and George Chapman. One play contained in the 1679 folio is *The Coronation*, a comedy that has been fairly conclusively identified as the work of James Shirley: three plays not contained in this folio may be added to the Beaumont and Fletcher canon (*Henry VIII*, by Shakespeare with Fletcher; and *Sir John van Olden Barnavelt* by Fletcher and Massinger and probably *A very Woman*). Attributions of plays in which Fletcher is not the sole author are as relatively uncertain as defining the correct share of Beaumont and Fletcher.

Attempts to disentangle the various shares of Beaumont and Fletcher in any given work are complicated by the

fact that Beaumont sometimes revised scenes by Fletcher, and Fletcher edited some of Beaumont's work. Attribution is based chiefly on clues from internal stylistic evidence: metrical tests, for example, have shown that Fletcher's literary mannerisms included a fondness for feminine endings (*i.e.*, with the last syllable of a line left unstressed).

Dramatic characteristics. Beaumont's unaided work, *The Knight of the Burning Pestle*, parodies a then popular kind of play—sprawling, episodic, with sentimental lovers and chivalric adventures. It opens with The Citizen and his Wife taking their places on the stage to watch "The London Merchant"—itself a satire on the work of a contemporary playwright, Thomas Dekker. Citizen and Wife interrupt, advise, and insist that their apprentice should take a leading part. In it, Beaumont indulgently satirizes bourgeois naïveté about art.

This was not immediately a popular play, and neither was Fletcher's first solo offering, *The Faithfull Shepheardesse*. A stylized tragicomedy in a pastoral setting, it displays a spectrum of attitudes to love ranging from the spirituality of Clorin, the Shepherdess (who is faithful to a dead lover), to the Sullen Shepherd, and finally to Chloe ("It is impossible to ravish me/ I am so willing"). The plot is derived from the pastoral drama *L'Aminta* (first performed 1573) by the Italian poet Torquato Tasso, whose theme ("that is lawful which doth please") is rejected by Fletcher, who suggests that nature and natural impulse must be controlled. (The play was to be an important source for John Milton's masque, *Comus*, whose theme is that of chastity.)

Theatrical formula of *Phylaster*

But in *Phylaster*, which is the most romantic, amiably improbable play of their partnership, the dramatists together found a successful theatrical formula. The principal characters fall into absurd situations that afford a virtual parody of the revenge drama: Prince Phylaster, in love with Arethusa, the usurping King's daughter, is loved by Euphrasia, who disguises herself as a page and acts as servant to both, a part she finally chooses to retain rather than make a suitable marriage. The courtly setting, the divided loyalties, the disguises, and the fascination with near sadomasochistic situations and states of mind (evidenced in Phylaster's wounding of Arethusa and Euphrasia's continuing servitude) were to be reworked in later plays. *The Maides Tragedy* (first performed 1611) is the most powerful of the dramatists' joint works. In it, the Maid, Aspatia, is betrothed to Amintor, who is forced by the King to break off his engagement and marry Evadne, sister of the King's general, Melantius. In an astonishing scene on the wedding night, Amintor gradually learns that the marriage is to remain unconsummated, being merely a screen for Evadne's affair with the King. Suspense is maintained by twisting the plot: the King's jealousy of Amintor, Melantius' dilemma of loyalty, Evadne's murder of the King. By the end most of the principal characters are murdered, including Aspatia, disguised, at the hands of Amintor. The protagonists show no intensified moral awareness, and critics have seen this play as an unrelated series of titillating problem situations. Yet Petrarchan notions of "pure" love, military honour, chastity, and the Divine Right of Kings are wittily and trenchantly mocked in Evadne's dry nihilism while the close is politically daring: Melantius does not suffer for his part in the regicide but remains as general to the new regime. Of approximately the same date, *A King And No King* is remarkable for symmetrical structure; the subplot neatly balances the main action. The play treats of supposedly incestuous love between brother and sister, Arbaces and Panthea, from which arises maximum casuistry and excitement.

Fletcher's unaided work

Among Fletcher's unaided work are two tragedies, *Bonduca* and the more interesting *Valentinian*. Like *The Maides Tragedy*, *Valentinian* proceeds from climax to climax, each one virtually autonomous and not part of a single ascendant rhythm. The play, a pseudo-history, is set in the decadence of the Roman Empire; the young emperor Valentinian is surrounded by sycophants eager to serve his lust for Lucina, wife of the Emperor's general, Maximus. An atmosphere of intense sensuality pervades the court, stressed by two accomplished lyrics, "Hear ye, Ladies that despise" and "Now the lusty Spring is seen," sung to soften Lucina before Valentinian's attempted seduction. Fletcher's songs, in beauty second only to Shakespeare's, are generally decorative but here possess thematic relevance. Fletcher's plots particularly emphasize how far identity, attitudes, and course of action are imposed upon men by circumstances and expectations of the social world; both Maximus and Lucina, after Lucina's rape by Valentinian, agree that she must kill herself. Neither grief nor private sense of violation demands this but rather a concern for reputation. Characters in other plays remain absorbed with the "image" they present. Fletcher's *The Wild-Goose Chase* anticipates the course of comedy for the following 100 years. Oriana, the heroine, is handicapped by having naïvely confessed her love for Mirabel to himself and his friends. She is, however, determined that Mirabel shall marry, not seduce, her; the ensuing sexual duel is worked out through witty argument and Oriana's various stratagems. The more farcical moments concern disguise, underlining the necessity for artifice in civilized society. Finally the wild goose tires of being chased and capitulates. While less savage than much of the Restoration comedy that succeeds it, the play is not sentimental; its wit is sustained; the plotting uncharacteristically severe, with scarcely a sentence unrelated to the central theme of marriage.

Fletcher succeeded Shakespeare as principal dramatist at Blackfriars, and both he and Beaumont were intensely conscious of Shakespeare's work. They sometimes invert, sometimes "outdo" the master: the Euphrasia of *Phylaster*, for example, owes something to the Viola of Shakespeare's *Twelfth Night*. Their own fluency, however, inevitably suffers by the side of Shakespeare's richly metaphorical language.

Reputation. The reputation of Beaumont and Fletcher has touched extremes. A small folio collection of 1647 and that of 1679 suggest that their work then ranked with Shakespeare and Jonson in distinction and popularity. Until the close of Charles II's reign in 1685, they were acted frequently and with applause. Like Shakespeare, they survived (in adaptation) through the 18th and up to the mid-19th century, when their plots and language were found offensive. Reaction, however, had begun earlier with Romanticism, the critic and poet Samuel Taylor Coleridge being particularly harsh. A number of 20th-century critics, such as T.S. Eliot and Muriel C. Bradbrook, have been equally unfriendly. The test of staging is now rare; yet their art anticipates much 20th-century practice. The Italian dramatist Luigi Pirandello found them absorbing, and they are acutely interested in several problems that fascinate the modern theatre: the relationships between fictions and truths; the problem of identity; the adoption of roles. Beaumont and Fletcher's work will always pose critical problems, but it embodies a vivid and witty, if limited, account of man's inadequacies.

MAJOR WORKS

The canon of the Beaumont and Fletcher plays was printed in the folio *Fifty Comedies and Tragedies* . . . (1679), which had been preceded by a smaller folio in 1647. The dates of composition given here for the individual plays are largely conjectural.

PLAYS BY BEAUMONT AND FLETCHER IN COLLABORATION: *The Woman Hater* (1606); *Phylaster* (1608–10); *The Coxcombe* (1608–10); *The Maides Tragedy* (1608–11); *The Captaine* (1609–12); *A King and No King* (1611); *Cupids Revenge* (1611); *The Scornful Ladie* (1613–17); *Loves Pilgrimage* (?1616); *The Noble Gentleman* (*c.* 1625).

PLAY BY BEAUMONT UNAIDED: *The Knight of the Burning Pestle* (1607).

PLAYS BY FLETCHER UNAIDED: *The Faithfull Shepheardesse* (1608–09); *Valentinian* (1610–14); *Monsieur Thomas* (1610–16); *The Womans Prize* (?1611); *Bonduca* (1611–14); *The Mad Lover* (1616); *The Chances* (?1617); *The Loyall Subject* (1618); *Women pleas'd* (?1619); *The Island Princesse* (?1619–21); *The Humorous Lieutenant* (?1619); *The Wild-Goose Chase* (1621); *The Pilgrim* (1621); *Rule A Wife And have a Wife* (1624); *A Wife for a Moneth* (1624).

PLAYS BY BEAUMONT, FLETCHER, AND PHILIP MASSINGER:

Thierry and Theodoret (date of composition unknown, printed 1621); *The Beggars Bush* (?1622); *Loves Cure* (?revived 1625).

PLAYS BY FLETCHER AND MASSINGER: *The Little French Lawyer* (1619–23); *Sir John van Olden Barnavelt* (1619); *The Custome of the Countrey* (1619); *The False One* (1620); *The double Marriage* (1621); *The Spanish Curat* (1622); *The Sea Voyage* (1622); *The Prophetesse* (1622); *The Lovers Progress* (1623–34); *The Elder Brother* (1625); *A very Woman* (1625–34).

PLAYS BY FLETCHER WITH VARIOUS OTHER COLLABORATORS: including plays written by Fletcher and Massinger together with a third and sometimes a fourth collaborator—with Shakespeare (no general agreement on these assignments but they probably represent majority opinion), *Henry VIII* (1613); *The Two Noble Kinsmen* (1613); with James Shirley, *The Night-Walker* (1633); with Nathaniel Field, *Foure Playes in One* (?1609–12); with Massinger and Field, *The Honest mans Fortune* (1613); *The Knight of Malta* (1616–18); *The Queene of Corinth* (1616–17); with an unknown reviser, *Wit With-Out Money* (1614); with Thomas Middleton, *The Nice Valour* (?1616); with William Rowley, *The Maid in the Mill* (1623); with John Ford, Massinger, and John Webster, *The Faire Maide of the Inne* (1626); with Massinger, Ben Jonson, and George Chapman, *The Bloody Brother, or, Rollo, Duke of Normandy* (?1617–30).

BIBLIOGRAPHY. S.A. TANNENBAUM, *Beaumont and Fletcher: A Concise Bibliography* (1938; *Supplement*, 1946); CYRUS H. HOY, "The Shares of Fletcher and His Collaborators in the Beaumont and Fletcher Canon," *Studies in Bibliography*, vol. 8–9, 11, 12–15 (1956–62).

Editions: *The Works of Francis Beaumont and John Fletcher*, 4 vol. Variorum edition, ed. by P.A. DANIEL, R. WARWICK BOND *et al.*, under the general editorship of A.H. BULLEN (1904–12), of the projected 12 volumes only four were completed; *The Works . . .*, ed. by ARNOLD GLOVER and A.R. WALLER, 10 vol. (1905–12, reprinted 1969), the standard text, now being superseded by the edition of Bowers; FREDSON BOWERS (ed.), *The Dramatic Works in the Beaumont and Fletcher Canon* (1966–), vol. 1 and 2 have so far appeared —the definitive edition; *Poems* by FRANCIS BEAUMONT (1640) with additions (1653) and in GEORGE DARLEY'S *Works* of 1840.

Biography and Criticism: U.M. ELLIS-FERMOR, *The Jacobean Drama: An Interpretation*, new ed. (1953); E.M. WAITH, *The Pattern of Tragicomedy in Beaumont and Fletcher* (1952), claims that Beaumont and Fletcher significantly stylize experience and relates their work to the Roman art of declamation and to the classical *controversiae;* J.F. DANBY, *Elizabethan and Jacobean Poets* (1964), brilliant pages of Beaumont and Fletcher criticism; M.C. BRADBROOK, *The Growth and Structure of Elizabethan Comedy* (1955), with a chapter on Fletcher's comedies; CLIFFORD LEECH, *The John Fletcher Plays* (1962), acute and sympathetic; T.B. TOMLINSON, *A Study of Elizabethan and Jacobean Tragedy*, ch. 12 (1964), deals with Beaumont and Fletcher, with some brilliant pages on *The Maides Tragedy;* IAN FLETCHER, *Beaumont and Fletcher* (1967), a brief, appreciative monograph.

(I.Fl.)

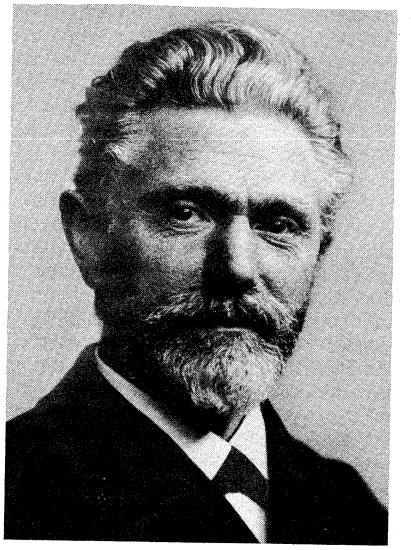
Bebel, August

August Bebel, cofounder of the German Social Democratic Party and its most influential and popular leader for more than 40 years, was one of the outstanding figures in the history of western European Socialism.

Bebel was born in Deutz near Cologne on February 22, 1840, the son of a Prussian noncommissioned officer. Growing up in extreme poverty at Wetzlar, where he learned the turner's craft, he began to travel as a journeyman through southern Germany and Austria and in the spring of 1860 settled in Leipzig, where he began his political career.

In 1861 Bebel joined the Leipzig Workers' Educational Association, which, like many others of its kind, was formed through the initiative of members of the liberal bourgeoisie; in 1865 he became its chairman. Political and economic circumstances, however, gave the workers' education movement an increasingly political orientation, which was to be significantly reflected in the development of Bebel's own political views. Like the other young workers in the new associations, Bebel had not yet heard anything of *The Communist Manifesto* or of its authors, Friedrich Engels and Karl Marx.

If in 1863 Bebel believed that the working classes were not ready for the vote, he was already changing his mind when he began his friendship with Wilhelm Liebknecht, who came to Leipzig from Berlin in 1865. Liebknecht, older than Bebel and a man with university training, became in many respects Bebel's mentor. But the more open-minded Bebel always maintained his independence. The Austro-Prussian War (1866), which divided German opinion between the advocates of a Kleindeutschland (Small Germany) and those of a Grossdeutschland (Large Germany), advocated by the Prussian prime minister Otto von Bismarck, drove the Saxon workers' associations into an alliance with the radical anti-Prussian democrats; for Bebel and Liebknecht, the workers' leaders, were implacable opponents of Bismarck. The Sächsische Volkspartei (Saxon People's Party) was thus brought into being, and in 1867 Bebel entered the constituent Reichstag of the North German confederation as a member for this party. Eventually, this and other like-minded parties united in 1869 in the Sozialdemokratische Arbeiterpartei (Social Democratic Labour Party) of Germany.



Bebel, *c.* 1898.
Archiv fur Kunst und Geschichte

Opposition to Bismarck

Already in 1867, as a member of the North German Reichstag, Bebel had protested against the Bismarckian "greater Prussia," believing that it meant "turning Germany into one great barracks." In parliament he untiringly continued this protest both before and after the founding of the German empire. He and Liebknecht were the only voices to speak against the war loan voted in the Reichstag on July 21, 1870; as a result, they were brought to trial on a charge of high treason at Leipzig in March 1872. Sentenced to two years' imprisonment, Bebel recovered from tuberculosis during this period of enforced idleness. Also, he was able at last to give himself a systematic education.

Beginning with an earlier sentence in 1869, Bebel spent a total of nearly five years in prison within less than 20 years, though he never faced any graver charge than that of "spreading doctrines dangerous to the state," "*lèse-majesté*," "libel of Bismarck," or "libel of the Bundesrat." These sentences were a serious threat to his livelihood. As the party itself could afford only the most essential expenditure and as a member of the Reichstag received no allowances, Bebel continued to rely on his income as a craftsman. He had established himself in Leipzig as a master turner and had married the daughter of a railway worker in 1864. Not until the end of the 1880s was he able to live by his writing.

As a writer Bebel had most success with *Die Frau und der Sozialismus* (1883; English translation, *Woman and Socialism*, 1904) which went through many editions and translations. This book was the most powerful piece of Social Democratic propaganda for decades. Above all, by its combination of science and prophecy, it served as a blueprint for German social democracy in the

conditions produced by Bismarck's Anti-Socialist Law (1878–90). Bebel himself never doubted that this period of repression under the emergency laws was anything more than an episode, declaring to his opponents in the Reichstag: "Your lances will be shattered in this struggle like glass on granite." His unshakable confidence gave his colleagues the courage to stand firmly together, but he opposed all tendencies toward retaliation by force, since terrorism or attempts at subversion might have endangered the very existence of the party.

These tactics were proved right when the emergency laws were allowed to lapse and when, in the elections of 1890, the Social Democrats received nearly 20 percent of the vote. Bebel's position at the head of the party was now uncontested, and in the Reichstag he was the most prominent opponent of the government. Within the party itself he opposed all the "opportunist" tendencies, which had come out into the open since the ending of the anti-Socialist laws. According to these, features of the existing social and political structure might be developed gradually until social democracy was attained. At the Erfurt congress of 1891 he reproached the leader of the Bavarian Social Democrats, Georg von Vollmar, with belying the "inspiration" of social democracy, without which "a party such as ours cannot exist."

The struggle against open reformism and the theoretical revisionism advocated by Eduard Bernstein at the end of the 1890s reached its climax at the Dresden congress of 1903. Just as he condemned all deviations from the party's official radical creed, so too was Bebel unwilling to yield to left-wing pressure to indulge in extraparliamentary experiments and thus perhaps to bring repression of the party again. His stand was justified, for in election after election the party gained new adherents, and Bebel lived to see the day when, in 1912, it became, with 110 seats, the strongest group in the Reichstag. He died at Passugg, near Chur, Switzerland, on August 13, 1913.

Assessment

Bebel, as no other, embodied the tradition of the German Social Democratic Party. Already in 1882 Engels had described him as "a unique manifestation of the German, indeed of the European working class." A member of the Reichstag from 1867 almost continuously until his death, he achieved his most celebrated triumphs as a parliamentarian. Even his opponents could not withhold their respect in the face of his passionate honesty. A shrewd contemporary, Hellmut von Gerlach, suggested that in politics Bebel lived from hand to mouth: "His political aims were for the most distant future or for the immediate present"; he did not concern himself with what might lie between. This is true. For him and for the leading body of Social Democratic thought he represented, political activity essentially consisted in promoting as effectively as possible the politico-social interests of the working classes. His contradictory combination of futuristic revolutionary sentiment and a social policy rooted in the present reflects the equivocal position of his party under the conditions of the new German empire. This explains to a great extent both the strength of Bebel's position within the party and the political passivity of German social democracy, already noticeable before his death and fully revealed when, on the fall of the empire, the party had to face its first great political test.

BIBLIOGRAPHY. Bebel's autobiography, *Aus meinem Leben*, 3 vol. (1910–14; new ed., 1953; Eng. trans. of 1st ed., *My Life*, 1912), extends only up to 1882. Recent research includes ERNST SCHRAEPLER, *August Bebel, Sozialdemokrat im Kaiserreich* (1966). Schraepler has also published an *August-Bebel-Bibliographie* (1962). For Bebel's significance, see the instructive book by GUENTHER ROTH. *The Social Democrats in Imperial Germany* (1963).

(E.Ma.)

Beccaria, Cesare

The author of a celebrated volume on the reform of criminal justice and a pioneer in economic analysis, Cesare Bonesana Beccaria was born in Milan on March 15, 1738. His parents, although members of the Milanese aristocracy (Beccaria inherited the title of marchese), possessed only modest means. The essential traits of his character emerged at an early age. A highly volatile temperament resulted in periods of enthusiasm followed by depression and inactivity. He was reserved and somewhat taciturn in his social contacts but placed great value on his personal and family relationships. At the age of eight he was sent to the Jesuit school in Parma. Beccaria later described the education he received there as "fanatical" and stifling to "the development of human feelings." Although he revealed a mathematical aptitude, little in his student days gave indication of the remarkable intellectual achievements that were soon to follow. In 1758 he received a degree in law from the University of Pavia.



Beccaria, engraving by Carlo Faucci, 1766.
By courtesy of Raccolta Delle Stampe Achille Bertarelli, Milan

In 1760 Beccaria's proposed marriage to the 16-year-old Teresa Blasco encountered the obdurate opposition of his father. The following year the marriage took place without parental consent, and the young couple began their life together in poverty. The breach between father and son was ultimately repaired, and Beccaria and his wife were received into the family home. In 1762 a daughter, the first of his three children, was born.

Upon completion of his formal training Beccaria returned to Milan and was soon caught up in the intellectual ferment associated with the 18th-century Enlightenment. He joined with Count Pietro Verri in the organization of a literary society and participated actively in its affairs. In 1762 his first writing appeared, a pamphlet on monetary reform. Later he associated himself with the periodical *Il Caffè*, a journal modelled on the English essayist Joseph Addison's *Spectator*, and contributed several anonymous essays to its pages.

First studies of criminal law

In 1763 Verri suggested that Beccaria next undertake a critical study of the criminal law. Although he had had no experience in the administration of criminal justice, Beccaria accepted the suggestion; and in 1764 his great work *Dei delletti e delle pene* (Eng. trans., J.A. Farrer, *Crimes and Punishments*, 1880) was published. Almost immediately Beccaria, then only 26 years of age, became an international celebrity. The work enjoyed a remarkable success in France, where it was translated in 1766 and went through seven editions in six months. English, German, Polish, Spanish, and Dutch translations followed. In the two centuries following its appearance it has been translated into many other languages.

Beccaria's treatise is the first succinct and systematic statement of principles governing criminal punishment. Although many of the ideas expressed were familiar, and Beccaria's indebtedness to such writers as the French philosopher Montesquieu (which he generously acknowledged) is clear, the work nevertheless represents a major advance in criminological thought. The argument of the book is founded on the utilitarian principle that governmental policy should seek the greatest good for the

greatest number. He lashed out at the barbaric practices of his day: the use of torture and secret proceedings, the caprice and corruption of magistrates, brutal and degrading punishments. The objective of the penal system, he argued, should be to devise penalties only severe enough to achieve the proper purposes of security and order; anything in excess is tyranny. The effectiveness of criminal justice depends largely on the certainty of punishment rather than on its severity. Penalties should be scaled to the importance of the offense. Beccaria was the first modern writer to advocate the complete abolition of capital punishment and may therefore be regarded as a founder of the abolition movements that have persisted in most civilized nations since his day.

Beccaria's treatise exerted significant influence on criminal-law reform throughout western Europe. In England, the utilitarian philosopher and reformer Jeremy Bentham advocated Beccaria's principles, and the Benthamite disciple Samuel Romilly devoted his parliamentary career to reducing the scope of the death penalty. Legislative reforms in Russia, Sweden, and the Habsburg Empire were influenced by the treatise. The legislation of several American states reflected Beccaria's thought.

Major work

Although nothing Beccaria achieved in later life approaches the importance of the treatise, his subsequent career was fruitful and constructive. In 1768 he accepted the chair in public economy and commerce in the Palatine School in Milan, where he lectured for two years. His reputation as a pioneer in economic analysis is based primarily on these lectures, published posthumously in 1804 under the title *Elementi di economia pubblica* ("Elements of Public Economy"). He apparently anticipated some of the ideas of Adam Smith and Thomas Malthus, such as the concept of division of labour and the relations between food supply and population.

In 1771 he was appointed to the Supreme Economic Council of Milan and remained a public official for the remainder of his life. In his public role Beccaria became concerned with a large variety of measures, including monetary reform, labour relations, and public education. A report written by Beccaria influenced the subsequent adoption of the metric system in France.

Beccaria's later years were beset by family difficulties and problems of health. He apparently did not relish the role of celebrity. In 1766 he journeyed to Paris, where he was warmly greeted by the most distinguished figures of the day, but cut short his visit because of acute homesickness. His wife died in 1774 after a period of declining health. Three months later he remarried. Property disputes initiated by his two brothers and sister resulted in litigation that distracted him for many years. Beccaria's last months were saddened by events in France: although he had initially welcomed the French Revolution enthusiastically, he was shocked by the excesses of the terror. He died of apoplexy in Milan on November 28, 1794. Beccaria's rationality, versatility, and insistence on the unity of knowledge were typical of the intellectual life of his time. His treatise, the most important volume ever written on criminal justice, is still profitably consulted two centuries after its first appearance.

BIBLIOGRAPHY. P. VILLARI, "Discorso sulla vita e le opere di Cesare Beccaria," in *Le opere* (1854); C. CANTU, *Beccaria e il diritto penale* (1862); C.A. VIANELLO, *La vita e l'opera di Cesare Beccaria con scritti e documenti inediti* (1938); C. PHILLIPSON, *Three Criminal Law Reformers: Beccaria, Bentham, Romilly* (1923); M.T. MAESTRO, *Voltaire and Beccaria As Reformers of Criminal Law* (1942). For an assessment of Beccaria's economic writings, see J.A. SCHUMPETER, *History of Economic Analysis* (1954).

(F.A.A.)

Becket, Thomas

Thomas Becket, chancellor of England under Henry II and subsequently archbishop of Canterbury, baffled his contemporaries and has set historians at variance by his sudden change from a devoted servant of the King into an obstinate opponent of his policy, from a worldly clerk into an austere archbishop. The drama of his long quarrel with Henry, which ended only with his murder in Canterbury Cathedral, one of the most familiar and arresting episodes in English history, remains an issue on which the judgments of historians are divided.

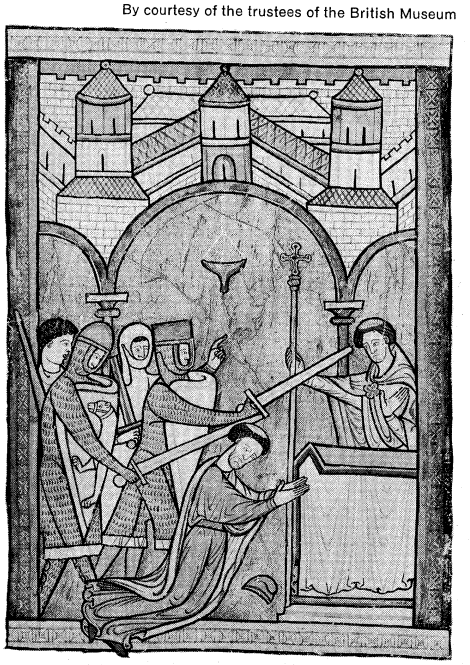
Thomas was born in 1118 in Cheapside, a busy market district of medieval London, to Norman parents of the merchant class. He was educated first at Merton priory, then in a City of London school, and finally at Paris. Deeply influenced in childhood by a devout mother who died when he was 21, Thomas entered adult life as a city clerk and accountant in the service of the sheriffs. After three years he was introduced by his father to Archbishop Theobald, a former abbot of Bec, of whose household he became a member. His colleagues were a distinguished company that included the political philosopher John of Salisbury, the Roman lawyer Vacarius, and several future bishops, including the Roger of Point l'Évêque, later archbishop of York. Thomas won Theobald's confidence, acted as his agent, and was sent by him to study civil and canon law at Bologna and Auxerre.

His contemporaries described Thomas as a tall and spare figure with dark hair and a pale face that flushed in excitement. His memory was extraordinarily tenacious, and though neither a scholar nor a stylist, he excelled in argument and repartee. He made himself agreeable to all around him, and his biographers attest that he led a chaste life—in this respect uninfluenced by the King.

In 1154 Theobald, as a reward of his services, appointed Thomas archdeacon of Canterbury, an important and lucrative post, and less than three months later recommended him to Henry as chancellor. Here Thomas showed to the full his brilliant abilities, razing castles, repairing the Tower of London, conducting embassies, and raising and leading troops in war. Trusted completely by the King, Thomas was compared by a biographer to Joseph under Pharaoh. To Henry himself Thomas was a welcome companion and intimate friend, both at court and in the chase, aiding the King in his policy of gathering all power into the hands of the monarchy, even when that policy went against claims of the church. Thomas, older than Henry by 15 years and celibate, may well have felt, at least initially, a quasi-paternal or elder-brother affection, mingled with admiration for Henry's talents and charm. He must also have enjoyed the satisfaction of moving in a rank of society to which he had not been born. Henry's attitude is less easy to identify, but the efficiency and intelligence of Thomas must have recommended him to a king surrounded by uneducated and at times truculent barons.

As chancellor

By courtesy of the trustees of the British Museum



Murder of Becket, illustration from an English psalter, c. 1200. In the British Museum.

Whether Becket was fully satisfied with his life as chancellor is another matter. Throughout his life Thomas gave with prodigality and acted with panache. The description of the procession of men, beasts, and carriages laden with objects of luxury that accompanied him as envoy to Paris in 1158 is one of the highlights of William FitzStephen's *Life of Thomas Becket*. This, and his customary splendour of clothing and furnishings, suited ill with his status as archdeacon. More serious in the eyes of contemporaries was his refusal to surrender his archdeaconry while neglecting its duties, and his extraction of scutage (payment in lieu of military service) at a high rate from ecclesiastical fiefs. Most serious to modern minds is his failure to visit the disapproving and dying Theobald when summoned. In general, there can be no doubt that in public affairs he was the King's man, even when Henry endeavoured to reassert what he claimed to be his ancestral rights.

Meanwhile, the great movement known as the Gregorian reform had spread from Italy to France and the Holy Roman Empire and had begun to influence English churchmen. In its program, free elections to clerical posts, inviolability of church property, freedom of appeal to Rome, and clerical immunity from lay tribunals were leading points. Under Henry I and Stephen, the archbishops had stood out for these reforms, sometimes with partial success. Henry II, however, undoubtedly aimed at a complete return to the practice of Henry I, who had strict control over the church. He had begun to press his claims, and his chancellor had aided him. With the death of Theobald in 1161, Henry hoped to appoint Thomas as archbishop and thus complete his program.

As archbishop

For almost a year after the death of Theobald the see of Canterbury was vacant. Thomas was aware of the King's intention and tried to dissuade him by warnings of what would happen. Henry persisted and Thomas was elected. Once consecrated, Thomas changed both his outlook and his way of life. He became devout and austere and embraced the integral program of the papacy and its canon law. This spectacular change has baffled historians, and several explanations have been attempted: that Thomas was intoxicated by his ambition to dominate or that he threw himself, as before, into a part he had agreed to play. It is simpler to suppose that he accepted at last the spiritual obligations he had ignored as chancellor and turned into a new channel his mingled energy, force of character, impetuosity, and ostentation. Greatly to Henry's displeasure, he immediately resigned the chancellorship but clung to the archdeaconry until forced by the King to resign. Henry had been in Normandy since August 1158, and on his return in January 1163 Thomas began the struggle by opposing a tax proposal and excommunicating a leading baron. More serious was his attitude in the matter of "criminous clerks." In western Europe, accused clerics for long had enjoyed the privilege of standing trial before the bishop rather than secular courts and usually received milder punishments than lay courts would assess. In England before the Conquest this was still the custom. If found guilty in an ecclesiastical court, clerics could be degraded or exiled but were not liable to death or mutilation. For 60 years after the Norman Conquest little is heard of clerical crime or its punishment, while on the Continent, Gregorian reformers were tending to emphasize the sole right of the church to try and punish clerks in major orders. The position of Thomas, that a guilty clerk could be degraded and punished by the bishop but should not be punished again by lay authority—"not twice for the same fault"—was canonically arguable and ultimately prevailed. Henry's contention that clerical crime was rife and that it was encouraged by the absence of drastic penalties commends itself to modern readers as a fair one. But it must be remembered that the King's motives were authoritarian and administrative rather than enlightened. Nevertheless, it may be thought that Thomas was ill-advised in his rigid stand on this point. The issue was joined in a council at Westminster (October 1163), but the crisis came at Clarendon (Wiltshire, January 1164), when the King demanded a global assent to all traditional royal rights, reduced to writing under 16 heads and known as the Constitutions of Clarendon. These asserted the King's right to punish criminous clerks, forbade excommunication of royal officials and appeals to Rome, and gave the King the revenues of vacant sees and the power to influence episcopal elections. Henry was justified in saying that these rights had been exercised by Henry I, but Thomas also was justified in maintaining that they contravened church law. Thomas, after verbally accepting the constitutions, revoked his assent and appealed to the Pope, then in France, who supported him while deprecating precipitate action.

The matter of the "criminous clerks"

Quarrel with Henry

Good relations between Thomas and Henry were now at an end; the Archbishop was summoned to trial by the King on a point of feudal obligation. At the Council of Northampton (October 6–13, 1164), it was clear that Henry intended to ruin and imprison or to force the resignation of the Archbishop. In this he was encouraged by some of the bishops, among them Gilbert Foliot, bishop of London. Thomas fled in disguise and took refuge with Louis VII of France. Pope Alexander III received him with honour but hesitated to act decisively in his favour in fear that he might throw Henry into the arms of the Holy Roman emperor Frederick I and his antipope, Paschal III.

Thomas' exile lasted for six years (November 2, 1164–December 2, 1170). He was joined by many of his distinguished household and lived ascetically, first at Pontigny Abbey and then, when Henry threatened the monks, at an abbey near Sens. Henry meanwhile had seized the properties of the Archbishop and his supporters and had exiled all Thomas' close relatives. In the following years several abortive attempts were made at reconciliation, but new acts of hostility by the King and declarations of excommunication hurled by Thomas at his opponents exacerbated the hostilities. The bishops were divided, but a majority of them, led by Foliot, were either hostile to Thomas or hesitant in supporting him. Papal legates more than once endeavoured to mediate, and the King and the Archbishop came together at Montmirail in 1169, only to part in anger. Thomas distrusted the King and was, in turn, hated by him. In the same year Henry put out additions to the Constitutions of Clarendon, virtually withdrawing England from papal obedience. Finally, in 1170, he had his eldest son crowned as co-king by the archbishop of York, Becket's old rival. This was a flagrant breach of papal prohibition and of the immemorial right of Canterbury to crown the king. Thomas, followed by the Pope, excommunicated all responsible. Henry, fearing an interdict for England, met Thomas at Fréteval (July 22), and it was agreed that Thomas should return to Canterbury and receive back all the possessions of his see. Neither party withdrew from his position regarding the Constitutions of Clarendon, which on this occasion were not mentioned. This "open-ended" concordat has remained an inexplicable event. Thomas returned to Canterbury (December 2) and was received with enthusiasm, but further excommunications of the hostile royal servants, refusal to lift the excommunication of Roger of York and Foliot, as well as his ready acceptance of tumultuous acclaim by the crowds infuriated Henry in Normandy. Some violent words of Henry were taken literally by four leading knights of the court, who proceeded swiftly to Canterbury (December 29), forced themselves into the Archbishop's presence, and, on his refusal to absolve the bishops, followed him into the cathedral. There, at twilight, after further altercation, they cut him down with their swords. His last words were an acceptance of death in defense of the church of Christ.

Martyrdom

Within a few days after Thomas' death, his tomb became a goal of pilgrimage, and he was canonized by Alexander III in 1173. In 1174 Henry did penance at Canterbury and was absolved. For almost four centuries, Becket's shrine was one of the most famous in Europe. Thomas was portrayed in illuminations and sculpture, and churches were dedicated to him throughout western Christendom.

Judgment on the character and actions of St. Thomas

has been varied. From his martyrdom till the reign of Henry VIII he was the "blisful martyr" of Chaucer's pilgrims, who had heroically defied a tyrant. Henry VIII despoiled his shrine, burned his bones, and erased his name from all service books. Thenceforth Thomas was a hero to Catholics and a traitor to Protestants. Many recent historians, impressed by the legal and administrative reforms of Henry II, have seen Thomas as an ambitious and fanatical nuisance. Certainly there is room for debate, for both Thomas and his king were remarkable men with complex characters. If Henry had moral failings and made private and political miscalculations, Thomas can rightly be accused, at various moments of his life, of worldly behaviour, ostentation, impetuosity, weakness, and violent language. If Henry was ill-advised in committing his claims to writing at Clarendon and in crowning his son, Thomas was equally ill-advised in needlessly opposing the King in 1163 and in wavering between compliance and intransigence when careful diplomacy might have won out. But his courage and sincerity cannot be doubted, and in the quarrel between church and state he gave his life for what he took to be a vital issue.

BIBLIOGRAPHY. The nine contemporary Latin *Lives*, mostly by men who knew the Archbishop, in *Materials for the History of Thomas Becket, Archbishop of Canterbury*, vol. 1–4, "Rolls Series" (1875–85), are described and translated in a selection by D.C. DOUGLAS and G.W. GREENAWAY, *English Historical Documents II*, pp. 698–776 (1953). Becket's letters and many addressed to him are in *Materials*, vol. 5–7 (not translated). There is an excellent short account by KATE NORGATE in the *Dictionary of National Biography* (1898, reprinted 1937–38); and brief lives, *Thomas Becket*, by R. WINSTON (1967), with bibliography; and D. KNOWLES (1970). See also D. KNOWLES, *The Episcopal Colleagues of Archbishop Thomas Becket* (1951) and *Archbishop Thomas Becket: A Character Study* (1949). The most detailed account, needing revision on some points, is RAYMONDE FOREVILLE, *L'Église et la royauté en Angleterre sous Henri II, 1154–1189* (1943).

(M.D.K.)

Beckett, Samuel

Among the major writers of the 20th century, Samuel Beckett stands out by virtue of the uncompromising austerity and purity of his approach to literature. His work, though seasoned with a bitter humour, was a dedicated and courageous exploration of his stark perceptions of the transitoriness and insignificance of human existence in a bleak universe. Whether his work took the form of narrative prose, poetry, criticism, or drama, this boldness in seeking ultimate truths made his writings invaluable documents of human experience and of the workings of man's consciousness. At the same time, the writings represented important contributions to existential philosophical thought, which is concerned with human experience at its most concrete and specific as the basis of all more general concepts and values. Writing in both French and English, and translating his works from one to the other, Beckett was an incomparable stylist in both. His output in verse was small, but his narrative and dramatic prose was so delicately phrased, so subtle in its rhythms, and so intricately structured that it, too, must be considered poetry.

The restless search. Samuel Beckett was born on April 13, 1906, at Foxrock, a suburb of Dublin. Like his fellow Irish writers George Bernard Shaw, Oscar Wilde, and William Butler Yeats, he came from a Protestant, Anglo-Irish background. At the age of 14 he went to the Portora Royal School, in what became Northern Ireland, a school that catered to the Anglo-Irish middle classes.

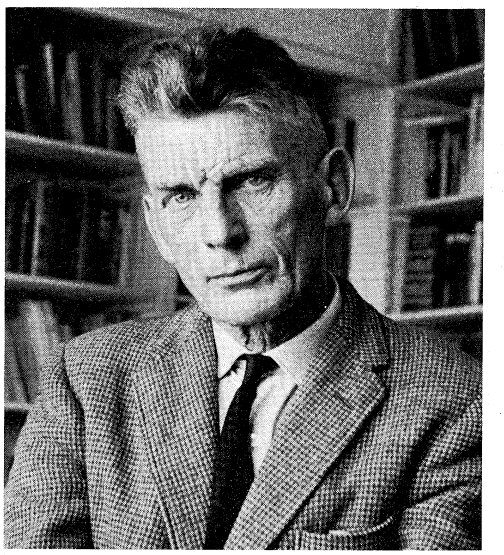
From 1923 to 1927, he studied Romance languages at Trinity College, Dublin, where he received his bachelor's degree. After a brief spell of teaching at a school in Belfast, he became a reader in English at the École Normale Supérieure in Paris in 1928. Here he met the self-exiled Irish writer James Joyce, the author of the controversial and seminally modern novel *Ulysses*, and became a member of his circle. Contrary to often-repeated reports, however, he never served as Joyce's secretary. He returned to Ireland in 1930 to take up a post as lecturer in French at Trinity College, but after only four terms he resigned, in December 1931, and embarked upon a period of restless travel in London, France, Germany, and Italy.

In 1937, Beckett decided to settle in Paris. As a citizen of a country that was neutral in World War II, he was able to remain there even after the occupation of Paris by the Germans, but he joined an underground resistance group in 1941. When, in 1942, he received news that members of his group had been arrested by the Gestapo, he immediately went into hiding and eventually moved to the unoccupied zone of France. Until the liberation of the country, he supported himself as an agricultural labourer. Beckett in Paris

In 1945, he returned to Ireland but volunteered for the Irish Red Cross and was back in France as an interpreter in a military hospital in Saint-Lô, Normandy. In the winter of 1945, he finally returned to Paris.

Production of the major works. There followed a period of intense creativity, the most concentratedly fruitful period of Beckett's life. His relatively few prewar publications included two essays on Joyce and the French novelist Marcel Proust. The volume *More Pricks than Kicks* (1934) contained ten stories describing episodes in the life of a Dublin intellectual, Belacqua Shuah, and the novel *Murphy* (1938) concerns an Irishman in London who escapes from a girl he is about to marry to a life of contemplation as a male nurse in a mental institution. His two slim volumes of poetry were *Whoroscope* (1930), a poem on the French philosopher René Descartes, and the collection *Echo's Bones* (1935). A number of short stories and poems were scattered in various periodicals. During his years in hiding in unoccupied France, Beckett also completed another novel, *Watt*, which was not published until 1953. After his return to Paris, between 1946 and 1949, Beckett produced a number of stories, the major prose narratives *Molloy* (1951; Eng. trans., 1955), *Malone meurt* (1951; Eng. trans., *Malone Dies*, 1956), and *L'Innommable* (1953; Eng. trans., *The Unnamable*, 1958) and two plays, the unpublished three-act *Eleutheria* and *En attendant Godot* (1952; Eng. trans., *Waiting for Godot*, 1954).

It was not until 1951, however, that these works saw the light of day. After many refusals, Mme Beckett, who had been in Beckett's resistance group during the war, finally succeeded in finding a publisher for *Molloy*. When this book not only proved a modest commercial success but also was received with enthusiasm by the French critics, the same publisher brought out the two other novels and *En attendant Godot*. It was with the amazing success of this play, however, at the small Théâtre de Babylone in Paris, in January 1953, that Beckett's rise to world fame began. Following this, Beckett continued writing, but more slowly than in the immediate postwar years. Plays First success



Beckett, 1965.

for the stage and radio and a number of prose works occupied much of his attention. He continued to live in Paris, but most of his writing was done in a small house secluded in the Marne Valley, a short drive from Paris. His total dedication to his art extended to his complete avoidance of all personal publicity, of appearances on radio or television, and of all journalistic interviews. When, in 1969, he received the Nobel Prize for Literature, he accepted the award but declined the trip to Stockholm to avoid the public speech at the ceremonies.

Continuity of his philosophical explorations. Beckett's writing reveals his own immense learning. It is full of subtle allusions to a multitude of literary sources as well as many philosophical and theological writers. The dominating influences on his thought are undoubtedly the Italian poet Dante, the French philosopher René Descartes, the 17th-century Dutch philosopher Arnold Geulincx—a pupil of Descartes who dealt with the question of how the physical and spiritual side of man interact—and, finally, his fellow Irishman and revered friend, James Joyce. But it is by no means essential for the understanding of Beckett's work to be aware of all the literary, philosophical, and theological allusions.

The widespread idea, fostered by the popular press, that Beckett's work is concerned primarily with the sordid side of human existence, with tramps and with cripples who inhabit trash cans, is a fundamental misconception. He deals with human beings in such extreme situations not because he is interested in the sordid and diseased aspects of life but because he concentrates on the essential aspects of human experience. The subject matter of so much of the world's literature—the social relations between individuals, their manners and possessions, their struggles for rank and position, or the conquest of sexual objects—appears to Beckett as mere external trappings of existence, the accidental and superficial aspects that mask the basic problems and the basic anguish of the human condition. The basic questions for Beckett seem to be these: How can we come to terms with the fact that, without ever having asked for it, we have been thrown into the world, into being? And who are we; what is the true nature of our self? What does a human being mean when he says "I"?

Basic questions of Beckett's work

What appears to the superficial view as a concentration on the sordid thus emerges as an attempt to grapple with the most essential aspects of the human condition. The two heroes of *Waiting for Godot*, for instance, are frequently referred to by critics as tramps, yet they were never described as such by Beckett. They are merely two human beings in the most basic human situation of being in the world and not knowing what they are there for. Since man is a rational being and cannot imagine that his being thrown into any situation should or could be entirely pointless, the two vaguely assume that their presence in the world, represented by an empty stage with a solitary tree, must be due to the fact that they are waiting for someone. But they have no positive evidence that this person, whom they call Godot, ever made such an appointment—or, indeed, that he actually exists. Their patient and passive waiting is contrasted by Beckett with the mindless and equally purposeless journeyings that fill the existence of a second pair of characters. In most dramatic literature the characters pursue well-defined objectives, seeking power, wealth, marriage with a desirable partner, or something of the sort. Yet, once they have attained these objectives, are they or the audience any nearer answering the basic questions that Beckett poses? Does the hero, having won his lady, really live with her happily ever after? That is apparently why Beckett has chosen to discard what he regards as the inessential questions and begins where other writing leaves off.

This stripping of reality to its naked bones is the reason that Beckett's development as a writer was toward an ever greater concentration, sparseness, and brevity. His two earliest works of narrative fiction, *More Pricks than Kicks* and *Murphy*, abound in descriptive detail. In *Watt*, the last of Beckett's novels written in English, the milieu is still recognizably Irish, but most of the action takes place in a highly abstract, unreal world. Watt, the hero, takes service with a mysterious employer, Mr. Knott, works for a time for this master without ever meeting him face to face, and then is dismissed. The allegory of man's life in the midst of mystery is plain.

Most of Beckett's plays also take place on a similar level of abstraction. *Fin de partie* (one-act, 1957; Eng. trans., *Endgame*, 1958) describes the dissolution of the relation between a master, Hamm, and his servant, Clov. They inhabit a circular structure with two high windows—perhaps the image of the inside of a human skull. The action might be seen as a symbol of the dissolution of a human personality in the hour of death, the breaking of the bond between the spiritual and the physical sides of man. In *Krapp's Last Tape* (one-act, first performed 1958), an old man listens to the confessions he recorded in earlier and happier years. This becomes an image of the mystery of the self, for to the old Krapp the voice of the younger Krapp is that of a total stranger. In what sense, then, can the two Krapps be regarded as the same human being? In *Happy Days* (1961), a woman, literally sinking continually deeper into the ground, nonetheless continues to prattle about the trivialities of life. In other words, perhaps, as one gets nearer and nearer death, one still pretends that life will go on normally forever.

Narrative prose works

In his trilogy of narrative prose works—they are not, strictly speaking, novels as usually understood—*Molloy*, *Malone Dies*, and *The Unnamable*, as well as in the collection *Stories and Texts for Nothing*, (1967), Beckett raises the problem of the identity of the human self from, as it were, the inside. This basic problem, simply stated, is that when I say "I am writing," I am talking about myself, one part of me describing what another part of me is doing. I am both the observer and the object I observe. Which of the two is the real "I"? In his prose narratives, Beckett tries to pursue this elusive essence of the self, which, to him, manifests itself as a constant stream of thought and of observations about the self. One's entire existence, one's consciousness of oneself as being in the world, can be seen as a stream of thought. *Cogito ergo sum* is the starting point of Beckett's favourite philosopher, Descartes: "I think; therefore, I am." To catch the essence of being, therefore, Beckett tries to capture the essence of the stream of consciousness that is one's being. And what he finds is a constantly receding chorus of observers, or storytellers, who, immediately on being observed, become, in turn, objects of observation by a new observer. Molloy and Moran, for example, the pursued and the pursuer in the first part of the trilogy, are just such a pair of observer and observed. Malone, in the second part, spends his time while dying in making up stories about people who clearly are aspects of himself. The third part reaches down to bedrock. The voice is that of someone who is unnamable, and it is not clear whether it is a voice that comes from beyond the grave or from a limbo before birth. As we cannot conceive of our consciousness not being there—"*I* cannot be conscious that *I* have ceased to exist"—therefore consciousness is at either side open-ended to infinity. This is the subject also of the play *Play* (first performed 1963), which shows the dying moments of consciousness of three characters, who have been linked in a trivial amorous triangle in life, lingering on into eternity.

The humour and mastery. In spite of Beckett's courageous tackling of the ultimate mystery and despair of human existence, he is essentially a comic writer. In a French farce, laughter will arise from seeing the frantic and usually unsuccessful pursuit of trivial sexual gratifications. In Beckett's work, as well, a recognition of the triviality and ultimate pointlessness of most human strivings, by freeing the viewer from his concern with senseless and futile objectives, should also have a liberating effect. The laughter will arise from a view of pompous and self-important preoccupation with illusory ambitions and futile desires. Far from being gloomy and depressing, the ultimate effect of seeing or reading Beckett is one of cathartic release, an objective as old as theatre itself.

Command of form

Technically, Beckett is a master craftsman, and his sense of form is impeccable. *Molloy* and *Waiting for Godot*, for example, are constructed symmetrically, in two parts

that are mirror images of one another. In his work for the mass media, Beckett also showed himself able to grasp intuitively and brilliantly the essential character of their techniques. His radio plays, such as *All That Fall* (1957), are models in the combined use of sound, music, and speech. The short television play *Eh Joe!* (1967) exploits the television camera's ability to move in on a face and the particular character of small-screen drama. Finally, his film script *Film* (1967) creates an unforgettable sequence of images of the observed self trying to escape the eye of its own observer.

Beckett's later works tended toward extreme concentration and brevity. *Come and Go* (1967), a playlet, or "dramaticule," as he called it, contains only 121 words that are spoken by the three characters. The prose fragment "Lessness" consists of but 60 sentences, each of which occurs twice. His series *Acts Without Words* are exactly what the title denotes. Such brevity is merely an expression of Beckett's determination to pare his writing to essentials, to waste no words on trivia.

MAJOR WORKS

NARRATIVE PROSE: *More Pricks Than Kicks* (1934), *Murphy* (1938), and *Watt* (written *c.* 1942–44, published 1953), all in English; a trilogy consisting of *Molloy* (Eng. trans. by Beckett and Patrick Bowles, 1955, *Malone meurt* (1951; *Malone Dies*, trans. by Beckett, 1956), and *L'Innomable* (1953; *The Unnamable*, trans. by Beckett, 1958); *Nouvelles et textes pour rien* (1955; *Stories and Texts for Nothing*, in *No's Knife: Collected Shorter Prose 1945–66*, 1967); *Comment c'est* (1961; *How It Is*, trans. by Beckett, 1964); *Imagination morte imaginez* (1965; *Imagination Dead Imagine*, trans. by Beckett, 1965), a short novel, also included in *Collected Shorter Prose; Assez* (1966; *Enough*, in *Collected Shorter Prose*); *Bing* (1966; *Ping*, in *Collected Shorter Prose*); *Têtesmortes* (1967).

PLAYS AND DRAMATIC PIECES: *En attendant Godot* (1952: *Waiting for Godot*, trans. by Beckett, 1954); *Fin de partie* (1957; *Endgame*, trans. by Beckett, 1958), one-act play; *Acte sans paroles I* (1957; *Act Without Words*, trans. by Beckett, 1958); *All That Fall* (1957; in *Krapp's Last Tape and Other Dramatic Pieces*, 1960), one-act play for radio; *Krapp's Last Tape* (1959), one-act play; *Embers* (1959), for radio; *Happy Days* (1961; French trans. by Beckett, *Oh les Beaux Jours*, 1963); *Play* (1964; French trans. by Beckett, *Comédie*, published in *Comédie et actes divers*, 1966); *Words and Music* (1964), for radio—music and voices; *Cascando* (1964), for radio—music and voices; *Come and Go*, first published in Beckett's French trans., *Va et vient* (1966; English original first published in a bilingual German collection 1966, separately 1967), a "dramaticule" lasting only a few minutes; *Eh Joe* (1966), for television, first published in Beckett's French trans., *Dis Joe* (1966, in English 1967); *Film* (1967), film script.

OTHER WORKS (VERSE): *Whoroscope* (1930); *Echo's Bones* (1935); *Poems in English* (1961). (ESSAYS): "Dante . . . Bruno. Vico . . . Joyce," (1929), in *Our Exagmination Round his Factification . . .; Proust* (1931); *Proust/Three Dialogues* (1965).

BIBLIOGRAPHY. RAYMOND FEDERMAN and JOHN FLETCHER, *Samuel Beckett: His Works and His Critics: An Essay in Bibliography, 1929–1966* (1970), is an exhaustive bibliography of the author's writings and writings about him. Many critical studies have been devoted to Beckett; among these HUGH KENNER, *Samuel Beckett: A Critical Study* (1962); RUBY COHN, *Samuel Beckett: The Comic Gamut* (1962); JOHN FLETCHER, *The Novels of Samuel Beckett* (1964); and on Beckett's plays, MARTIN ESSLIN, *The Theatre of the Absurd* (1961), can serve as a basis for an approach to Beckett. Two anthologies of shorter articles of important Beckett criticism are *Samuel Beckett: A Collection of Critical Essays*, ed. by MARTIN ESSLIN (1965); and *Samuel Beckett Now*, ed. by MELVIN J. FRIEDMAN (1970). Beckett's early writings are the subject of RAYMOND FEDERMAN, *Journey to Chaos: Samuel Beckett's Early Fiction* (1965); and LAWRENCE E. HARVEY, *Samuel Beckett: Poet and Critic* (1970). JOHN FLETCHER, *Samuel Beckett's Art* (1967), is a valuable study of Beckett's technique as a writer. *Beckett at Sixty* (1967), is a collection of personal reminiscences by friends and admirers. RICHARD N. COE, *Beckett* (1964), is a very useful brief introduction. Two books devoted to the interpretation of single works are *Casebook on Waiting for Godot*, ed. by RUBY COHN, a collection of early reviews and critical essays on the play (1967); and *Twentieth Century Interpretations of Endgame*, ed. by B.G. CHEVIGNY (1969).

(M.J.E.)

Becquerel, Henri

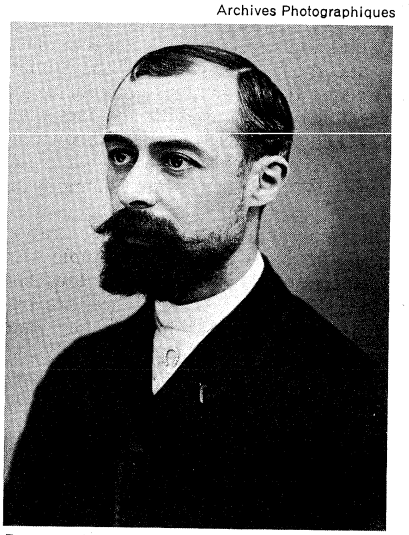
As the discover of radioactivity and as a member of a scientific family extending through several generations, Henri Becquerel was probably the most prominent French physicist of the decade spanning the turn of the century. He was born in Paris on December 15, 1852, and raised in the centre of French scientific activity, for his father, Alexandre-Edmond Becquerel, was professor of applied physics at the Muséum National d'Histoire Naturelle. Henri succeeded to his father's professorship in the museum in 1892, the chair of which his grandfather, Antoine-César Becquerel, had been the first occupant; Henri's son, Jean Becquerel, extended the chain to four generations.

Early years

After his early schooling at the Lycée Louis-le-Grand, Henri received his formal scientific education at the École Polytechnique (1872–74) and engineering training at the École des Ponts et Chaussées (Bridges and Highways School; 1874–77). In addition to his teaching and research posts, Becquerel was for many years an engineer in the Department of Bridges and Highways, being appointed chief engineer in 1894. His first academic situation was in 1876 as assistant teacher at the École Polytechnique, where in 1895 he succeeded to the chair of physics. Concurrently, he was assistant naturalist to his father at the museum, where he also assumed the physics professorship upon his father's death.

Electricity, magnetism, optical phenomena, and energy were major areas of physical investigation during the 19th century. For several years the young man's research was concerned with the rotation of plane-polarized light by magnetic fields, a subject opened by Michael Faraday and to which Henri's father had also contributed. Henri then concerned himself with infrared radiation, examining among other things, the spectra of different phosphorescent crystals under infrared stimulation. Of particular significance, he extended the work of his father by studying the relation between absorption of light and emission of phosphorescence in some uranium compounds.

Archives Photographiques



Becquerel.

By 1896, Henri was an accomplished and respected physicist—a member of the Académie des Sciences since 1889—but more important than his research thus far were his expertise with phosphorescent materials, his familiarity with uranium compounds, and his general skill in laboratory techniques, including photography. Together, these were to place the discovery of radioactivity within his reach.

Work with radiation

At the end of 1895, Wilhelm Röntgen discovered X-rays. Becquerel learned that the X-rays issued from the area of a glass vacuum tube made fluorescent when struck by a beam of cathode rays. He undertook to investigate whether there was some fundamental connection between this invisible radiation and visible light such

that all luminescent materials, however stimulated, would also yield X-rays. To test this hypothesis, he placed phosphorescent crystals upon a photographic plate that had been wrapped in opaque paper so that only a penetrating radiation could reach the emulsion. He exposed his experimental arrangement to sunlight for several hours, thereby exciting the crystals in the customary manner. Upon development, the photographic plate revealed silhouettes of the mineral samples, and, in subsequent experiments, the image of a coin or metal cutout interposed between the crystal and paper wrapping. Becquerel reported this discovery to the Académie des Sciences at its session on February 24, 1896, noting that certain salts of uranium were particularly active.

He thus confirmed his view that something very similar to X-rays was emitted by this luminescent substance at the same time it threw off visible radiation. But the following week Becquerel learned that his uranium salts continued to eject penetrating radiation even when they were not made to phosphoresce by the ultraviolet in sunlight. To account for this novelty he postulated a long-lived form of invisible phosphorescence; when he shortly traced the activity to uranium metal, he interpreted it as a unique case of metallic phosphorescence.

During 1896 Becquerel published seven papers on radioactivity, as Marie Curie later named the phenomenon; in 1897, only two papers; and in 1898, none. This was an index of both his and the scientific world's interest in the subject, for the period saw studies of numerous radiations (*e.g.*, cathode rays, X-rays, Becquerel rays, "discharge rays," canal rays, radio waves, the visible spectrum, rays from glowworms, fireflies, and other luminescent materials), and Becquerel rays seemed not especially significant. The far more popular X-rays could take sharper shadow photographs and faster. It required the extension in 1898 of radioactivity to another known element, thorium (by Gerhard Carl Schmidt and independently by Marie Curie), and the discovery of new radioactive materials, polonium and radium (by Pierre and Marie Curie and their colleague, Gustave Bémont), to awaken the world and Becquerel to the significance of his discovery.

Further contributions

Returning to the field he had created, Becquerel made three more important contributions. One was to measure, in 1899 and 1900, the deflection of beta particles, a constituent of the radiation, in both electric and magnetic fields. From the charge to mass value thus obtained, he showed that the beta particle was the same as Joseph John Thomson's recently identified electron. Another discovery was the circumstance that the allegedly active substance in uranium, uranium X, lost its radiating ability in time, while the uranium, inactive when freshly prepared, regained its activity. When Ernest Rutherford and Frederick Soddy found similar decay and regeneration in thorium X and thorium, they were led to the transformation theory of radioactivity, which explained the phenomenon as a subatomic chemical change in which one element spontaneously transmutes into another. Becquerel's last major achievement concerned the physiological effect of the radiation. Others may have noticed this before him, but his report in 1901 of the burn caused when he carried an active sample of the Curies' radium in his vest pocket inspired investigation by physicians, leading ultimately to medical use.

For his discovery of radioactivity, Becquerel shared the 1903 Nobel Prize for Physics with the Curies; he was also honoured with other medals and memberships in foreign societies. His own Académie des Sciences elected him its president and one of its permanent secretaries. He died on August 25, 1908.

BIBLIOGRAPHY. There is no full-length biography of Becquerel in English. ALFRED ROMER's article on him in the *Dictionary of Scientific Biography* provides further details. Aspects of Becquerel's work are discussed by LAWRENCE BADASH in the following articles: "Chance Favors the Prepared Mind: Henri Becquerel and the Discovery of Radioactivity," *Archs. Int. Hist. Sci.*, 18:55–66 (1965); "Becquerel's 'Unexposed' Photographic Plates," *Isis*, 57:267–269 (1966); and "Radioactivity Before the Curies," *Am. J. Phys.*, 33:128–135 (1965). Becquerel summarized his own work on radioactivity in vol. 46 of *Mémoires de l'Académie des Sciences* (1903). Good historical surveys of the development of radioactivity are ALFRED ROMER, *The Restless Atom* (1960); and (ed.), *The Discovery of Radioactivity and Transmutation* (1964), and *Radiochemistry and the Discovery of Isotopes* (1970); the latter two contain English translations of numerous original paper.

(L.Ba.)

Beekeeping

Beekeeping is the care and management of swarms of honeybees. They may be kept for their honey and other products or their services as pollinators of fruit and vegetable blossoms or as a hobby. Honeybees may be kept in large cities and villages, on farms and rangelands, in forests and deserts, from the Arctic to the Equator.

In antiquity man knew that bees produce delicious honey, that they sting, and that they increase their numbers by swarming. By the 17th century he had learned the value of smoke in controlling them and had developed the screen veil as protection against stings. From the 17th to the 19th century, the key discoveries upon which modern beekeeping is founded were made. These included the mystery of the queen bee as the mother of all the occupants of the hive, her curious mating technique, parthenogenetic development, the movable frame hives, and the fact that bees rear a new queen if the old one disappears.

Given this knowledge man was able to divide a colony instead of relying on natural swarming. Then the development of the wax-comb foundation, the starter comb on which bees build straight, easily-handled combs, and the discovery that honey can be centrifuged or extracted from them and the combs reused, paved the way for large-scale honey production and modern commercial beekeeping. The identification of bee diseases and their control with drugs, the value of pollen and pollen substitutes in producing strong colonies, and the artificial insemination of queens have increased the honey-production efficiency of colonies.

HONEYBEES AND THEIR COLONIES

Honeybees. Honeybees belong to the order Hymenoptera and to one of the *Apis* species. (For a complete discussion of honeybees, see the article HYMENOPTERA.) Honeybees are social insects noted for providing their nests with large amounts of honey. A colony of honeybees is a highly complex cluster of individuals that functions virtually as a single organism. It usually consists of the queen bee, a fertilized female capable of laying a thousand or more eggs per day; from a few to 60,000 sexually undeveloped females, the worker bees; and from none to 1,000 male bees, or drones. The female of most species of bees is equipped with a venomous sting.

Honeybees are not domesticated; those living in a man-made domicile called a beehive or hive are no different from those living in a colony in a tree.

Honeybees collect nectar, a sugary solution, from nectaries in blossoms and sometimes from nectaries on the leaves or stems of plants. Nectar may consist of 50 to 80 percent water, but when the bees convert it into honey it will contain only about 16 to 18 percent water. Sometimes they collect honeydew, an exudate from certain plant-sucking insects, and store it as honey. The primary carbohydrate diet of bees is honey. They also collect pollen, the dustlike male element, from the anthers of flowers. Pollen provides the essential proteins necessary for the rearing of young bees. In the act of collecting nectar and pollen to provision the nest, the bees pollinate the flowers they visit. Honeybees also collect propolis, a resinous material from buds of trees, for sealing cracks in the hive or for covering foreign objects in the hive that they cannot remove. They collect water to air-condition the hive and to dilute the honey when they consume it.

A populous colony in a desirable location may, in a year's time, collect and carry into the hive as much as 1,000 pounds of nectar, water, and pollen.

Honeycomb

Bees secrete beeswax in tiny flakes on the underside of the abdomen and mold it into honeycomb, thin-walled,

back-to-back, six-sided cells. The use of the cell varies depending on the needs of the colony. Honey or pollen may be stored in some cells, while the queen lays eggs, normally one per cell, in others. The area where the bees develop from the eggs is called the broodnest. Generally, honey is stored toward the top of the combs and pollen in cells around the broodnest below the honey.

The bees maintain a uniform temperature of about 93° F (34° C) in the broodnest regardless of outside temperature. The colony can survive daily maximum temperatures of 120° F (49° C) if water is available with which they can air-condition the cluster. When the temperature falls below about 57° F (14° C) the bees cease flying, form a tight cluster to conserve heat, and await the return of warm weather. They can survive for several weeks in temperatures of −50° F (−46° C).

When summer flowers bloom in profusion, the queen's egglaying is stimulated, the cluster expands, and honey accumulates in the combs. When the large number of young bees emerge, the domicile becomes crowded.

Swarming. When the colony becomes crowded with adult bees and there are insufficient cells in which the queen can lay large numbers of eggs, the worker bees select a dozen or so tiny larvae that would otherwise develop into worker bees. These larvae are fed copiously with royal jelly, a whitish food with the consistency of mayonnaise, produced by certain brood-food glands in the heads of the worker bees. The cell in which the larva is developing is drawn out downward and enlarged to permit development of the queen. Shortly before these virgin queens emerge as adults from their queen cells, the mother queen departs from the beehive with the swarm. Swarming usually occurs during the middle of a warm day, when the queen and a portion of the worker bees (usually from 5,000 to 25,000) suddenly swirl out of the hive and into the air. After a few minutes' flight, the queen alights, preferably on a branch of a tree but sometimes on a roof, a parked automobile, or even a fire hydrant. All the bees settle into a tight cluster around her while a handful of scouts reconnoitre a new homesite.

When the scout bees have located a new domicile, the cluster breaks, the swarm takes to the air and in a swirling mass proceeds to the new home. Swarming is the bees' natural method of propagation or increase.

Queen bee. Back in the parent colony, the first queen to emerge after the mother queen departs with the swarm immediately attempts to destroy the others. If two or more emerge at the same time, they fight to the death. When the surviving virgin is about a week old, she soars off on her mating flight; she frequently mates with more than one drone while in the air. She may repeat the mating flights for two or three successive days, after which she begins egg laying. She rarely ever leaves the hive again except with a swarm. Sufficient sperm are stored in her sperm pouch, or spermatheca, to fertilize all the eggs she will lay for the rest of her life. The drones die in the act of mating.

The queen usually lives about a year, although occasionally one will live for several years. If she is accidentally killed or begins to falter in her egg-laying efficiency, the worker bees will rear a "supersedure" queen that will mate and begin egg laying without a swarm emerging. She ignores the mother queen, who soon disappears from the colony.

Worker bees. Worker bees live about six weeks during the active season but may live for several months if they emerge as adults in the fall and spend the winter in the cluster. As the name implies, worker bees do all of the work of the hive, except the egg laying.

Drones. Drones are reared only when the colony is populous and there are plentiful sources of nectar and pollen. They usually live a few weeks, but are driven from the hive to perish when fall or an extended period of adversity comes upon the colony. The only duty of the drone is to mate with the queen.

The queen can lay drone (unfertilized) eggs in the drone cells. If she is not allowed to mate or if her supply of sperm is exhausted, she will lay unfertilized eggs in worker cells. The development of unfertilized eggs into adult drones is known as parthenogenesis. Occasionally a colony may become queenless and unable to develop another queen. Then, some of the worker bees begin to lay eggs, often several to a cell, and these develop into drones. A colony that has developed laying workers is difficult to requeen with a laying queen.

COLONY MANIPULATION

The yearly work cycle. The beekeeper's year starts in early fall. At that time he requeens the colonies whose queens are not producing adequate amounts of brood and makes sure that each colony has sufficient stores: at least 50 pounds of honey and several frames filled with pollen. Some beekeepers also feed the drug fumagillin to reduce possible damage to the adult bees by nosema disease (see below *Disease and Pest Control*). The colonies need a sunny exposure and protection from cold winds. Some beekeepers in northern and mountainous areas wrap their colonies with insulating material in winter. A few beekeepers kill their bees in the fall, harvest the honey, store the empty equipment, then restock with a two or three pound package of bees and a young queen the following spring.

Wintering

If the colonies are well prepared in the fall they need little attention during the winter. But in early spring, an examination of the colonies by the beekeeper is important. Frequently, strong colonies exhaust their food supply and starve only a few days before flowers begin to bloom in abundance. Only a few pounds of sugar syrup, 50–50 sugar water, or a honey-filled comb from another more prosperous colony might save such a starving colony. Again fumagillin may be fed to the colony, and some beekeepers also feed a cake of pollen substitute or pollen supplement. Honey is not fed to the colonies unless the beekeeper is sure about its source. Honey from colonies affected by the brood disease American foulbrood could infect his colonies and cause a serious loss.

As the spring season advances, the cluster size increases from the low population of 10,000 to 20,000 bees that survived the winter. To accommodate the increased size of the cluster and broodnest, the keeper adds more supers, or boxes of combs. If the combs are so manipulated that the queen can continually expand her egglaying area upward, the colony is unlikely to swarm. This can be achieved by placing empty combs or combs in which brood is about ready to emerge at the top of the cluster and combs filled with eggs or young brood toward the lower part of the broodnest. The beekeeper wants the colony to reach its peak of population, 50,000 to 60,000 bees, at the beginning of the major nectar flow.

The bees in a swarm, having departed the hive with a full stomach of honey, rarely sting. The usual way to capture them is to place a hive or upturned box beneath or nearby, then shake or smoke the bees to force the queen and a majority of the bees into it. The others follow. After the swarm is safely inside the box it can be removed to a permanent location.

Regulations governing the keeping of bees usually require the bees to be kept in hives with movable combs. If the bees are captured in a box they are generally transferred into a movable-frame hive within a few days so the new honey and comb will not be lost in the transfer.

Requeening a colony. When a beekeeper requeens a colony, he removes the failing or otherwise undesirable queen and places a new one in a screen cage in the broodnest. After a few days the colony becomes adjusted to her and she can be released. A strange queen placed in the cluster without this temporary protection usually will be killed at once by the workers. Queens usually are shipped in individual cages of about three cubic inches with about half a dozen attendant bees and a ball of specially prepared sugar candy plugging one end of the cage. When the cage is placed in the hive, the bees from both sides eat the candy. By the time the candy is consumed and the bees reach each other, their odours have become indistinguishable, the queen emerges from the cage into the colony and begins her egglaying duties.



A beekeeper, wearing a veil and holding a hive tool, lifts a frame from a super. A smoker is attached to the side of the hive and a super has been removed.
By courtesy of the U.S. Department of Agriculture

Beekeeping equipment. Standard tools of the beekeeper are: the smoker to quell the bees; a veil to protect his face; gloves for the novice or the person sensitive to stings; a blunt steel blade called a hive tool, for separating the frames and other hive parts for examination; the uncapping knife, for opening the cells of honey; and the extractor, for centrifuging the honey from the cells (see the Figure).

Bee stings. The worker bee sting is barbed, and in the act of stinging it is torn from the bee. It has a venom-filled poison sac and muscles attached that continue to work the sting deeper into the flesh for several minutes and increase the amount of venom injected. To prevent this, the sting should be scraped loose at once. Bee stings are painful, and no one becomes immune to the pain. Immunity to the swelling is usually built up after a few stings, however.

Normal reaction to a bee sting is immediate, intense pain at the site of the sting. This lasts for a minute or two and is followed by a reddening, which may spread an inch or more. Swelling may not become apparent until the following day. Occasionally, acute allergic reactions develop from a sting, usually with persons who have other allergic problems. Such a reaction becomes evident in less than an hour and may consist of extreme difficulty in breathing, heart irregularity, shock, splotched skin, and speech difficulty. Such persons should obtain the services of a medical doctor immediately.

BEE PRODUCTS

Honey production. Honey is marketed in several different forms; liquid honey, comb honey, and creamed honey. Sometimes the predominant floral type from which the honey was collected is indicated.

Liquid honey. If liquid (strained, extracted) honey is desired, additional supers are added directly above the broodnest. When one is largely filled, it is raised and another is placed underneath. This may continue until several have been filled, each holding from 30 to 50 pounds, or until the nectar flow has ended. After the bees have evaporated the water until the honey is of the desired consistency and sealed in the cells, the combs are removed, the cells uncapped with the uncapping knife, and the honey extracted. The removed honey is immediately heated to about 140° F (60° C), which thins it and destroys yeasts that can cause fermentation. It is then strained of wax particles and pollen grains, cooled rapidly, and packaged for market.

Comb honey. In production of honey in the comb, or comb honey, extreme care is necessary to prevent the bees' swarming. The colony must be strong, and the bees must be crowded into the smallest space they will tolerate without swarming. New frames or sections of a frame with extra-thin foundation wax, added at exactly the right time for the bees to fill without destroying them, are placed directly above the broodnest. The bees must fill and seal the new comb honey to permit removal within a few days or it will be of inferior quality. As rapidly as sections are removed new sections are added, until the nectar flow subsides; then these are removed and the colony given combs to store its honey for the winter.

Creamed honey. Almost all honey will granulate or turn to sugar. Such honey can be liquefied without materially affecting its quality by placing the container in water heated to about 150° F (66° C). Liquid and granulated honey is sometimes blended, homogenized, and held at a cool temperature, which speeds uniformly fine granulation. If properly processed, the granules will be extremely fine; the honey, which has a smooth, creamy appearance, is referred to as creamed honey.

Floral types. Some honeys are sold by floral type; that is, they are given the name of the predominant flowers visited by the bees when they accumulated the honey. The beekeeper has no way to direct his bees to a particular source of food, but through experience he learns which plants are his major sources of honey. Different flowers produce different colours and flavours of honey. It may be heavy-bodied or thin-bodied, dark or light, mild-flavoured or strong-flavoured. Most honey has been blended by the beekeeper to a standard grade that he can supply and market year after year.

World honey production statistics. World production of honey, excluding mainland China, is about 900,000,000 pounds annually. North America produces about 300,000,000 pounds from about 5,000,000 colonies; the U.S.S.R., about 300,000,000 from about 11,000,000 colonies; and the remaining countries of the world, about 300,000,000 pounds from an estimated 15,000,000 colonies. The largest exporting countries are Mexico, Argentina, Australia, and the United States. West Germany is the greatest importer of honey. Although honey production per colony in the U.S. amounts to about 50 pounds, when the colonies are properly manipulated and in good locations they frequently produce several times this amount. An average annual production of several hundred pounds per colony has been reported for a small isolated area of southwestern Australia.

Beeswax. Beeswax is a by-product of beekeeping in most areas. When the beekeeper uncaps or breaks honey combs or has unusable combs, he endeavours to salvage the beeswax. First he recovers as much honey from the combs as possible by drainage or extraction. Then he places the material in water heated to slightly over 145° F. This melts the wax, which rises to the surface. After it cools and hardens, the cake of wax is removed and refined for reuse in comb foundation. Beeswax has many other uses: in quality candles, cosmetics, agriculture, art, and industry. In some areas bees are manipulated primarily for wax production. Wax is a highly stable commodity that can be transported long distances under unfavourable conditions without damage.

Bees reared for sale. Queens are reared for sale to other beekeepers for requeening established colonies or for adding to a two- or three-pound package of 8,000 to 10,000 live bees to form new colonies or replenish weak ones. The queens are produced when the beekeeper cages the reigning queen in a colony, then inserts into the cluster from 30 to 60 queen cell bases into which he has transferred young (one-day old) worker larvae. More than 1,000,000 queens are produced in this way and sold each year in the United States. Queens can be artificially inseminated with sperm from drones of a known source, but most beekeepers let the queens mate naturally.

The live bees are shaken from the combs of the colony through a funnel into screen-wire cages. About 500 tons of live bees are produced for sale annually in the United States, primarily in the Southern states and California. Several tons are shipped annually from the United States to foreign countries, primarily to Canada.

Pollination. The greatest value of bees is in their service as pollinators. Some 90 crops grown in the United States alone are dependent on insect pollination, performed primarily by the honeybee. The average colony of bees is worth from 20 to 40 times as much in the pollination of crops as it is in the production of honey. The value of bees in the pollination of ornamental plants has never been calculated. Bees are also valuable in the pollination of some forest and range plants that produce seeds on which birds and other wildlife feed.

Bees in pollination

When bees are used in the pollination of crops, the beekeeper places his colonies within or adjacent to the field to be pollinated. The majority of the roughly 1,000,000 colonies that are used for pollination are used in alfalfa-seed fields and almond and apple orchards. The colonies are distributed at the rate of two or more per acre in groups every 0.1 mile throughout alfalfa fields. Two colonies per acre are recommended for almond orchards and about one colony per acre in apple orchards.

Some growers prefer to have the colonies placed alongside the orchard; others want them distributed in small groups within the orchard. Bees are also used regularly by growers of many other crops: blueberries, cantaloupes, cherries, clovers, cucumbers, cranberries, cutflower seed, plums and prunes, vetch, and watermelon.

DISEASE AND PEST CONTROL

Honeybees have diseases and enemies: diseases of the brood; diseases that affect only the adult bees; insect enemies of the adults and of the comb; and other enemies, including toads, lizards, birds, mice, skunks, and bears.

Diseases. American foulbrood, caused by the spore-forming bacterium, *Bacillus larvae*, is the most serious brood disease. It occurs throughout the world wherever bees are kept and affects workers, drones, and queens. The spores are highly resistant to heat and chemicals. A comb containing brood severely infected with this disease has a mottled appearance caused by the mixture of healthy capped brood interspersed with diseased or empty cells formerly occupied by diseased brood. The decayed mass has a typical ropiness when dug into, which is one of its identifying characteristics.

American foulbrood can be spread to healthy colonies by transferring equipment or allowing the bees to feed on honey from infected colonies. Sulfathiazole and Terramycin are widely used to control the disease. Many countries and most states in the U.S. require the destruction by fire of diseased colonies and have apiary inspectors to enforce the regulations.

European foulbrood is caused by a nonsporeforming bacterium, *Streptococcus pluton*, but *Bacillus alvei* and *Achromobacter eurydice* are often associated with *Streptococcus pluton*. This disease is similar in appearance to American foulbrood. In some instances it severely affects the colonies, but they recover so that colony destruction is not necessary. Terramycin can control the disease.

Sacbrood is caused by a virus and is superficially similar to the foulbrood diseases. It can appear and disappear spontaneously but is seldom serious. No chemical control is needed, but if the problem persists the beekeeper usually requeens the colony.

Chalk brood is caused by the fungus *Ascosphaera apis*. The larvae victims of this disease have a chalky, white appearance.

Stonebrood, which affects both brood and adults, is also caused by a fungus, *Aspergillus flavus*, which can usually be isolated from bees that have stonebrood.

Nosema disease, caused by the protozoan *Nosema apis*, is the most serious disease of adult bees. It is widespread, causes heavy losses in honey production, and severely weakens colonies. The external symptoms of bees with nosema disease are not apparent. The disease is transmitted from adult to adult by ingestion of the spores that soon germinate in the ventriculus, or main, stomach. An infected ventriculus is normally swollen, soft, and grayish white. A degree of control may be obtained by feeding the colony the drug fumagillin.

Acarine disease is caused by the mite *Acarapis woodi* that gets into the tracheae of the bee through its breathing holes or spiracles in its thorax or midsection. Bees affected by this mite are unable to fly, have disjointed wings and distended abdomens. There is presently no good control for this mite. The only U.S. federal law pertaining to bees was passed to prevent the importation of adult bees carrying this mite into the United States. Two other mites, *Varroa jacobsoni* and *Tropilaelops clareae*, are serious problems of Asian beekeepers, but they do not occur in Europe or North America.

There are other minor diseases of adult bees, but they seldom cause serious problems.

Pests. The greater waxmoth, *Galleria mellonella*, is a lepidopterous insect that, in its larval stage, destroys combs. It does not attack adult bees, but may begin destruction of combs of a weak colony long before the bees are gone. It can also destroy stored combs of honey. When the larvae are ready to pupate they often eat out a place to spin their cocoons in the soft wood of the beehive, damaging frames and other hive parts. The best control for this pest is keeping colonies strong. Stored combs are fumigated, kept in a cold room or stacked in such a way that a strong air draft flows around them.

The larvae of the lesser waxmoth, *Achroia grisella*, cause damage to stored combs similar to that of the greater waxmoth. The Mediterranean flour moth larva, *Anagasta kuehniella*, feeds on pollen in the combs and causes some damage. Control for both of these moths is the same as for the greater waxmoth.

The bee louse, *Braula caeca*, is a tiny, wingless member of the fly family that is occasionally found on bees, but feeds on nectar or honey from the mouthparts of its host. Its larvae burrow in the cappings of honey combs.

Ants sometimes invade hives and disrupt or kill the bees.

Termites can damage hive parts placed on the soil.

Other insects, such as dragonflies (Odonata), robberflies (Diptera), praying mantises (Orthoptera), ambush bugs (Hemiptera), and certain wasps and yellow jackets (Hymenoptera) are natural enemies of the honeybee.

Predators. Mice frequently enter the hive in winter when the bees are clustered, or they get into stored combs and despoil or damage them by chewing the frames and combs to construct their nest.

Skunks devour large numbers of bees at the hive entrance, usually at night. Fences, traps, and poison are used against them.

Bears eat the honeybees and brood in the hive, and usually destroy it and its contents in the process. In bear country, electric fences and traps are used to protect bee colonies.

At times bees become their own deadly enemy. If honey is exposed to them when no flowers are in bloom and the weather is mild, the bees from different colonies will fight over it. Sometimes this fighting, or robbing, becomes intense and spreads from hive to hive in moblike action. If all the bees in one colony are killed, the honey is quickly stolen and carried into other hives. This further intensifies the robbing so that a cluster that was carrying honey into its hive a few minutes earlier is attacked, all of its occupants killed, the honey again stolen and the process repeated. Usually, once robbing becomes intense, only darkness or foul weather will stop it.

BIBLIOGRAPHY. Books that concern the life history of the individual bee and the colony, honey and wax production, diseases of bees, flora that provide nectar and pollen, and economics of beekeeping include: A. BUDEL and E. HERALD, *Biene und Bienenzucht* (1960); J.E. ECKERT and F.R. SHAW, *Beekeeping* (1960); R.A. GROUT (ed.), *The Hive and the Honey Bee* (1963); H. MUXFELDT, *Apicultura Para Todos* (1965); G.S. ORDETZ and D.E. PEREZ, *La Apicultura en Los Trópicos* (1966); C.R. RIBBANDS, *Behaviour and Social Life of Honeybees* (1953); and F.G. SMITH, *Beekeeping in the Tropics* (1960). A.I. ROOT *et al.*, *The ABC and XYZ of Bee Culture*, 33rd ed. (1966), is an alphabetical encyclopaedia dealing with a multitude of subjects on beekeeping.

Periodicals, such as the *American Bee Journal*, *Gleanings in Bee Culture*, *Journal of Apicultural Research*, *Bee World*, *Canadian Bee Journal*, and the *Australian Bee Journal* cover

news, nontechnical reports, and technical papers with a practical slant.

The story of the language of bees is told in K. VON FRISCH, *Bees: Their Vision, Chemical Senses, and Language* (1950); and M. LINDAUER, *Communication Among Social Bees* (1961).

The U.S. DEPARTMENT OF AGRICULTURE, *Handbook No. 335* (1967) on "Beekeeping in the United States," contains sections on special subjects dealing with bees and honey. The Department's Statistical Reporting Service also periodically issues reports on honey production, prices, and numbers of colonies owned by beekeepers.

(S.E.McG.)

Beethoven, Ludwig van

A universal genius widely regarded as the greatest composer who ever lived, Ludwig van Beethoven dominates a period of musical history as no one else before or since. Rooted in the Classical traditions of Haydn and Mozart, his art reaches out to encompass the new spirit of humanism expressed in the works of Goethe and Schiller, his elder contemporaries in the world of literature, and above all in the ideals of the French Revolution, with its passionate concern for the freedom and dignity of the individual. He revealed more vividly than any of his predecessors the power of music to convey a philosophy of life without the aid of a spoken text; and in certain of his compositions is to be found the strongest assertion of the human will in all music, if not in all art. Though not himself a Romantic, he became the fountainhead of much that characterized the work of the Romantics who followed him, especially in his ideal of program or illustrative music, which he defined in connection with his *Sixth* (*Pastoral*) *Symphony* as "more an expression of emotion than painting." In musical form he was a considerable innovator, widening the scope of sonata, symphony, concerto, and quartet; while in the *Ninth Symphony* he combined the worlds of vocal and instrumental music in a manner never before attempted. His personal life was marked by a heroic struggle against encroaching deafness, and some of his most important works were composed during the last ten years of his life when he was quite unable to hear. In an age that saw the decline of court and church patronage, he not only maintained himself from the sale and publication of his works; he was also the first musician to receive a salary with no other duties than to compose how and when he felt inclined.



Beethoven, oil painting by Ferdinand Schimon, 1819. In the Beethoven-Haus, Bonn.
By courtesy of the Beethoven-Haus, Bonn, West Germany

The early years. Baptized on December 17, 1770, in Bonn, northwest Germany, Beethoven was the eldest surviving child of Johann and Maria Magdalena van Beethoven. The family was Flemish in origin and can be traced back to Malines. It was Beethoven's grandfather who had first settled in Bonn when he became a singer in the choir of the Archbishop-Elector of Cologne. He eventually rose to become Kappellmeister—an unusual feat for one who was not a composer. His son Johann was also a singer in the electoral choir; thus, like most 18th-century musicians, Beethoven was born into the profession. Though at first quite prosperous, with the death of his grandfather in 1773 and the decline of his father into alcoholism, the Beethoven family became steadily poorer. By the age of 11 Beethoven had to leave school; at 18 he was the breadwinner of the family.

Having observed in him signs of a talent for the piano, Johann had tried to make of his son a child prodigy like Mozart but without success. It was not until his adolescence that Beethoven, on his own account, began to attract mild attention.

When, in 1780, Joseph II became sole ruler of the Holy Roman Empire, he appointed his brother Maximilian Francis as adjutant and successor-designate to the archbishop-elector of Cologne. Under Maximilian's rule, Bonn was transformed from a minor provincial town into a thriving and cultured capital city. A liberal Roman Catholic, he endowed Bonn with a university; limited the power of his own clergy; and opened the city to the full tide of the German literary renaissance, associated with Lessing, Klopstock, and the young Goethe and Schiller. A sign of the times was the nomination as court organist of Christian Gottlob Neefe, a Protestant from Saxony, who became Beethoven's teacher. Although somewhat limited as a musician, Neefe was nonetheless a man of high ideals and wide culture, a man of letters as well as a composer of songs and light theatrical pieces; and it was to be through Neefe that Beethoven in 1783 would have his first extant composition (*Variations on a March by Dressler*) published at Mannheim. By June 1782 Beethoven had become Neefe's assistant as court organist.

In 1783 he had also been appointed continuo player to the Bonn opera. By 1787 he had made such progress that Maximilian Francis, archbishop-elector since 1784, was persuaded to send him to Vienna to study with Mozart. The visit was cut short when, after only two months, Beethoven received the news of his mother's death. According to tradition, Mozart was highly impressed with Beethoven's powers of improvisation and told some friends that "this young man will make a great name for himself in the world." For the next five years Beethoven remained at Bonn. To his other court duties was added that of playing viola in the theatre orchestra; and, although the archbishop for the time being showed him no further mark of special favour, he was beginning to make valuable acquaintances. Sometime previously he had come to know the widow of the chancellor, Joseph von Breuning, and she engaged him as music teacher to two of her four children. From then on the Breunings' house became for him a second home, far more congenial than his own. Through Mme von Breuning, Beethoven acquired a number of wealthy pupils. His most useful social contact came in 1788 with the arrival in Bonn of Count Ferdinand von Waldstein, a member of the highest Viennese aristocracy and a music lover. Waldstein became a member of the Breuning circle, where he heard Beethoven play and at once became his devoted admirer. At a fancy dress ball given in 1790, the ballet music, according to the *Almanach de Gotha* (a journal chronicling the social activities of the aristocracy), had been composed by Count Waldstein; but it was generally known that Beethoven had written it for him. The same year saw the death of the emperor Joseph II. Through Waldstein again, Beethoven was invited to compose a funeral ode for soloists, chorus, and orchestra; but the scheduled performance was cancelled because the wind players found certain passages too difficult. He then added to it a complementary piece celebrating the accession of Joseph's brother Leopold II; but there is no record that either was ever performed until the end of the 19th century, when the manuscripts were rediscovered in Vienna and pronounced authentic by Johannes Brahms. But in 1790 another great composer had seen and admired them: that year Haydn, passing through Bonn on his way to London, was feted by the elector and his musical establishment; when shown Beethoven's score, he was sufficiently impressed by it to offer to take Beethoven as a pupil when he returned from London. Beethoven accepted Haydn's

Study with Mozart

Study with Haydn

offer and in the autumn of 1792, while the armies of the French Revolution were storming into the Rhineland provinces, Beethoven left Bonn, never to return. The album that he took with him (preserved in the Beethoven-Haus in Bonn) indicates the wide circle of his acquaintances and friends in Bonn. The most prophetic of the entries, written shortly after Mozart's death, runs:

> The spirit of Mozart is mourning and weeping over the death of her beloved. With the inexhaustible Haydn she found repose but no occupation. With the help of unremitting labour you shall receive Mozart's spirit from Haydn's hands. Count Waldstein.

The compositions belonging to the years at Bonn—excluding those probably begun at Bonn but revised and completed in Vienna—are of more interest to the Beethoven student than to the ordinary music lover. They show the influences in which his art was rooted as well as the natural difficulties that he had to overcome and that his early training was inadequate to remedy. Three piano sonatas written in 1783 demonstrate that, musically, Bonn was an outpost of Mannheim, the cradle of the modern orchestra in Germany, and the nursery of a musical style that was to make a vital contribution to the classical symphony. But at the time of Beethoven's childhood, the Mannheim school was already in decline. The once famous orchestra was, in effect, dissolved after the war of 1778 between Austria and Prussia. The Mannheim style had degenerated into mannerism, which took the form of trivial and often inappropriate experimenting with dynamic contrasts as reflected even in Mozart's *Piano Sonata in C major*, K. 309. The preoccupation with extremes of piano (soft) and forte (loud), often in contradiction to the musical phrasing, is found in Beethoven's early sonatas and in much else written by him at that time—which is not surprising since the symphonies of later Mannheim composers formed the staple fare of the Bonn court orchestra. But what was for Mozart only a deviation from his normal style was to remain a fundamental element in that of Beethoven. The sudden pianos, the unexpected outbursts, the wide leaping arpeggio figures (chord notes played rapidly up or down over several octaves) known as "Mannheim rockets"—all these are central to Beethoven's musical personality and were to help him toward the liberation of instrumental music from its dependence on vocal style. Beethoven may, indeed, be described as the last and finest flower on the Mannheim tree.

Early influences

Like other composers of his generation, Beethoven was subject to the influence of popular music and of folk music, influences particularly strong in the Waldstein ballet music of 1790 and in several of his early songs and unison choruses. Heavy Rhineland dance rhythms can be found in many of his mature compositions; but he could assimilate other local idioms as well—Italian, French, Slavic, and even Celtic. Although never a nationalist or folk composer in the 20th-century sense, he often allowed the unusual contours of folk melody to lead him away from traditional harmonic procedure.

French music impinged on him from two main directions: from Mannheim, whose artistic links with Paris had always been strong; and from the Bonn Nationaltheater, which relied mainly for its repertory on comic operas translated from the French. In fashionable Bonn society, sympathy with the French Revolution was very strong, and the flavour of the French Revolutionary march is present in many of Beethoven's symphonic allegros. The jigging rhythms to be found in several of his scherzos are also clearly of French provenance.

Like all pianists of the late 18th century, Beethoven was raised on the sonatas of Carl Philipp Emanuel Bach, the chief exponent of "expressive" music at a time when music was regarded as the art of pleasing sounds. These sonatas, with their quirks of rhythms and harmony and their occasional wordless recitative, were equally familiar to Haydn and Mozart; but in Beethoven they evoked a much readier response, not only for reasons of temperament, but also because of the intellectual climate in which he himself was reared. The favourite literary fare of the Breunings and their friends was associated with the *Sturm und Drang*, a reaction against the rationalism of the early 18th century, an exaltation of feeling and instinct over reason. Its gospel was enshrined in Goethe's early novel, *Die Leiden des jungen Werthers* (*The Sorrows of Werter*), the language of which finds an echo in certain of Beethoven's letters and especially in the "Heiligenstadt Testament" (see below).

In such a movement music took on a new importance as an art of feeling. The sharp conflicts of mood that characterize the sonatas of C.P.E. Bach appear much more powerfully again in Beethoven, to whom "feeling" was as important in practice as it was in theory to his master Neefe, who proclaimed it the only condition of artistic value. All of this does not make Beethoven a Romantic, although Romantics attempted to claim him as one of themselves. His literary world—he read widely and voraciously despite a formal education that in arithmetic had not carried him as far as the multiplication table—was rooted in the German classics, above all Goethe and Schiller. Like them he was to achieve in music a balance of form and emotion that can only be called classical.

The Bonn compositions of most enduring interest date, as might be expected, from the last years: a *Rondino* and an *Octet*, for wind instruments, composed in 1792, probably for the elector's *harmonie* (wind band); a *Trio in G Major for Flute, Bassoon, and Piano* (1791); and the two cantatas. The songs—doubtless written under Neefe's inspiration—show no great feeling for the solo voice. This is strange in one whose father and grandfather had been singers, but it remained a limitation that pursued Beethoven throughout his career. Of particular interest are 24 variations on a theme by Vincenzo Righini, an Italian composer, which, like the *String Trio in E Flat Major*, Opus 3, Beethoven revised and published at a much later date. These variations, representing a compendium of Beethoven's piano technique, for a long time were to serve as the mainstay of his repertory in the salons of Vienna.

Beethoven as pianist

Vienna. Before Beethoven left Bonn he had acquired a very considerable reputation in northwest Germany as a piano virtuoso, with a particular talent for extemporization. Mozart had been one of the finest improvisers of his age; by all accounts Beethoven surpassed him. In the age of sensibility he could move an audience to tears more easily than any other pianist of the time. For this reason especially he was taken up by the Viennese aristocracy almost from the moment he set foot in Vienna. Count Waldstein had, of course, prepared the way with his talk of a successor to Mozart; and it is significant that Beethoven's earliest patrons in Vienna were Baron van Swieten and Prince Karl Lichnowsky, who alone among the aristocracy had remained Mozart's supporters until his death. In the Vienna of the 1790s, music had become more and more the favourite pastime of a cultured aristocracy, for whom politics under the reactionary emperor Francis II were now discreditable and dangerous and who had, moreover, never shown a like appreciation of any of the other fine arts. Many played instruments themselves well enough to be able to take their place beside professionals. Probably at no other time and in no other city was there such a high standard of amateur and semi-professional music making as in the Vienna of Beethoven's day.

As a composer, however, Beethoven still had many technical problems to overcome, and it soon became clear that Haydn was not the best person to help him. Outwardly their relations remained cordial; but Beethoven soon began taking extra lessons in secret. One of his teachers was the organist of St. Stephen's Cathedral, Johann Georg Albrechtsberger, a learned contrapuntist of the old school who equipped him with the comprehensive technique that he needed. He also studied vocal composition with Antonio Salieri, the imperial Kappellmeister. By 1794, when Haydn had left for his second visit to London, there was no longer any question of Beethoven's returning to Bonn, which was then in French hands. The elector himself had left, and consequently Beethoven's subsidy came to an end. But he had no need

to worry for, apart from what he was able to earn by teaching and playing, he received free board and lodgings from Prince Lichnowsky. The year 1795 marked Beethoven's first public appearance as a pianist in Vienna. He played a concerto (No. 2, Opus 19) of his own and one by Mozart and also took part in a benefit concert for Haydn. More important still, his *Three Trios for Piano, Violin and Cello*, Opus 1, were published with a long list of aristocratic subscribers. In the next three years he undertook concert tours in Berlin and Prague and might have travelled more widely still had the international situation permitted. In 1800 he launched a public concert on the grand scale, in which one of his own piano concerti, the *Septet* (Opus 20), and *First Symphony* were given, together with works by Haydn and Mozart. The event did much to spread Beethoven's fame abroad.

The turn of the century concluded what is generally referred to as Beethoven's first period, a period during which his art stayed within the bounds of 18th-century technique and ideas. Most of his published works during that time are for the piano, alone or with other instruments, important exceptions being the *String Trio in E Flat Major*, Opus 3; the *Three String Trios*, Opus 9; the *Six String Quartets*, Opus 18; and the *First Symphony*. Beethoven extended his range slowly and methodically, but he was still a piano composer par excellence.

Approaching deafness. The change in direction occurred with Beethoven's gradual realization that he was becoming deaf. The first symptoms had appeared even before 1800, yet for a few years his life continued unchanged: he still played in the houses of the nobility, in rivalry with other pianists, and performed in public with such visiting virtuosos as violinist George Bridgetower (to whom the *Kreutzer Sonata* was originally dedicated). But by 1802 he could no longer be in doubt that his malady was both permanent and progressive. During a summer spent at the (then) country village of Heiligenstadt he wrote the "Heiligenstadt Testament." Ostensibly intended for his two brothers, but never sent to them, the document begins:

The "Heiligenstadt Testament"

> O ye men who think or say that I am malevolent, stubborn or misanthropic, how greatly do you wrong me. You do not know the cause of my seeming so. From childhood my heart and mind was disposed to the gentle feeling of good will. I was ever eager to accomplish great deeds, but reflect now that for six years I have been in a hopeless case, made worse by ignorant doctors, yearly betrayed in the hope of getting better, finally forced to face the prospect of a permanent malady whose cure will take years or even prove impossible.

He was tempted to take his own life, "But only Art held back; for, ah, it seemed unthinkable for me to leave the world forever before I had produced all that I felt called upon to produce. . . ." There is a Werther-like postscript:

> As the leaves of autumn wither and fall, so has my own life become barren: almost as I came, so I go hence. Even that high courage that inspired me in the fair days of summer has now vanished.

More significant, perhaps, are his words in a letter to his friend Franz Wegeler: "I will seize fate by the throat. . . ." Elsewhere he remarks, "If only I were rid of my affliction I would embrace the whole world." He was to do both, though the condition he hoped for was not fulfilled.

From then on his days as a virtuoso were numbered. Although it was not until about 1819 that his deafness became total, making necessary the use of those conversation books in which friends wrote down their questions while he replied orally, his playing degenerated as he became able to hear less and less. He continued to appear in public from time to time, but most of his energies were absorbed in composing. He would spend the months from May to October in one or another of the little villages near Vienna. Many of his musical ideas came to him on long country walks and were noted in a sketchbook.

Beethoven's sketchbooks

These sketchbooks, many of which have been preserved, reveal much about Beethoven's methods of work. The man who could improvise the most intricate fantasies on the spur of the moment took infinite pains in the shaping of a considered composition. In the sketchbooks such famous melodies as the adagio of the *Emperor Concerto* or the andante of the *Kreutzer Sonata* can be seen emerging from a trivial and characterless beginning into their final form. It seems, too, that Beethoven worked on more than one composition at a time and that he was rarely in a hurry to finish anything that he had on hand. Early sketches for the *Fifth Symphony*, for instance, date originally from 1804, although the finished work did not appear till 1808. Sometimes the sketches are accompanied by verbal comments as a kind of *aide-mémoire*. Sometimes, as in the sketching of the *Third (Eroica) Symphony*, he would leave several bars blank, making it clear that the rhythmic scheme had preceded the melodic in his mind. Many of the sketches consist merely of a melody line and a bass—enough, in fact, to establish a continuity. But in many works, especially the later ones, the sketching process is very elaborate indeed, with revisions and alterations continuing up to the date of publication. If, in general, it is only the primitive sketches and jottings that have survived, this is because Beethoven kept them beside him as potential sources of material for later compositions. The working out of a musical composition in all its detail ceased to interest him once the piece had been completed.

Beethoven and the theatre. The next few years were those of Beethoven's short-lived connection with the theatre. In 1801 he had provided the score for the ballet *Die Geschöpfe des Prometheus* (*The Creatures of Prometheus*). Two years later he was offered a contract for an opera on a classical subject with a libretto by Emanuel Schikaneder, who had achieved fame and wealth as the librettist of Mozart's *Magic Flute* and who was then impresario of the Theater an der Wien. Two or three completed numbers show that Beethoven had already begun work on it before Schikaneder himself was ousted from the management and the contract annulled—somewhat to Beethoven's relief, as he had found Schikaneder's verses "such as could only have proceeded from the mouths of our Viennese applewomen." When the new management re-engaged Beethoven the following year, it was largely on the strength of his now almost forgotten oratorio, *Christus am Ölberg* (*Christ on the Mount of Olives*), which had been given in an all-Beethoven benefit concert, together with the first two symphonies and the *Third Piano Concerto*. The year 1804 was to see the completion of the *Third Symphony*, regarded by most biographers as a landmark in Beethoven's development. It is the answer to the "Heiligenstadt Testament": a symphony on an unprecedented scale and at the same time a prodigious assertion of the human will. The work was to have been dedicated to Napoleon, one of Beethoven's heroes, but Beethoven struck out the dedication on hearing that Napoleon had taken the title of emperor. Outraged in his republican principles, he later substituted the words "for the memory of a great man." From then on the masterworks followed hard on one another's heels: the *Piano Sonata in F Minor*, Opus 57, known as the *Appassionata;* the *Piano Concerto No. 4 in G Major*, Opus 58; the three *Razumovsky Quartets*, Opus 59; the *Fourth Symphony*, Opus 60; the *Violin Concerto*, Opus 61. To this period also belongs his one opera, *Fidelio*, commissioned for the winter season of 1805. The play concerns a wife who disguises herself as a boy in order to rescue her imprisoned husband, and, in setting this to music, Beethoven was influenced by Paer and by Luigi Cherubini, composer of similar "rescue" operas and a musician whom he greatly admired. *Fidelio* enjoyed no great success at first, partly because the presence of French troops, who had occupied Vienna after the Battle of Austerlitz, kept most of the Viennese away. With great difficulty Beethoven was persuaded to make certain changes for a revival in the following spring, with modified libretto. This time the opera survived two performances and would have run longer, but for a quarrel between Beethoven and the management, after which the composer in a fury withdrew his score. It was not until eight years later that *Fidelio*, heavily revised by Beethoven himself and a new librettist, returned to the Vienna stage, to become one of the classics of the Ger-

Mature masterpieces

man theatre. Beethoven later turned over many other operatic projects in his mind but without bringing any to fruition.

The established composer. During all this time, Beethoven, like Mozart, had maintained himself without the benefit of an official position—but with far greater success insofar as he had no family to support. His reputation as a composer was steadily soaring both in Austria and abroad. The critics of the Leipzig *Allgemeine musikalische Zeitung*, the most authoritative music journal in Europe, had long since passed from carping impertinence to unqualified praise, so that, although there were as yet no copyright laws to ensure a system of royalties, Beethoven was able to drive far more favourable bargains with the publishing firms than Haydn and Mozart before him or Schubert after him. Despite the restrictions on Viennese musical life imposed by the war with France, Beethoven had no difficulty in getting his most ambitious works performed, largely because of the generosity of such patrons as Prince Lichnowsky, who at one point made him a regular allowance of 600 florins a year. Others would pay handsomely for a dedication; *e.g.*, Count Oppersdorf, for the *Fourth Symphony*. Also, Beethoven's pupils included the archduke Rudolf, youngest brother of the emperor. Consequently, poverty was never a serious threat. But, doubtless because of increasing deafness combined with a habitual readiness to take offense, Beethoven's relations with the Viennese musicians, on whose cooperation he depended, became steadily worse; and in 1808, at a benefit concert where the *Fifth* and *Sixth* symphonies were first performed, together with the *Choral Fantasia*, Opus 80, there occurred a quarrel so serious that Beethoven thought of leaving Vienna altogether. But the threat of his departure was sufficient to stir his patrons into action. The archduke Rudolf, Prince Lobkowitz, and Prince Kinsky banded together to provide him with an annuity of 4,000 florins, requiring only that he should remain in Vienna and compose. The agreement remained in force until Beethoven's death, though it was to be affected by circumstances, one of which was the devaluation of 1811; although the Archduke increased his contribution accordingly, it was some time before his partners could do the same. Nevertheless, from 1809 onward Beethoven remained adequately provided for, although his habits of life often gave visitors the impression that he was miserably poor. Inevitably, his public appearances became less frequent.

Beethoven and his patrons

Beethoven and women. In this period, too, he considered more seriously than before the idea of marriage. As early as 1801, letters to his friend Wegeler refer to "a dear sweet girl who loves me and whom I love." This is thought to have been the Countess Giulietta Guicciardi, a piano pupil and the cousin of two other pupils, Therese and Josephine, daughters of Count von Brunsvik. It was to the Countess Giulietta that he dedicated the *Piano Sonata in C Sharp Minor*, Opus 27, No. 2, known as the *Moonlight Sonata*. But Giulietta married Count Gallenberg in 1803, and in later years Beethoven seems to have remembered her only with mild contempt. It seems clear, however, that he did propose marriage to her cousin Josephine, whose elderly husband, Count von Deym, died in 1804; and the understanding appears to have continued for about three years, until it was brought to an end partly by Beethoven's own indecisiveness and partly by pressure from Josephine's family. The prospective bride of 1810 is thought to have been Therese Malfatti, daughter of one of Beethoven's doctors, but, like the other marriage projects, this, too, lapsed, and Beethoven remained a bachelor. A curious item, however, was found among his effects, locked away in a drawer, at the time of his death: three letters, written but never sent, to the "Immortal Beloved." The content, which varies from high-flown poetic sentiments to banal complaints about his health and discomfort, makes it clear that this is no literary exercise but was intended for a real person. The month and day of the week are given, but not the year. The periods 1801–02, 1806–07, and 1811–12 have been proposed, but the last is the most probable. The identity of the person addressed is uncertain.

Wider recognition. In 1810 E.T.A. Hoffmann in Berlin produced a famous appreciation of the *Fifth Symphony*, which undoubtedly did much to launch that work on its triumphant career throughout the world and, above all, to interest the Romantics in its composer. The same year, Beethoven made the acquaintance of the writer Bettina Brentano, the sister of the German poet and novelist Clemens Brentano and, later, wife of Achim von Arnim, the two compilers of that famous collection of German folk poetry, *Des Knaben Wunderhorn*. Of the letters that Bettina gave out as having been written to her by Beethoven, only one can be accepted as genuine; at least one of the others, in which the composer is made to philosophize on music in the most uncharacteristically romantic terms, must be dismissed as spurious. Bettina also performed the questionable service of bringing together Beethoven and Goethe at Teplitz in 1812. The admiration had been all on Beethoven's side; to Goethe, Beethoven was little more than a famous name. The meeting was not a success. "Goethe is too fond of the atmosphere of the courts," Beethoven wrote to Breitkopf and Härtel, the music publishers, "more so than is becoming to a poet . . . ." Goethe considered Beethoven to be "an utterly untamed personality, who is not altogether in the wrong in holding the world to be detestable, but surely does not make it any the more enjoyable either for himself or for others by his attitude." He showed a certain interest in the incidental music written in 1810 for *Egmont* "out of pure love for the subject."

Meeting with Goethe

The chief compositions of 1811–12 were the *Seventh* and *Eighth* symphonies, the first of which had its premiere in 1813. Another novelty at the same concert was the so-called *Battle Symphony*, written to celebrate Arthur Wellesley's (later duke of Wellington) decisive victory over Joseph Bonaparte at Vitoria. Composed originally for a mechanical musical instrument, the Panharmonicon, invented by J.N. Maelzel, Beethoven later scored the work for orchestra. He frankly admitted it was program music of the worst kind, so different from the ideals of "mehr Ausdruck der Empfindung als Malerei" ("more as an expression of feeling than painting") expressed in his own *Pastoral Symphony;* but in view of its success he was ready enough to score it for orchestra and even to send a copy of the score to the English prince regent, who, much to Beethoven's annoyance, made no acknowledgment. The concert, profitable as it was for the composer, led to a bitter quarrel with Maelzel, from which Beethoven emerges with little credit.

Despite the difficulties over the annuity caused by the devaluation of 1811, the years 1813–14 were profitable ones for Beethoven. The first performance of the *Seventh Symphony* was a huge success, and the audience insisted on the allegretto being repeated. When the Congress of Vienna assembled in 1814, Beethoven's music was universally known, and he himself was courted by the crowned heads of Europe. *Fidelio* was revived with tumultuous success, and Beethoven celebrated the fall of France with a grand patriotic cantata, *Der glorreiche Augenblick* (*The Glorious Moment*). In 1814, after years of war, Vienna was to enjoy a brief hour of glory before the Austrian economy collapsed and the city sank into a state of dowdy provincialism that lasted for nearly 40 years.

The last years. With the start of Metternich's long reign and the so-called Biedermeier period, marked by simplicity and homeliness in art and design, Beethoven's creative life entered its third and final phase. Because of his deafness he became more of a recluse than ever. His rate of composition, too, began to decrease. The works written between 1815 and 1827 comprise a mere fraction of his output after 1792; but they have a density of musical thought far surpassing anything that he had composed before. Though he now went less into society, he concerned himself more and more with business matters, not always with happy results.

It was about this time that he was brought in touch with the Philharmonic Society of London. Earlier, in 1803, he had been approached by the Edinburgh publisher George Thomson with a proposal that he should write

sonatas based on Scottish folk tunes. Although nothing came of this, Thomson somewhat later succeeded in contracting him to arrange national folk melodies for voice, violin, cello, and piano, each with an introduction and coda. These remained an easy and profitable source of income to Beethoven for many years. It was in 1815, however, when Beethoven's pupil Ferdinand Ries settled in London and became one of the founder-members of the Philharmonic Society, that English music lovers began to take an active interest in the promotion of Beethoven's works. Another society member, Charles Neate, visited Beethoven in Vienna and later brought about the commission of three new overtures to be performed by the society. The overtures *König Stephan*, *Namensfeier*, and *Die Ruinen von Athen* were, however, late in arriving, and the discovery that they were not new, after all, caused considerable bad feeling; for a time, relations became strained on both sides. Ries did much to effect a reconciliation, but a visit to London, planned as early as 1813, never materialized, though Beethoven continued to hope that it would. The Philharmonic Society never ceased to interest itself in Beethoven's music and it undoubtedly played an important part in the genesis of the *Ninth Symphony*, which in a sense it commissioned. The society's archives contain an autograph of the first movement with a dedication by the composer. The first performance of the work was not, however, given in London but in Vienna, and the printed edition was dedicated to Frederick William III, king of Prussia. Beethoven, on his deathbed, received from the society a gift of £100, which moved him profoundly.

Commissions from England

In 1815 all prospects of foreign travel were cut short for Beethoven by the death of his brother Caspar Anton Carl, who left a widow, Johanna, and a son, Karl, aged nine. The will, which appointed Beethoven and the widow as joint guardians, was contested by Beethoven on the grounds of the widow's immorality; and after three years of litigation he won his case. But, for all the affection that he lavished on young Karl, Beethoven was far from being an ideal guardian. Quarrels between uncle and nephew were frequent and bitter and came to a head in 1826 when, just before sitting for his university examination, Karl attempted suicide. He recovered in a hospital, and Beethoven, on the advice of friends, agreed reluctantly that the boy should be launched on an army career. Once away from his uncle, Karl seems to have led a successful, law-abiding life. But the events of 1826 upset Beethoven profoundly and almost certainly hastened his death.

Late masterworks

The important compositions of the final period begin with the *Two Sonatas for Piano and Cello*, Opus 102, the *Piano Sonata in A Major*, Opus 101, and the *Piano Sonata in B Flat Major*, Opus 106, the latter known as the *Hammerklavier*. Beethoven then reverted to sketches he had begun for the *Ninth Symphony*. This was broken off when the news came that the archduke Rudolf was to be appointed archbishop of Olmütz, and Beethoven decided to write a large-scale solemn mass for the installation ceremony. Work on this progressed slowly, and, like the early cantata for Joseph II, it was not completed in time for the intended occasion. Not until 1823, three years after the enthronement, was Beethoven able to send to the new archbishop the completed manuscript of the *Missa Solemnis*.

In the meantime, Beethoven had written the three final piano sonatas (1820–22) and had worked desultorily on the symphonic sketches. The mass was followed by his last important piano work (completed 1823), variations on a theme that the publisher and composer Anton Diabelli had sent to a number of composers, Beethoven among them. Most of them, including Schubert and the archduke Rudolf himself, obliged; Beethoven at first declined, then changed his mind and decided to write a complete set of 33 variations himself.

The *Ninth Symphony* had begun to take shape; by the following year (1824) it was finished and was performed, together with movements from the *Missa Solemnis* and the overture from Opus 124, with great success at the Kärnthnerthor Theatre. The composer, following with the score, remained unaware of the applause until one of the soloists made him turn to face the audience. The *Ninth Symphony* was Beethoven's last work for large-scale forces. His final commission came in 1823 from Prince Nikolas Galitzin, who offered 50 ducats each for three string quartets. Beethoven accepted with alacrity, though only in 1825 was the first of the three, the *String Quartet in E Flat Major*, Opus 127, completed. Not two but four more followed, including an extra movement, which was substituted for the original fugal finale (*Grosse Fuge*) of the *String Quartet in B Flat Major*, Opus 130. The last quartet was finished in 1826, about the time of Karl's attempted suicide. Beethoven spent that summer on the estate belonging to his surviving brother, Nikolaus Johann. On his return to Vienna he contracted pneumonia, from which he never fully recovered. He remained bedridden and died from cirrhosis of the liver in Vienna on March 26, 1827. The funeral three days later was attended by 20,000 people. Pallbearers included the famous pianist Hummel; Schubert was among the torchbearers; Franz Grillparzer, Austria's greatest living dramatist, wrote the funeral oration.

Beethoven's achievement. Beethoven's greatest achievement was to raise instrumental music, hitherto considered inferior to vocal, to the highest plane of art. During the 18th century, music, being nonimitative, was ranked below literature and painting. Its highest manifestations were held to be those in which it had the aid of a text—that is, cantata, opera, and oratorio—the sonata and the suite being relegated to a lower sphere. A number of factors combined to bring about a gradual change of outlook: the instrumental prowess of the Mannheim Orchestra, which made possible the development of the symphony; the reaction on the part of writers against pure rationalism in favour of feeling; and the works of Haydn and Mozart. But, above all, it was the example of Beethoven that made possible the late-Romantic dictum of the English essayist and critic Walter Pater: "All arts aspire to the condition of music."

After Beethoven it was no longer possible to speak of music merely as "the art of pleasing sounds." His instrumental works combine a forceful intensity of feeling with a hitherto unimagined perfection of design. He carried to a further point of development than his predecessors all the inherited forms of music (with the exception of opera and song), but particularly the symphony and the quartet. In this he was the heir of Haydn rather than of Mozart, whose most striking achievements lie more in opera and concerto.

Division of works into three periods

It was his biographer Wilhelm von Lenz who first divided Beethoven's output into three periods, omitting the years of his apprenticeship in Bonn. The first period begins with the completion of the *Three Trios for Piano, Violin and Cello*, Opus 1, in 1794, and ends about 1800, the year of the first public performance of the *First Symphony* and the *Septet*. The second period extends from 1801 to 1814, from the *Piano Sonata in C Sharp Minor* (*Moonlight*) to the *Piano Sonata in E Minor*, Opus 90. The last period runs from 1814 to 1827, the year of his death. Though the division is a useful one, it cannot be applied rigidly. A composition begun in one period may often have been completed in another, hence the existence of such transitional works as the *Third Piano Concerto* and the *Second Symphony*, which belong partly to the first period and partly to the second. Again, the tide of Beethoven's maturity advanced at a rate that varied according to his familiarity with the medium in which he happened to be writing. The piano was his home ground; therefore, it is in the piano sonatas that the middle-period characteristics first make their appearance, even before 1800. The mass, on the other hand, was unfamiliar territory, so that the *Mass in C Major*, written during the same period as the *Fourth Piano Concerto* and the Razumovsky string quartets, sounds in many ways like an early work.

First period. Apart from the *First Symphony* and first two piano concerti, the works of the first period consist entirely of chamber music, most of it based on Beethoven's own instrument, the piano. All show a preoc-

cupation with craftsmanship in the 18th-century manner. The material, for the most part, has a family likeness to that of Haydn and Mozart but, in keeping with the contemporary style, is slightly coarser and more blunt. Beethoven's treatment of the forms in current use is usually expansive. The expositions are long and polythematic; the developments are relatively short. Slow movements are long and lyrical with copious decoration. The third movement, though sometimes called a scherzo, remains true to its minuet origins, though its surface is often disturbed by unminuet-like accents. Finales are at once high-spirited and elegant. Two characteristics, however, mark Beethoven out strongly from other composers of the time: one is an individual use of contrasted dynamics and especially the device of crescendo leading to a sudden piano; the other, most noticeable in the piano sonatas, is the gradual infiltration of techniques derived from improvisation—unexpected accents, rhythmic ambiguities designed to keep the audience guessing, and especially the use of apparently trivial, almost senseless material from which to generate a cogent musical argument.

Second period. The second period may be said to begin in the piano music with two sonatas "quasi una fantasia," Opus 27, of 1801, but in the symphony and concerto it is not fully apparent before the *Eroica* (1804) and the *Fourth Piano Concerto* (1806). Here the use of improvisatory material is more and more marked; but, whereas in the earlier period Beethoven was more concerned to show how it could fit naturally into a traditional 18th-century framework, here he explores in greater detail the logical implication of every departure from the norm. His harmony remains basically simple—much simpler, for instance, than much of Mozart's; what is new is the way it is used in relation to the basic pulse. From this Beethoven creates in his main themes an infinite variety of stress and accent, out of which the form of each movement is generated. The result is that, of all composers, Beethoven is the least inclined to repeat himself; all his works, but especially those of the middle and late period, inhabit their own individual formal world. Other characteristics of the middle period include shorter expositions and longer developments and codas; slow movements, too, become much shorter, sometimes vanishing altogether. The third movement is now always a scherzo, not a minuet, with frequent use of unexpected accents and syncopation. Finales tend to take on much more weight than before and in certain cases become the principal movement. Decoration begins to disappear as each note becomes more functional, melodically and harmonically. Another feature of these works is their immediacy. Here Beethoven's power is most evident; and the majority of the repertory works belong to this period.

Third period. The third period is marked by a growing concentration of musical thought combined with an increasingly wider range of harmony and texture. Beethoven's enthusiasm for Handel began to bear fruit in a much more thoroughgoing use of counterpoint. But he never lost touch with the simplicity of his earliest manner, so that the range of expression and mood in these last works is something that has never been surpassed. A form to which he gave more and more attention at this time was that of the variation. As an improviser he had always found it congenial, and, though some of the sets he had published in earlier years are merely decorative, he had created such outstanding examples of the genre as the finale of the *Eroica* and the *Prometheus* variations, both on the same theme. It is this type of variation that Beethoven began to pursue in his final period. A unique feature of the sets that occur in his last string quartets and sonatas is the sense of cumulative growth, not merely from variation to variation but within each variation itself. In the quartets, everything in the composer's musical equipment is deployed—fugue; variation; dance; sonata movement; march; even modal and pentatonic, or five-tone, melody.

Use of variations

Structural innovations. Beethoven remains the supreme exponent of what may be called the architectonic use of tonality. In his greatest sonata movements, such as the first allegro of the *Eroica*, the listener's subconscious mind remains oriented to E-flat major even in the most distant keys, so that when, long before the recapitulation, the music touches on the dominant (B flat), this is immediately recognizable as being the dominant. Of his innovations in the symphony and quartet, the most notable is the replacement of the minuet by the more dynamic scherzo; he enriched both the orchestra and the quartet with a new range of sonority and variety of texture.

The same is true of the concerto, in which, strictly speaking, he introduced no formal innovations, the entry of solo instrument before an orchestral ritornello in the *Fourth* and *Fifth* piano concerti having been already anticipated by Mozart. Although, in the finale of the *Ninth Symphony* and the *Missa Solemnis*, Beethoven shows himself a master of choral effects, the solo human voice gave him difficulty to the end. His many songs form, perhaps, the least important part of his output. His one opera, *Fidelio*, owes its pre-eminence to the excellence of the music, rather than to any real understanding of the operatic medium. But even this lack of vocal sense could be made to bear fruit, in that it set his mind free in other directions. A composer such as Mozart or Haydn, whose conception of melody remained rooted in what could be sung, could never have written anything like the opening of the *Eroica*, in which the melody takes shape from three instrumental strands each giving way to the other. Wagner was not far wrong when he hailed Beethoven as the discoverer of instrumental melody.

Beethoven holds an important place in the history of the piano. In his day, the piano sonata was the most intimate form of chamber music that existed—far more so than the string quartet, which was often performed in public. For Beethoven, the piano sonata was the vehicle for his most bold and inward thoughts. He did not anticipate the technical devices of such later composers as Chopin and Liszt, which were designed to counteract the percussiveness of the piano, partly because he himself had a pianistic ability that could make the most simply laid-out melody sing; partly, too, because the piano itself was still in a fairly early stage of development; and partly because he himself valued its percussive quality and could turn it to good account. Piano tone, caused by a hammer's striking a string, cannot move forward, as can the sustained, bowed tone of the violin, although careful phrasing on the player's part can make it seem to do so. Beethoven, however, is almost alone in writing melodies that accept this limitation, melodies of utter stillness in which each chord is like a stone dropped into a calm pool. Beethoven was less successful in combining the piano with one other instrument, and his duo sonatas remain on a slightly lower level. But it is above all in the piano sonata that the most striking use of improvisatory techniques as an element of construction is found. Among later composers it was chiefly Liszt who extended Beethoven's principle of transferring structural weight from the first movement to the finale, making it the basis of his symphonic poems as well as of his two concertos. The two works of Beethoven that undoubtedly had most influence over succeeding generations were the *Fifth* and *Ninth* symphonies, with their progression from storm and stress to triumph. Brahms' *Symphony No. 1 in C Minor*, Tchaikovsky's *Symphony No. 5 in E Minor*, César Franck's *Symphony in D Minor*, and Mahler's *Symphony No. 2 in C Minor* are all examples of Beethoven's spiritual progeny, though few will grant that they equal, let alone surpass, their models.

MAJOR WORKS

Orchestral music

SYMPHONIES: *No. 1 in C Major*, op. 21 (1800); *No. 2 in D Major*, op. 36 (1802); *No. 3 in E Flat Major*, op. 55 (*Eroica;* 1804); *No. 4 in B Flat Major*, op. 60 (1806); *No. 5 in C Minor*, op. 67 (1808); *No. 6 in F Major*, op. 68 (*Pastoral;* 1808); *No. 7 in A Major*, op. 92 (1812); *No. 8 in F Major*, op. 93 (1812); *No. 9 in D Minor*, op. 125 (*Choral;* 1824). *Wellington's Victory*, op. 91 (also known as *The Battle of Vitoria* and the *Battle Symphony;* 1813).

CONCERTOS (PIANO): *"No. 1" in C Major*, op. 15 (1798),

"No. 2" in B Flat Major, op. 19 (in fact composed first 1795, revised 1798); *No. 3 in C Minor*, op. 37 (1800); *No. 4 in G Major*, op. 58 (1806); *No. 5 in E Flat Major*, op. 73 (*Emperor;* 1809). (VIOLIN): *Violin Concerto in D Major*, op. 61 (1806); *Triple concerto in C Major*, op. 56 (violin, cello, piano; 1804).

OTHER ORCHESTRAL COMPOSITIONS: 2 romances for violin and orchestra; various overtures, including *Coriolan*, op. 62 (1807); *Leonore No. 1*, op. 138, *2*, op. 72A, and *3*, op. 72B.

Chamber music

STRING QUARTETS: No. 1–6, op. 18 (1798–1800); No. 1–3, op. 59 (*Razumovsky;* 1806); op. 74 (*Harp* 1809); op. 95 (1810); and the late quartets (1824–26); op. 127, 130, 131, 132, 133 (*Grosse Fuge*, originally the finale to 130) and op. 135.

OTHER CHAMBER WORKS: *Octet*, op. 103 (winds; 1792); *Septet* (strings and wind; 1800); *Sextet for Horns and String Quartet*, op. 81B (1795); *Quintet for Piano and Winds*, op. 16 (1796); *String Quintet in C Major*, op. 29 (1801); 7 piano trios; 5 string trios; 10 sonatas for violin and piano, including *Sonata in A Major* (*Kreutzer;* 1803); 5 sonatas for cello and piano; sonata for horn and piano.

Piano music

32 sonatas, including *Sonata in C Sharp Minor*, op. 27, no. 2 (*Moonlight;* 1801); and *Sonata in F Minor*, op. 57 (*Appassionata;* 1804); 3 sets of Bagatelles; 20 sets of variations; 4 rondos.

Vocal music

Missa Solemnis (mass in D major; 1823); *Mass in C Major*, op. 86 (1807); *Christus am Ölberg* (oratorio 1803); various smaller works for chorus and orchestra including *Choral Fantasia*, op. 80 for piano, chorus, and orchestra (1808); songs, including the cycle *An die ferne Geliebte*, op. 98 (1816), and Goethe and Gellert settings; Scottish, Irish, and Welsh folk-song settings.

Theatre music

One opera, *Fidelio* (1805; revised versions, 1806, 1814—the final version is the one usually heard today); one ballet, *Die Geschöpfe des Prometheus* (1801); incidental music to four plays; *Egmont*, op. 84 (1810), *Die Ruinen von Athen*, op. 113 (1811), *König Stephan*, op. 117 (1811), *Die Weihe des Hauses*, op. 124 (1822).

BIBLIOGRAPHY

Works: The standard complete edition, *Ludwig van Beethovens Werke*, ed. by GUIDO ADLER *et al.* (1864–90), is being replaced by an entirely new edition, *Beethoven: Werke* (1961–), under the general editorship of JOSEPH SCHMIDT-GORG. The important works for the most part have opus numbers allocated by Beethoven himself. Lists of those of Beethoven's works without opus numbers (*Werke ohne Opuszahl*) may be found in the catalogs of Kinsky and Hess: G. KINSKY, *Das Werk Beethovens: Thematisch-bibliographisches Verzeichnis seiner sämtlichen vollendeten Kompositionen*, ed. by H. HALM (1955); W. HESS, *Verzeichnis der nicht in der Gesamtausgabe veröffentlichten Werke Ludwig van Beethovens* (1957). Neither of the above is complete in its information. An exhaustive catalog of completed works, projected works, and sketches, many of them transcribed with commentary, may be found in *Beethoven: A Symposium*, ed. by ATES ORGA (1972).

Letters and conversation books: *The Letters of Beethoven*, 3 vol., collected, trans., and ed. by EMILY ANDERSON, (1961), is the standard edition of Beethoven's letters; a selection from these has been issued, with additional notes by ALAN TYSON (1967). *New Beethoven Letters*, trans. and annotated by D.W. MACARDLE and L. MISCH (1957); G. SCHUNEMANN, *Ludwig van Beethovens Konversationshefte*, 3 vol. (1941–43); *Ludwig van Beethovens Konversationshefte*, ed. by KARL-HEINZ KOHLER and GRITA HERRE, with PETER POTSCHNER, 5 vol. (1968–). The Conversation Books represent Beethoven's only way of keeping contact with his friends after the onset of complete deafness. For obvious reasons, they represent mostly the nonBeethoven side of the conversation.

Life: A.W. THAYER, *The Life of Ludwig van Beethoven*, 2 vol., ed. by E. FORBES (1964, 1967), is the standard biography, representing the third completed edition, revised and brought up-to-date, of Thayer's original work. It is however, considerably condensed; and students are recommended to consult in addition the earlier American edition, *The Life of Ludwig van Beethoven*, ed. and trans. by H. KREHBIEL, 3 vol. (1926); and the German *Ludwig van Beethovens Leben*, ed. by H. DEITERS and H. RIEMANN (1907–17). A. SCHINDLER, *Biographie von Ludwig van Beethoven* (1860; trans. into English by C.S. JOLLY as *Beethoven As I Knew Him*, ed. by D.W. MACARDLE, 1966), has the value of a detailed life written by someone who knew the composer intimately, and its errors and distortions of fact are corrected in some excellent annotations. *Beethoven: Impressions of Contemporaries*, ed. by O.G. SONNEK (1926), is a useful anthology of opinions and accounts given by those with whom Beethoven came into contact. G.R. MAREK, *Beethoven: Biography of a Genius* (1969), gives a balanced, readable, and above all, up-to-date account of the composer's life without going into as much detail as Thayer. E. and R. STERBER, *Beethoven and His Nephew: A Psychological Study of Their Relationship*, trans. by W.R. TRASK (1954), is a controversial exercise in posthumous psychoanalysis.

Studies of the music: D.F. TOVEY, *Beethoven*, ed. by H. FOSS (1944), is a series of penetrating essays on various aspects of Beethoven's work; it was intended to form the basis of a full study which the author never lived to complete. Equally valuable is his *A Companion to Beethoven's Piano Sonatas* (1931), which provides a close structural analysis of all 32 works. E. BLOM, *Beethoven's Piano Sonatas Discussed* (1938), is compiled from a set of notes written for the famous recordings made by Artur Schnabel. J. KERMAN, *The Beethoven Quartets* (1967), offers a comprehensive and stimulating treatment of the music. P. RADCLIFFE, *Beethoven's String Quartets* (1965), is a shorter study but very concentrated. In the symphonic field G. GROVE, *Beethoven and His Nine Symphonies* (1896), is an established classic.

(J.M.Bu.)

Begoniales

The order Begoniales (begonia order) consists of two families of flowering plants: Datiscaceae, distributed mostly in the northern tropics and subtropics, has three genera and three or four species; Begoniaceae, mostly of the moist tropics, has three to five genera, depending on the authority consulted, and about 1,000 species. There is little resemblance between plants of the two families, members of the Datiscaceae being mostly trees (except for *Datisca*, an herb) and those of the Begoniaceae mostly succulent herbs or subshrubs.

General features. *Size range and diversity of structure.* Begonias vary in size from a few centimetres in height (such as *Begonia steyermarkii* of Venezuela, an herb) to the five- to six-metre (16- to 20-foot) height of somewhat woody shrubs or what are virtually small trees (such as *B. pentaphylla* and *B. digitata* of Brazil).

The growth habits of the plants of the order varies as widely as their size. Included in the family Begoniaceae are plants with stems that are upright, creeping, or developed into underground rhizomes (horizontal rootlike stems); stemless plants with basal tubers; and plants that climb by means of true roots or prop roots or by hooks developed from bent and elongated axillary buds—buds located in the upper angle between stems and leafstalks.

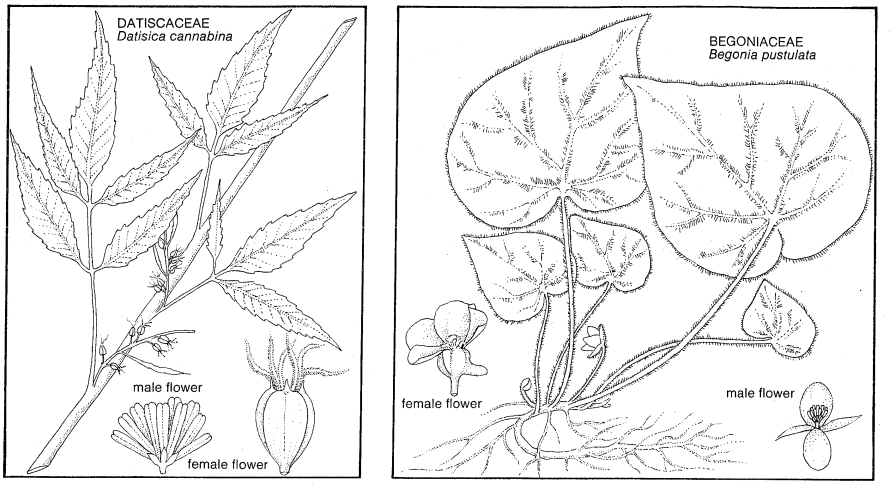
Characteristic begonia leaf shapes

The leaves of begonias vary from simple to compound (*i.e.*, single bladed or with several leaflets, respectively) and essentially are smooth margined to much lobed. They generally have one lobe noticeably larger than the other. Characteristically, all leaves on one side of a branch have the larger lobe on the same side of the leaf, and those on the other side have the larger lobe on the opposite side; the leaves, in effect, form mirror images of each other.

In the family Datiscaceae, *Tetrameles* and *Octomeles* are large trees, 30 metres (100 feet) or more in height. *Datisca*, the other genus in the family, is a perennial herb resembling *Cannabis* (hemp), for which one of its two species, *D. cannabina*, is named.

Distribution and abundance. The family Begoniaceae is distributed generally throughout the moist tropics—being noticeably absent only from Australia and from Fiji to the Galápagos Islands. Some species of Begoniaceae are abundant in moist, cool tropical areas, such as occur in Central America to altitudes of 2,000 metres (6,500 feet) or more and in the wet eastern Himalayas to 3,600 metres (11,800 feet). Others, with thick hairy leaves or stout tuberous bases, occupy niches in drier areas that are sandy or rocky. The plants may be found singly, in small colonies, or, occasionally, covering valleys or new road cuts.

The only species of *Tetrameles*, *T. nudiflora*, is found in India, Sri Lanka (formerly Ceylon), Java, and elsewhere in deciduous forests of medium rainfall, such as those in



Representative plants from the two families of the begonia order.
Drawing by M. Moran

the Western Ghāts; the region from Sikkim to Tenasserim, Burma; the Andaman Islands; and the middle and eastern portions of Java in teak forests.

Two species of *Octomeles* have been described, but they are considered synonymous by most botanists. Specimens of *Octomeles* are among the tallest trees in the moist evergreen ebony forests of the Malay Archipelago and Papuasia (Territory of Papua, New Guinea and the Solomon Islands), where they are locally abundant.

The two species of *Datisca* are perennial herbs that are widely separated geographically. *D. cannabina* is a native of the Old World, ranging from western Asia to the south side of the Himalayas, where it is found in the dry western portion, as well as in the temperate damp parts of the middle Himalayas. *D. glomerata* (durango root) is known from California to northwestern Mexico but is not common.

Economic importance. The family Datiscaceae is of minor local importance and is of scientific interest chiefly because of its greatly reduced floral structure and its interrupted geographical distribution. The wood of *Tetrameles* and *Octomeles* is used in making packing cases especially for tea. The root of *Datisca cannabina* contains a compound, datiscin, used in the Orient as a yellow dye for silk.

The family Begoniaceae is of considerable horticultural interest, many species and hybrids being prized as house or garden plants. The leaves of some African and Asian *Begonia* species are eaten; in South America, some species are eaten to prevent scurvy. Rhizomes of certain species in South America are used for their medicinal qualities as astringents, as treatment for fever, as antisyphilitics, and as laxatives. The family Begoniaceae is also of evolutionary interest. The family seems to be the end point of a line of development that has involved the structure of its various organs. Patterns of evolutionary progression, apparent from the diversity of the species, may be traced in the floral parts and the fruits.

Botanical and horticultural features

The ornamental value of the plants is also based on diversity in growth habit; leaves, including colour patterns; inflorescence (flower-cluster) type; flower size and colour; and length of blooming period. Classifications of begonias for horticultural purposes are quite different from those for botanical studies. In fact, the species are valued for their potential in producing hybrids. Three groups of species are of interest to horticulturists. One is cultivated for its leaves (including especially *Begonia rex*). Another, valued for its flowers, is subdivided into a group of small-flowered plants centred on *B. cucullata* (more commonly called *B. semperflorens*) and a group of large-flowered plants including the tuberous begonias such as *B. rosaeflora*, *B. veitchii*, and *B. boliviensis*. A third group has both ornamental flowers and decorative leaves and includes *B. scharffiana* and *B. metallica* from which was produced the hybrid *B.* × *credneri*, now more widespread than either of its parents.

Modern growers use four group classifications for begonias: (1) fibrous-rooted, including the wax (*B. semperflorens*), the cane stemmed, and the hirsute, or hairy; (2) the rhizomatous; (3) the rex; and (4) the tuberous rooted. Such horticultural classifications are not used by botanists because they often separate related taxa (genera, species, groups of species, etc.).

Natural history. Because of the comparative rarity of the family Datiscaceae and because two genera of the family are tall trees with flowers at great distances from the ground, not many observations have been made on pollination mechanisms. Both *Tetrameles* and *Octomeles* are dioecious (*i.e.*, male and female flowers on separate plants) and have unisexual flowers. An abundance of flowers suggests that pollination by wind is most likely in *Tetrameles*. Wind pollination is indicated in the other two genera because of their smooth ovoid pollen grains, but the possibility of insect pollination cannot be excluded. *Datisca glomerata* is also wind-pollinated.

Flowers of all the Begoniaceae are unisexual. There are almost no records of abnormalities in nature involving abortive members of one sex appearing in flowers of the other. The plants are usually monoecious—with flowers of both sexes developing on the same plant—but normally the two sexes come to maturity at different times. In some cases the plants are dioecious. In monoecious plants, flowers of both sexes may, depending on the species, either develop on the same inflorescence (flower cluster) or the inflorescences may bear flowers of only one sex.

Pollination

Even though flowers of both sexes are often borne on the same plant, self-pollination does not usually occur because of the difference in the flowering time of the sexes. Some inflorescences are protandrous (*i.e.*, the male flower parts mature before the female of the same inflorescence), and the pollen is completely shed by the time the female flowers are receptive. Other inflorescences, however, which have the staminate (male) flowers on upper branches and carpellate (female) ones on lateral branches, are protogynous—the female parts mature first—and the fruits are completely mature before the pollen is shed.

Few records have been made of pollinating agents though three different genera of bee visitors (*Apis*, *Bombus*, and *Podalirius*) have been reported. In cultivation it is thought that self-pollination may occur occasionally by a trickling down of pollen to the stigmas (the sticky pollen-receiving surfaces of the female flower parts).

Mature seeds, mostly less than one millimetre long, are borne in large numbers, usually in capsules, which become dry and split open. The small, characteristically patterned, light-weight seeds seem not to be transported over very great distances, though there are some records of bird dispersal.

Many members of the Begoniaceae contain free organic acids (including oxalic and malic) in the cell sap. It has

been suggested that the presence of these acids may be an active deterrent to insect attack. Other cell contents of interest include crystals of calcium oxalate.

Evolution and paleontology. No unanimity has been reached on the actual relationships of the family Datiscaceae, so that evolutionary theorizing has been impractical and unsatisfactory. Affinity to the Begoniaceae has been rather generally accepted, with the seeds perhaps the most uniform character in the two families.

The interrupted distribution of *Datisca* has been suggested as an evidence of its former occurrence in Europe, but a fossil record is lacking.

Based on the present distribution of the family Begoniaceae, the present-day representatives of the probable prototype of the genus *Begonia* are the American and Asiatic members of *Begonia* (section *Begoniastrum*). From this group have originated the other numerous, highly specialized, endemic (*i.e.*, restricted to one region) members. The two chief centres of diversity are from Brazil and the Andes mountains to Mexico; and in the eastern Himalayas, the mountains of Indochina, the Malay Archipelago, and especially the Philippines and New Guinea. Smaller numbers of species have reached other areas. Representatives of the family are absent in Europe and are relatively scarce in Africa, but distribution patterns are thought to indicate an earlier worldwide dispersal. Fossil evidence, however, is nonexistent, and proof, therefore, is lacking.

Classification. *Distinguishing features.* At present the genus *Begonia* consists of more than 60 sections to which its hundreds of species have been assigned. The primary divisions are geographical, and the sections, though distinguished by morphological characters, are rather strictly confined by geographical limits. Since many species are narrow endemics, the system of subdividing the genus into sections is more satisfactory in *Begonia* than it might be in many other genera. Even here, however, many true relationships are obscured, and some species have been described numerous times from different localities.

Basis of begonia classification system

Twelve sections are described from Africa, 20 from Asia, 27 from America; one with three subsections is American–Asiatic. A few additional sections of uncertain relationship from each area have also been designated.

Although the plants vary in size, habit, and floral structure, they are usually recognizable as begonias because of their generally succulent (fleshy, thick) leaves and stems and the slightly to markedly oblique leaf outline characteristic of most species.

Botanically the family Begoniaceae offers many taxonomic problems of interest. No complete modern monograph is yet available, the most detailed studies having been prepared for particular geographical areas, usually as parts of more inclusive floras. Phytogeographical (plant distribution) studies would produce many interesting and worthwhile results because of the high percentage of endemism in the family. If combined with monographic work, such research would probably result in a reduction of the number of species now recognized.

Annotated classification. The order Begoniales is presented here in a form generally accepted by botanists; *i.e.*, including two families, Begoniaceae and Datiscaceae, the latter containing the genera *Datisca*, *Tetrameles*, and *Octomeles*. In 1965, however, the last two genera were accorded family status (Tetramelaceae) and thus were separated from the Datiscaceae. This newer arrangement seems to be a more satisfactory classification than any previously proposed.

ORDER BEGONIALES

Mostly succulent herbs, but perennial shrublike herbs and large woody trees also occur. Leaves simple, pinnate (with a midrib) or palmate (with radiating leaflets or veins); alternate. Stipules (small leaflike appendages at the base of leafstalks) present, early deciduous, or absent. Flowers usually unisexual (but some bisexual flowers in *Datisca*). Sepals (green or coloured petallike outer whorl of flower parts) often petallike, 2 to 9. Petals 2 to 8 or more, or none in a few. Stamens (male reproductive structures) 4 to 25 or more. Carpels (simple ovaries or segments of a compound ovary) 3, rarely 5 or 2. Styles (the narrow upper part of the pistil) 2 to 5, ovules numerous. Seeds small, numerous, and with a dotted or net-patterned (reticulate) seed coat. Fruit a capsule, rarely a berry. Two families, very dissimilar in character and number of species, about 1,000 in one and only 4 in the other.

Family Begoniaceae

Plants mostly monoecious (male and female flowers on one plant) succulent herbs or subshrubs. Leaves oblique—asymmetrical, one lobe larger than the other—with stipules. Flowers unisexual, the staminate (male) bilaterally symmetrical with 2 to 4 (sometimes to 8) free to somewhat fused tepals (sepals and petals considered together because they look alike). Stamens mostly numerous and free or the filaments somewhat fused. Carpellate (female) flowers with 2 to 5 (sometimes to 8 to 10) free to somewhat fused tepals. Ovary inferior (enclosed within the basal portions of tepals), except half-inferior in *Hillebrandia*, 3-locular (chambered) with ovule placentation (attachment) usually axile (located along central axis of ovary) or rarely parietal—attached to the ovary wall. Ovules very numerous, anatropous (inverted and straight). Styles 2 or 3, sometimes to 4 or 6, more or less fused at base, two-branched, the branches sometimes twisted and with stigmatic papillae (nipple-like bumps) in spiral bands or, rarely, in rings or covering the variously shaped surfaces. Fruit mostly capsular, 3-locular and 3-winged or 3-horned, opening along the inner edge of the wings; less often, wingless, fusiform (spindle-shaped), or berrylike and not splitting open or, in *Hillebrandia*, opening between the styles. Three genera: *Hillebrandia* (1 species, Hawaii), *Begonia* (including the sometimes-recognized genera *Begoniella* and *Semibegoniella*, about 1,000 species throughout the tropics except Australia and Fiji to the Galápagos Islands), and *Symbegonia* (12 species, New Guinea).

Family Datiscaceae

Flowers radially symmetrical, dioecious or sometimes polygamous, sepals and petals present or petals absent. Staminate (male) flowers with 4 to 9 free and very uneven sepals or these fused to a long, broad tube and only the tips free; petals lacking or equal in number to the sepals. Stamens either equal in number to the calyx tips and opposite them or indefinite in number and irregular in position. Anthers 2-locular, opening by lateral slits, attached near the base. Carpellate (female) flowers with sepals usually fused, attached to the ovary and exceeding it, terminating in 3 to 8 small, upright tips; petals and stamen rudiments lacking. Styles 2-parted, clavate (club-shaped) or bearing a capitate (headlike) stigma. Ovary 1-locular with 3 to 8 parietal placentae. Fruit a capsule opening either between the styles or laterally; seeds very numerous and very small. Three genera and 4 species: *Tetrameles* (1 species, a tree) is distributed in India, Ceylon, and Java; *Octomeles* (1 species, a tree) is found in Malaysia; and *Datisca* (2 species, tall perennial shrubs) with 1 species distributed from the Mediterranean to the Himalayas and in Central Asia, the other in southwestern United States and northwestern Mexico.

Critical appraisal. Problems remain concerning the relationship of the families Datiscaceae and Begoniaceae and the true affinity of each in the general classification system. Casual observation of the plants in the two families does not give much insight into their possible relationship because of the enormous differences in the habit and aspect of their members.

A need exists for more detailed observations, utilizing modern techniques, on a family-wide basis of trichomes (plant hairs), pollen, chromosomes, and seeds. A broad-scale anatomical survey might serve to clarify and perhaps to modify relationships between species groups.

BIBLIOGRAPHY. B.R. BUXTON, *Begonias and How to Grow Them* (1946), drawings and photographs; A. DE CANDOLLE, "Mémoire sur la famille des Bégoniacées, *Annls. Sci. Nat. (a) Botanique*, 4th Series, 11:93–149 (1859), a classic monographic treatment, in French, still of much interest; C. CHEVALIER, *Les Bégonias* (1938), a classic work in French on cultivated begonias; E. GILG, "Datiscaceae," in A. ENGLER and K. PRANTL (eds.), *Die Natürlichen Pflanzenfamilien*, 2nd ed., 21:543–547 (1925), the only general and inclusive treatment of this small family, in German; E. IRMSCHER, "Begoniaceae," *ibid.*, pp. 548–588, the most comprehensive review available—treats the family as a whole, the genera and sections, and names some species in each section but does not give species descriptions; and "Fam. Begoniáceae, Schiefblattgewächse," in *Parey's Blumengärtnerei*, 2nd ed., 2:67–98 (1960), a German work of extreme usefulness for horticulturists and amateurs—deals with species in cultivation as well as hybrids and cultivars on the basis of botanical characters and includes a key to sections and a synoptic key to species considered; HELEN K. KRAUSS, *Begonias for American Homes and Gardens* (1947), line drawings and photographs; R.A.H. LEGRO and J.

DOORENBOS, "Chromosome Numbers in Begonia," *Neth. J. Agric. Sci.*, 17:189–202 (1969), the most modern and best-documented chromosome study for *Begonia;* R. WILCZEK, "Begoniaceae," in *Flore du Congo, du Rwanda et du Burundi* (1969), a modern floristic treatment in French for tropical Africa.

(B.G.S./L.B.Sm.)

Behaviour, Animal

Animal behaviour (ethology) includes any activity of an intact organism. A living animal behaves constantly in order to survive, and all animals must solve the same basic problems. They must, for instance, periodically replace their energy source (consume food), avoid dehydration (drink), avoid becoming another animal's energy source (avoid being eaten), maintain their body surfaces (clean and groom), and reproduce.

THE VARIETY OF ANIMAL BEHAVIOUR

Any animal may be regarded as an agglomeration of interacting and interdependent structures and behaviours that are responses to environmental conditions. The behavioral features of modern animals are the accumulated results of millennia of selective pressures acting on small variations inherent in individuals. This selection is relentless because environments are constantly changing.

An understanding of comparative behaviour is helpful in understanding human behaviour, just as an understanding of comparative anatomy is helpful in understanding human anatomy. The reason for this is that both behaviour and anatomy have a genetic basis. All vertebrates, for example, share certain anatomical features that distinguish them as vertebrates; and smaller groupings such as fish, amphibians, reptiles, birds, and mammals may be distinguished from one another on the same basis. This type of distinction holds true between species and even between individuals. Consequently, an understanding of the anatomy or behaviour of any species is helpful in understanding other species, including man himself. In general, the closer the relationship between any two species, the more similar are the structures and behaviours of the two species. The converse is also true. Exceptions, however, do exist. Humans and chimpanzees, for instance, are closely related genetically, but, because of historic differences in environment, the behaviour of man is, in many ways, more like that of wolves, which experience many problems similar to those of ancient man. Such convergences and divergences are commonplace in biological evolution. Convergence occurs when unrelated animals independently evolve similar responses to similar environmental conditions—*e.g.*, the similar body shapes of porpoises and sharks; the similar social behaviour of wolves and men (see SOCIAL BEHAVIOUR, ANIMAL). Divergence occurs when closely related species are adapted to different conditions, with a resultant difference in behaviour and structure. This is the usual type of response; sometimes, however, divergence is extreme enough to obscure a close relationship. The males of many species of closely related hummingbirds, birds of paradise, pheasants, and ducks, for example, are superficially so different from one another that many of these species were formerly assigned to different genera.

Behaviour as a taxonomic aid

The study of behaviour has provided valuable information about relationships among animals. The Greek philosopher Aristotle was one of the first to use behaviour as a taxonomic aid, but only in recent times have behavioral features been important in animal taxonomy. Aristotle regarded pigeons and doves as closely related to the sand grouse, basing his view partly on their similar way of drinking. Pigeons, doves, and sand grouse, unlike most other birds, keep the bill in the water and drink with a pumping action.

Behaviour may be quite simple, as in taxis (movement toward or away from a stimulus) and kinesis (undirected response proportional to the intensity of a stimulus). These two types of behaviour—most often descriptive of invertebrates—may be further subdivided. Orthokinesis, for example, is a response that involves change in the speed of movement of the body as a whole. Klinokinesis involves changes in the rate of turning from side to side. Klinotaxis is a type of orientation to stimuli in which, in alternate body movements, external stimuli are received with equal intensity. In tropotaxis the orientation of the animal is similar to that in klinotaxis, but it depends upon stimuli acting simultaneously upon two receptors or upon two parts of one receptor. These are stimulated unequally if the animal is not oriented directly toward or away from the source of stimulation. In telotaxis the animal orients to one or the other of conflicting stimuli affecting the same sensory mechanism. In menotaxis, or light compass response, animals (*e.g.*, honeybees, ants) do not orient either directly away from or toward a source of stimulation but assume a constant angle to the direction of the stimulus. Complex behaviours such as nest building, courting, and fighting do not lend themselves to the simple labelling of taxes or kineses and are classified according to other systems, which are dealt with below (see REPRODUCTIVE BEHAVIOUR; AGGRESSIVE BEHAVIOUR).

CLASSIFICATION OF BEHAVIOUR

Types. Behaviour may be described according to the nature of the muscular contractions involved or in terms of the consequences of the behaviour. Because muscle contractions in specific behaviours are often complex, a kind of shorthand is commonly used to describe them. Terms such as tail flick, head bob, and threat posture are of this nature. In describing behaviour in terms of its consequences, terms such as avoidance, courtship, nest building, and burrowing are often employed. Sometimes a behaviour must be described both in terms of the type of muscle contraction and in terms of the consequences of the behaviour.

Three types of behavioral classification are used most commonly: (1) on the basis of immediate causation, (2) on the similarity of evolutionary history, and (3) on the similarity of function.

Immediate causation. Classification based on immediate causation requires that the causal factors first be identified. All behaviours triggered by similar causal factors are then grouped together. Whether or not behaviours share the same causal factor can be determined by either of two methods. One method is to administer the causal factor and see if all of the behaviours are elicited and affected similarly. The other method, used when the causal factor is not known, is to examine the chronological correlations between the activities in question. Two activities that consistently occur together are likely to be causally related. This method is often used in studies of agonistic (attack–escape) courtship, feeding, egg laying, and similar complex behaviours.

Evolutionary history. Types of behavioral evolution are often classified according to similarities in biological evolution. Similarities between patterns of muscular contractions are compared among species believed to be related. The degree of difference between presumably homologous behaviours provides a criterion for measuring the degree of evolutionary affinity and for determining the direction of evolutionary changes. Data useful in classification may sometimes also be obtained by observing the appearance or disappearance of behaviours during ontogeny, or individual development.

Function. Classification based on similarity of function depends on the identification of behaviours with similar evolutionarily adaptive values. Such behaviours may or may not be homologous. The flying behaviour of a bird, for example, is not homologous with that of a butterfly, but both have similar functional and adaptive significance and may be classified together on this basis. A functional classification often closely agrees with a causal classification; this is because evolutionarily functional and causal mechanisms are commonly associated. The more distantly animals are related, the less likely it is that their functional and causal classifications will overlap; the converse also tends to be true: the more closely animals are related, the more likely it is that the two classifications will overlap.

In the case of behaviour in juveniles and adults of the same species, causal and functional categories may some-

times not overlap; an example is penis erection in juvenile mammals, in which reproduction is not a causal factor, compared with that in adults, in which it is.

Innate versus learned behaviour

The influence of genetics and experience. Behaviour is sometimes classified according to the nature of the changes occurring during evolution or ontogeny. From these bases groupings such as learned, innate (*i.e.*, unlearned, or instinctive), and ritualized behaviour are derived. This classification scheme is useful in studies such as those dealing with the acquisition of communication behaviour (see INSTINCT; LEARNING, ANIMAL).

Every aspect of an animal—behavioral as well as structural—is influenced to some degree by heredity as well as by experience. Although a muscle can become larger, harder, and more vascularized (*i.e.*, supplied with blood vessels) or smaller, softer, and less vascularized depending upon the nature of the experience to which it is subjected, the possibility of such modification is not infinite because of genetic limitations. Some behaviours are inherited in the same sense that an organ is inherited, and they exhibit little or no capacity to be modified by experience. Other types of behaviour achieve definitive form and employment only with experience.

Various intermediate conditions exist. If the environment is predictable relative to the appropriateness of a given behaviour in an animal, it is biologically advantageous for this behaviour to be innate, because presumably there is no need for a variety of response. Variability would, in fact, be disadvantageous, because an inappropriate response could threaten the animal's chance of survival. On the other hand, if the environment is unpredictable in certain respects, it is biologically advantageous to have some degree of inherent variability based upon response to experience. In this case, however, it is likely that the animal would sometimes respond with the wrong behaviour, resulting in a lowered survival value. In behaviours that may be modified by experience, the nature and extent of the modification are not limitless. The nature, intensity, timing, and duration of experience that may influence a given behaviour vary with the type of behaviour involved. The kind of response to these aspects of experience is determined genetically and has evolved to the extent that the maximum possible survival value is achieved.

Some behaviours are innate in the same sense that an organ of the body is innate. Their functions are relatively fixed and predictable. Other behaviours have an inherent resiliency and may respond radically to changes in the environment. Genetically determined limits, however, are always imposed upon even the most variable behaviours. An animal would otherwise be capable of limitless modifications of its behaviour, within its structural limitations, and be able to associate arbitrarily any environmental stimulus with any behaviour.

A general evolutionary trend exists for more and more behaviours to be modified by environmental stimuli as the phylogenetic scale ascends (*i.e.*, as the animal becomes more complex and "advanced"). This tendency has reached its extreme expression in vertebrates, particularly mammals, and especially in humans.

COMPONENTS OF BEHAVIOUR

Fixed action patterns. A behaviour that is independent of environmental stimuli for its form is known as a fixed action pattern (FAP). An environmental stimulus may, however, be responsible for the elicitation and proper orientation of the FAP and may have an influence on the completeness of the response. Common examples of FAP's include displays (visible and audible signals), nest-building movements, various food-gathering and food-preparation movements, thermoregulatory movements, and attack and escape movements.

Species specificity of FAP's

Because FAP's are often specific for particular species, they are frequently useful in taxonomic and evolutionary studies. The homologous relationship among FAP's of related species is often easily determined, and their qualitative and quantitative differences can be evaluated.

FAP's are ordinarily quite constant in form, but this stereotypy is not a defining characteristic. Some vary considerably in the degree of completeness, even though the proportions of the components of the response may remain quite constant, relative to one another. Some FAP's share a single environmental stimulus. Sight of a rival male, for example, may elicit flight, attack, or any of a variety of agonistic (attack–escape) displays.

It is sometimes difficult to distinguish an FAP from its orienting movements. In such cases the eliciting stimulus can be manipulated, and the subsequent effects on the behaviour can be observed. The graylag goose (*Anser anser*), for example, retrieves eggs displaced from the nest by means of a highly stereotyped behaviour. While sitting on the nest, the goose extends its head beyond the egg so that the undersurface of the bill is against it. The head is then pulled toward the body until the egg is again safely in the nest. While the head is being pulled toward the body, it makes balancing adjustments that compensate for the tendency of the egg to roll to either side. If the egg is removed during retrieval, the head continues its movement toward the nest, but compensatory movements to counter the erratic roll cease. The FAP is elicited by the egg-out-of-nest stimulus and, once triggered, goes to completion whether or not the egg is still balanced against the underside of the bill. The balancing movements, however, depend upon continuing stimulation by the egg itself and are not part of the head-withdrawing FAP. The balancing movements also cease if a smoothly rolling cylinder is substituted for the egg.

This egg-retrieving FAP illustrates other features of FAP's in general. Every species can perform a finite number of FAP's and have a limited capacity, or none at all, for developing new ones. The limitations may be determined by physical structure: a woodchuck cannot fly, nor can a pelican burrow. Yet neural organization may provide restrictions as ubiquitous and rigid as anatomical ones. The egg-retrieving graylag goose is physically capable of retrieving an egg with its broadly webbed feet or with a wing used as a kind of broom. This never happens, however, because the goose's neural organization permits egg retrieving only in the manner described. Alternative behaviours can sometimes be employed. Parrots of the genus *Agapornis*, for instance, always scratch the head by bringing a foot forward over the wing on the same side; however, when a foot is brought forward for cleaning the bill, it always passes below the wing on the same side. The bird has the physical structure as well as the neural organization necessary for both movements, but each method is specific for its particular stimulus.

Key stimuli. *Examples.* An animal reacts to relatively few of the stimuli present in its environment. This is a basic characteristic of behaviour. It is seldom known whether an animal actually perceives but does not react to the various available stimuli or whether it fails to perceive stimuli of no significance to it—at least of no significance within a given context. Both situations are probably true, at least for some species, under given circumstances. The human eye, for instance, focusses on the retina a detailed picture of all that the eye is directed toward. The human observer, however, is unaware of much within his field of vision. The same is true of other sensory modalities. At some point, insignificant stimulation is filtered out. An animal could not function if it had to respond to all of the stimuli its sensory organs are capable of receiving at any moment. Each species has, therefore, evolved responses only to those stimuli significant to itself and at such times that responses to such stimuli are relevant. This simpler world that actually falls within the animal's perception at any particular moment is termed its *Umwelt*.

Umwelt

The tick's response to its relatively simple *Umwelt* graphically illustrates how an animal selectively responds to only those stimuli pertinent to its immediate requirements. The mature female tick responds to light falling on her photosensitive body surface by moving to the tip of a twig or some suitable substitute, where she waits for a mammal to approach. The tick is capable of waiting for years until that occurs. When a mammal passes close by, she releases her grip on the twig and, if successful, falls onto the body of the mammal. The key stimulus

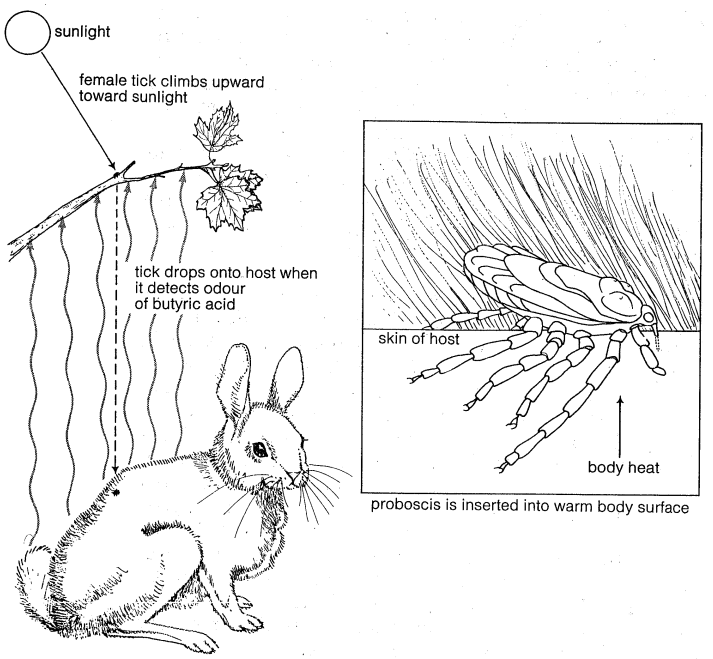


Figure 1: Response of the tick to environmental stimuli. Before obtaining a blood meal from the proper host, it must successively react to the stimuli of light, butyric acid, and body warmth.

causing her to release her grip is the scent of butyric acid, which naturally issues from the body of any mammal. When the female tick is on the host's body, she reacts to its body heat by inserting her proboscis (feeding organ) into the mammal's skin and sucking herself full of blood. This behavioral sequence depends on three stimuli: light, the odour of butyric acid, and warmth. The tick will relinquish its perch for anything with the scent of butyric acid, sink its proboscis into any warm surface—even a balloon filled with warm water—and fill herself with whatever fluid is within. The taste of blood and mammalian characteristics are not meaningful to the tick.

Aggressive behaviour in the male stickleback fish (*Gasterosteus aculeatus*) occurs when the fish sees the red belly of other males. Crude dummies are attacked as long as they have red bellies; realistic models or actual fish without red bellies are not attacked.

Various species of predacious animals utilize lures that simulate the natural foods of prey species. The alligator snapping turtle (*Macrochelys temmincki*) has two slender, reddish structures projecting from the tip of its tongue that move like small worms. The turtle, with its mouth open, rests on a river bottom. A fish attracted by the false worm is snapped up by the turtle. Several species of deep-sea anglerfish have a long slender projection from the forepart of the dorsal fin. This "fishing rod" terminates in a wriggling wormlike structure that is dangled close to the anglerfish's mouth. When another fish investigates the lure, it is easily snapped up by the anglerfish. Different species of anglerfish possess different lures that are specific for different prey species.

The male swordtail characin fish attracts female mates by means of an organ resembling a daphnia, a favourite prey. The gill covers of the male are modified into a long, slender projection terminating in this "daphnia," which is moved within view of the female. When she has been lured close enough, the male copulates with her.

Female fireflies of the genus *Photurus* lure males of other genera by imitating their flashing codes, then seize and eat them. A certain petal in various species of fly orchids is modified to resemble the females of various wasp species, and the odour of these female wasps is also imitated. When a male wasp is attracted to a model female of the same species and attempts to copulate with it, the pollen of the flower adheres to the male and is, in turn, carried to the next flower that attracts the wasp.

Males of certain African cichlid fish (genus *Haplochromis*) deceive females by means of a body pattern that resembles cichlid eggs. The females, which are mouth breeders, take into their mouths the eggs they have laid before the male fertilizes them. Visual models of these eggs are part of the pattern of the ventral fins of the male. After the female has placed the eggs in her mouth, the male spreads his ventral fin before her. The male emits sperm as the female snaps at the false eggs, thus permitting the real eggs in her mouth to be fertilized.

Fertilization behaviour in cichlid fish

Another mouth-breeder fish, *Tilapia macrochir*, ensures fertilization by means of a different deception stimulus. The male produces sperm in filament-like packets (spermatophores), which are shed into the water. Later, they are picked up by the female. She may not always find them, however, and the male has evolved long filament-like spermatophore models that project from his genital region. For some reason, these models exert a stronger stimulus than the real spermatophores. The female takes them into her mouth and at the same time receives real spermatophores that have been placed among the models.

Visual behaviour in the frog

The eyes of the meadow frog, *Rana pipiens*, have five types of cells, each of which responds to a different kind of stimulus. One type responds briefly when a light is turned on or off; it also responds to the passing of the leading and trailing edges of an image moving across the retina. It does not respond to a stationary image, however. A second type of cell responds to the passing of straight or curved edges. A third type does not respond to changes in light intensity but to the passage of the image of a small object, in contrast to its background, across the retina. The fourth type of cell measures a decrease in illumination, and the fifth type measures light intensity.

The frog is capable, therefore, of receiving information about the size, shape, movement, and illumination of objects and is particularly well equipped to perceive small moving objects—its normal food. Much of the stimulus filtering in the frog's vision takes place peripherally at the retinal level. Studies of other subjects such as cats, rabbits, and moths reveal that the processing and integrating of sensory data occurs at various levels in the central nervous system (brain and spinal cord).

Releasing mechanisms. Hundreds of types of responses to a few key stimuli have been identified in various animals. These responses are mediated in the central nervous system by a so-called releasing mechanism (RM) and are responsible for triggering the specific motor response appropriate to the stimulus. If the releasing mechanism is innate, it is termed an IRM. If it is acquired through individual experience, it is termed an ARM. Some innate releasing mechanisms may be modified as a result of individual experience; such a mechanism is termed an IRME.

A given stimulus does not always prompt the same response in the same individual. Such differences are due to internal factors. Some changes are seasonal and are brought about by internal conditions that may, for example, be related to reproduction and associated aggressive behaviour. In the spring, a male wood thrush (*Hylocichla mustelina*) responds to a female with courtship behaviour and to another male with aggression. In the winter the same thrush fails to respond in this way. Relatively short-term changes in responsiveness also occur. An animal that has just fed, for example, shows no further interest for a time in food. In such an instance a short-term internal change has taken place.

The strength of a stimulus necessary to evoke a response of standard intensity also varies with time. The longer an animal is deprived of food, for example, the more unappetizing the food can be and still be accepted. The converse is also true. The more recently an animal has eaten, the more appetizing the food must be to be accepted.

The intensity of a noxious stimulus or the degree of difficulty of an obstacle that an animal will attempt to surmount varies with time. The longer an animal has been without food (short of physical debilitation), the more difficult an obstacle can be and still be surmounted by the hungry animal. Again, the converse is true.

Drive and motivation. Internal changes that initiate behavioral changes are commonly termed drive or motivation. These terms are usually applied to short-term reversible changes in response to a constant stimulus. They are not applied to long-term changes that are the result of learning or to short-term changes that result from muscular fatigue, sensory accommodation, and sensory adaptation.

"Searching" behaviour

When internal conditions are intense enough to initiate a particular drive, an animal commonly behaves as if it were searching for the correct environmental stimulus necessary to trigger the appropriate response. Such searching, or appetitive, behaviour is often highly variable. The true nature of a particular appetitive behaviour can, as a rule, be ascertained only as the act progresses. The drive that motivates a robin to search a lawn, for example, cannot be determined until the search nears culmination. If the robin seizes an earthworm, it is evident that hunger was the activating drive. If it picks up mud or grass, the appetitive phases of nest-material gathering are apparent. Appetitive behaviour tends to become less and less variable as the appropriate terminating situation becomes more and more likely to occur. A hunting falcon flies a search pattern until a potential prey is sighted. The bird dives upon it, after which the exact flight path is determined by whatever evasive action the prey may take. If the falcon is successful, the prey is struck and killed, carried to a perch, and systematically pulled apart and eaten. When the hunger is satisfied, the drive state no longer exists, and some other activity follows. The closer the appetitive sequence is to termination, the more stereotyped the falcon's behaviour becomes. The consummatory act is the most stereotyped behaviour of all.

Courtship behaviour culminating in successful copulation provides many examples of characteristic appetitive–consummatory chains of behaviour. In such cases copulation itself is the terminal appetitive behaviour and is highly stereotyped. Ejaculation, or discharge of sperm by the male, in certain species is completely stereotyped and is followed by temporary cessation of the sex drive.

Learned behaviour is important in appetitive sequences in many animals. A jay, for instance, quickly discovers the best places to find particular foods, and it learns to begin its search in such places rather than search at random until a suitable forage area is found.

"Supernormal" stimuli. A major subject of investigation in animal behaviour has been the determination of key stimuli necessary to trigger particular behaviours. In order to determine those characteristics of an egg by which an incubating bird identifies it as such, a selection of model eggs can be presented to the bird, each model differing in one respect from a normal egg. The reactions to variations in colour, pattern, shape, size, and texture vary according to the species. Generally, differences in shape do not seem important to an incubating bird; but models with more rounded contours appear to be favoured. Differences in colour, pattern, and size are important, but differences in texture do not seem to be.

A model in which the key stimulus has been exaggerated to an extreme degree may be chosen in preference to a normal model. The oystercatcher, for example, prefers a "supernormal" egg, several times the usual size. It also prefers an abnormal clutch of five eggs to the normal clutch of three.

Feeding behaviour of herring gull chicks

Chicks of the herring gull (*Larus argentatus*) are stimulated by a red spot on the lower bill of the adult. When the chick pecks at this spot, the adult regurgitates food for it. By presenting the chick with various models of beaks, it has been found that differences in the colour of the head and bill are not significant; but the red spot, narrowness of the bill, movement, low position of the head, and a downward pointing of the bill are all important in eliciting a response. A thin rod with a red band near the tip moved in a low position provides a supernormal set of stimuli, which elicit a positive response.

When more than one stimulus elicits a given response, the stimuli may supplement one another. If two or more stimuli are required to evoke a response, a weakness of one stimulus may be counterbalanced by the strength of another. Such a compensatory effect is termed the law of heterogeneous summation. In higher animals, learned behaviour may play an important role in this phenomenon. A response may originally depend upon one or a few key stimuli, but, as a result of experience, an animal may come to regard previously irrelevant conditions as among those stimuli necessary for the response, resulting in a kind of gestalt response, in which several stimuli are perceived as an integrated whole. Animals lower on the phylogenetic scale may be more apt to respond to heterogeneous summation, and those higher on the scale may be more apt to respond to a gestalt, however acquired.

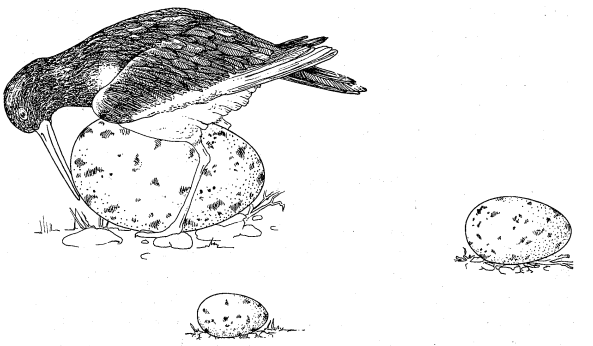


Figure 2: Oystercatcher (*Haematopus ostralegus*) reacting to giant egg in preference to normal egg (foreground) and herring gull's egg (right).
From N. Tinbergen, *The Study of Instinct*, copyright © 1962; the Clarendon Press, Oxford

The supernormal stimuli discussed above have been observed experimentally. The tendency of an animal to respond more vigorously to enhanced stimulation may be of evolutionary significance, because animals with a genetically based tendency to prefer more advantageous variants of a stimulus would tend to have a higher probability of survival. In social situations in which the relevant stimuli are part of an animal's body, such as the herring gull's bill, those structures offering a favourable departure from the normal would enhance the probability of survival in the offspring.

Movement. *Form.* The form of movement of a response is not determined by either the eliciting stimulus or by the properties of the musculature involved. It is possible, however, that the form of movement may be determined by either of two other factors. First, the sequence in which muscles contract to produce a movement may be determined solely by the properties of central nervous system mechanisms responsible for the movement. In this case the movement, once initiated, is independent of further sensory stimulation.

Second, the muscle contractions near the completion of a movement may also be influenced through a feedback mechanism provided by the earlier contractions. The form of the movement would thus be continuously monitored by sensory control. Fixed action patterns, although independent of further stimulation once elicited, may depend upon such internal feedback mechanisms.

Distinction between external and internal movement. The way by which animals are able to distinguish between movements of the environment and movements of the sense organs is not fully understood. When the human eye views a moving object, the object appears to move. If the eye is moved while looking at a stationary object, the object appears stationary, though in both cases the image moves across the retina. But if a stationary object is viewed while the eye is displaced slightly by pushing with a fingertip, the object appears to move.

Equilibration in fishes

If a resting fish is tilted to one side, the statolith (organ of equilibrium) on that side shifts position, thereby activating sensory endings; these set in motion muscular action that restores the fish to an upright position. A fish often deliberately tilts sideways, however, and, in this case, the automatic reflex does not pull the fish upright. It was formerly believed that the righting reflex is blocked during spontaneous movement. Studies have shown, however, that such blocking does not occur. If a

fish is whirled in a centrifuge, the deliberate tilting movements made by the free-swimming fish are of lower intensity. The tilting movements become less because the statoliths are made heavier. The righting reflex is not blocked during deliberate tilts, therefore, but is dependent upon the feedback caused by the tilts.

During spontaneous movement, the stimuli that otherwise release postural reflexes are not inactivated but must be neutralized in another way. The principle of reafference has been hypothesized to account for this. By this hypothesis the functional system is visualized as a feedback loop, whereby afferent nerves carry impulses toward the central nervous system and efferent ones carry impulses away from the central nervous system to the motor areas. Afferences can be divided into receptor excitations caused by internal changes in the musculature (reafference) and those produced passively by external stimulation (exafference). Reafference and exafference are integrated in some manner in the higher centres of the nervous system. The reafference hypothesis postulates that, with each voluntary movement, a copy of the efferent motor impulse is stored in a subordinate nervous centre. The efferent impulse continues to the effector, and movement results. The sense organs then report the result of this movement as a reafference—a feedback of information. This reafference is matched with the efferent copy and is cancelled. If the total afference is too much or too little, as the result of external stimulation, there remains a plus or minus value as compared to the efferent copy stored in the subordinate centre. The discrepancy is reported to the higher centre, which then strengthens or weakens the initial command. This hypothesis provides an explanation for the righting response of fish and for similar phenomena.

Behavioral chains. When an animal responds to a stimulus, the releasing situation is often altered because the animal has progressed to a new position, in which other stimuli are effective. For example, when a female three-spined stickleback enters the territory of a male, he performs a zigzag dance. She responds to this with a signal of her own, which, in turn, releases a behaviour in the male that causes her to follow him. The male shows her the nest opening, which she enters. The male trembles with his snout against her tail, stimulating her to spawn, after which she leaves the nest. The male then fertilizes the eggs. Each of these behaviours depends upon the appropriate stimulus. If one is omitted, the chain terminates without a productive conclusion.

Scent and vision in wasps

The behaviour of the bee-hunting wasp *Philanthus triangulum* illustrates another such chain. This wasp flies from flower to flower as it searches for bees. It responds initially to the visual stimulus afforded by any moving bee-size object; during this time it is indifferent to bee scent. After the wasp perceives the visual stimulus, it hovers about 10 to 15 centimetres (four to six inches) downwind of the bee, and then is sensitive to bee scent; if the scent is appropriate, the wasp attacks the bee and seizes it. Following seizure, bee scent is no longer an effective stimulus. Moving models of the appropriate size attract the wasp, but they will not be seized unless they have bee scent. The behaviour depends upon a succession of stimuli that must occur in a precise sequence.

Many such behavioral chains are known for vertebrates as well as invertebrates. They are not always precisely ordered, and variations may occur. Many such behavioral chains do not exhibit a succession of stimuli made available as a result of the responses. Single causal factors may stimulate several responses. Activities occurring near the end of a chain may require a higher intensity of stimulus than earlier ones. If a causal factor proves inadequate at some point, the behaviour reverts to an earlier stage in the sequence. This is probably true in the courtship of the three-spined stickleback, in which all the activities of both sexes depend upon common endocrine factors (*i.e.*, hormones), on short-term states of heightened responsiveness, and on the nature of the external stimuli.

Simultaneous stimulation. Causal factors for many types of behaviour are usually present at any given moment. A male stickleback may be simultaneously confronted with several stimuli: a ripe female, food, a rival male, and a predator. Animals usually respond to one stimulus at a time and according to a certain priority of sequence. In the male stickleback, escape from a predator almost always takes priority over concurrent stimuli.

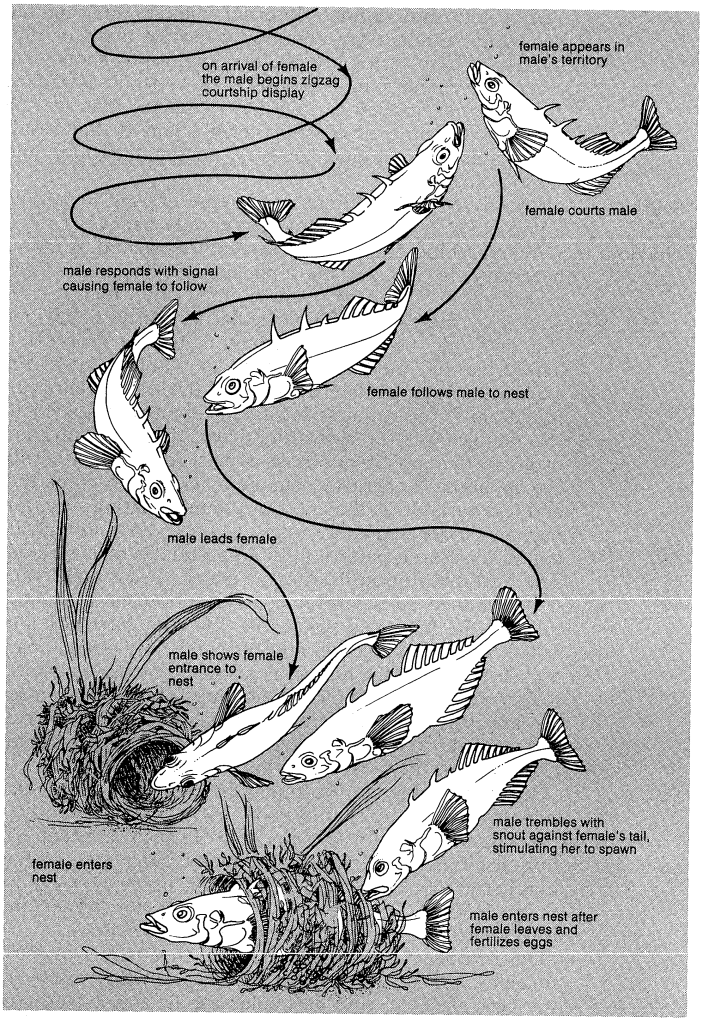


Figure 3: Mating behaviour of the stickleback.
From N. Tinbergen, *The Study of Instinct*, copyright © 1962; the Clarendon Press, Oxford

More than one drive is often activated simultaneously by the same situation. A conflict between the drives occurs, and the situation must be resolved. The resolution may occur in any one of several ways. Sometimes two drives may be expressed simultaneously. Pecking and head turning, when activated together, often occur simultaneously in chickens. Chickens that are in conflict between watching out with an elongated neck and making wide sweeping movements of the head as they peck, elongate the neck even further but reduce the extent of the head sweeping. Such conflicts can also be resolved by alternately performing the activities appropriate to the conflicting drives. The separate activities are often incomplete. A housewife, for example, may perform in this manner if the telephone begins ringing just as something begins to boil over on the stove.

Redirection and displacement. Sometimes an animal has a drive to perform a particular behaviour but is prevented from doing so and directs the behaviour to another object. If an animal is prompted to attack another but is prevented by fear of the opponent or by a reluctance to leave its territory, it might attack a harmless companion, the ground, vegetation, or even itself. Such behaviour is termed redirection. Displacement is the resolution of a conflict situation in which a seemingly irrelevant activity is performed. When an animal is obvi-

ously in conflict between, for example, sex and aggression or between aggression and fear, it will often perform an apparently irrelevant activity such as grooming, feeding, or scratching, or the animal may go to sleep. It is as if the two activated drives neutralize one another and the surplus energy is fed into another system. In disinhibition, two drives that appear to independently inhibit a third mutually inhibit each other in a conflict situation and lose their inhibiting effect on the third, which then becomes free to activate its own behaviour.

Disinhibition hypothesis

The probable reason that displacement activities are so often comfort behaviours (*e.g.*, preening) is that such behaviours have a typically low threshold of performance; ordinarily, they do not have a particularly high priority and are, therefore, easily elicited when more urgent demands are not being made upon the animal. This explanation is consistent with the fact that, though the frequent performance of these displacement activities is essential to survival, they are seldom associated with any condition of urgency.

The nature of the displacement activity may also depend upon the immediate environment or upon the effects of autonomic nervous activation on the animal; sometimes the activity depends on both factors. A wood thrush caught, while sitting in a horizontal position, by a strong attack–escape conflict will wipe its bill on its perch. The same bird caught by the same conflict, while sitting almost vertically, will make perfunctory preening movements directed at its upper breast. The relationship of the bird to its immediate environment makes it easier to perform a particular comfort activity. Autonomic responses—the result of fear or aggression in man, for example—may cause an expansion or constriction of blood vessels, a tingling sensation in the skin, a tendency to urinate or defecate, and so on. Scratching and temperature-adjustment behaviours may sometimes be prompted by such physiological changes.

Transitional activity. In transitional activity, another type of conflict resolution, the animal is stimulated to perform a particular behaviour, but the required environmental stimulus becomes unavailable during the course of the response. The animal discontinues its initial behaviour and substitutes another behaviour that it initially had not "intended" to perform.

The common grackle demonstrates a transitional activity in the form of a threat behaviour termed a spread–squeak; the bird ruffles its plumage, then utters a squeak as it compresses its feathers. If a rival approaches, prompting a spread–squeak, and turns away only at the plumage-ruffling stage, the bird will then, instead of squeaking, often shake its body—a normal comfort activity customarily preceded by plumage ruffling. When a man offers his hand to be shaken and it is not accepted, he often behaves as if he had really intended to gesture or perform a similar action. Transitional activities may be cases of displacement in which the nature of the behaviour is determined by the immediate environment. A grackle involved in the ruffling stage of a spread–squeak, after the rival leaves, may simply shake its body because, in the absence of further stimuli for fear and aggression, it is left with the stimulus for shaking, which may be ruffling. Conflict situations are of great interest from an evolutionary standpoint, because they are often the raw material from which signals have evolved.

BEHAVIORAL EVOLUTION AND DEVELOPMENT

Behaviours are believed to evolve in the same way as structures. The recombination of genes afforded by sexual reproduction ensures that each individual differs in some degree from all others of its species. Even slight variations from the norm increase or decrease the probability of survival. Advantageous features tend to be conserved and disadvantageous ones eliminated. Marked changes are the result of the slow accumulation of small variations over long periods of time, representing many generations. A species never becomes totally adapted to its environment because the environment is constantly changing. As a result, selective pressures always exist, and the process of evolution continues.

Selection in domestic animals. Animals can be selectively bred for specific behavioral changes. Many domestic animals differ markedly in behaviour from their wild progenitors. Domestic breeds such as fighting cocks and Siamese fighting fish are hyperaggressive, but most domestic animals tolerate greater crowding and are more docile than their wild ancestors. House mice have been selectively bred in the laboratory to produce unusually aggressive, as well as unusually timid, strains. Mating selection in domestic animals is usually less restrictive than in their wild counterparts. Reciprocal signalling systems between animals become less precise with domestication, and behavioral components may be omitted or lost altogether. Promiscuity may replace pairing in certain animals.

Courtship in pigeons

Marked differences in courtship displays occur between wild pigeons (*Columba livia*) and domestic breeds. Wild males loudly clap the wings over the back in flight and then glide with the wings held well above the horizontal position. In pouter pigeons (a breed of *C. livia*), the wings are clapped so frequently that two-thirds of the length of the primary wing feathers may thus be worn off, with the result that flying becomes very difficult. The elevation of the wings during the glide phase becomes so exaggerated that the wing tips touch, and the bird quickly loses altitude. In roller pigeons, another breed, the gliding flight has become a series of backward somersaults. On the ground the male wild pigeon, while cooing, twirls and makes a small hop when the female walks away. The German ringbeater breed elaborates on this behaviour by performing a wing-clapping flight around the female, who remains on the ground.

Domesticated zebra finches (*Poephila guttata*) show marked loss of specificity in their mating interactions and in care of the young, when compared with their wild counterparts. Wild chickens will kill their own chicks that lack specific colour patterns. Domestic chickens, on the other hand, will care for almost any chick regardless of colour and pattern; yet, they retain specific reactions to the species-specific calls of chicks so that they do not ordinarily accept other young birds, such as ducklings. Highly domesticated chicken breeds, such as Plymouth Rocks and barred Plymouth Rocks, however, will rear even ducklings. The least domesticated breeds, such as certain game bantams, still show much of the specificity characteristic of wild birds. Wild graylag geese form pairs only after a very long courtship period and remain monogamous. Domestic derivatives, on the other hand, pair quickly with any member of the opposite sex and are not monogamous. All of these differences between wild animals and their domesticated derivatives have a genetic basis.

Behaviour in hybrids. Two closely related species of small African parrots, the peach-faced lovebird (*Agapornis roseicollis*) and Fischer's lovebird (*A. personata fischeri*), have completely different methods of carrying nesting material. The females of both species prepare nesting material by cutting long, narrow strips of bark, leaves, or paper. The peach-faced lovebird tucks each strip, after she cuts it, into the feathers of the lower back, or rump. When she has accumulated about six strips, she flies to the nest cavity, retrieves the strips, and places them in her nest. Fischer's lovebirds carry each strip in the bill, one at a time, to the nest cavity.

Female hybrids between these two species initially tuck nest material into their rump feathers, but the strips fall out before the birds reach the nest. The birds gradually develop, through learning, an increased tendency to carry each strip singly in the bill. About four months after the onset of the tucking behaviour, they are utilizing both behaviours about equally. Although the tendency to carry in the bill continues to increase after this point and the tendency to tuck continues to decrease, the rate of divergence between the two methods becomes much slower. By the end of the third year the hybrids carry all strips in the mouth, but they make small intention movements to tuck. These intention movements consist of little ticlike, side movements of the head just before the bird flies off to the nest.

The courtship behaviour of male hybrids, paired with female hybrids of this same cross, is intermediate between that of the two parental-species males. When the hybrid males are paired with parental-species females, their courtship behaviour, in most cases, is closer to that of the parental species of the female, although it sometimes remains intermediate. The species-typical behaviour of the females is thus seen to influence the pattern of male courtship. The courtship behaviours of some bird hybrids are not so greatly modified; for them, no permissible variability has been inherited.

Behaviour inheritance of crickets

Two cricket species, *Gryllus campestris* and *G. bimacularis*, are so similar morphologically that they can be distinguished from one another only with great difficulty. Their behaviours, on the other hand, differ markedly. If the two species are crossed, however, the inheritance patterns may be traced by means of behaviour. Four behaviour patterns—antennal vibration in the post-courtship period, pendulum movements of the thorax, stridulation (rubbing one body part against another to produce sound), and fighting by young adults—have been investigated in particular detail. It has been found that antennal vibration and juvenile fighting in the hybrids have a monofactorial inheritance (*i.e.*, are caused by a single gene). The pendulum-like movement of the thorax during mating is found only in *G. campestris* and has a polygenic basis (*i.e.*, is caused by more than one gene). The stridulating sounds preceding courtship, performed only by *G. bimacularis*, are seemingly based on one pair of alleles (*i.e.*, different forms of a single gene).

Two races of honeybees are distinguished from one another by the presence or absence of hygienic behaviour. The race exhibiting hygienic behaviour opens comb cells containing dead pupae, which are removed. The nonhygienic race leaves dead pupae in their cells. The first generation (F_1) of hybrids contain only nonhygienic bees. One F_1 queen produces four kinds of drones, or males. When the F_1 is backcrossed with the hygienic form, a second generation (F_2) is obtained, which is made up of four different types of bees. One group is hygienic; one group opens the cells of dead pupae but does not remove the dead pupae; one group does not open the cells of dead pupae but removes the dead pupae if the cells are open; and the remaining group is nonhygienic.

The influence of experience on behaviour. The adaptive change of behaviour as the result of experience—usually known as learned behaviour or experience-dependent behaviour—may be observed in all higher vertebrate forms. Several types of such learning in animals are recognized: habituation, classical conditioning (CR Type I), trial and error (CR Type II, instrumental conditioning, or operant conditioning), latent learning, insight learning, and imprinting.

Habituation. Habituation is learning to disregard stimuli that are without significance to the animal. In many respects it is the simplest form of learning, and it is sometimes regarded as a fundamental property of all living matter. Most animals inherit a response to be frightened by sudden and strong stimuli such as loud sounds, flashes of light, and the sudden intrusion of anything foreign into the animal's sensory field. Yet, if an animal reacts nonselectively to all phenomena such as rustling leaves, thunder, snapping twigs, and the sudden appearance of harmless animals, those phenomena that are significant to the animal's well-being will not receive the intensity of response that they often require. All animals, then, quickly habituate to such harmless stimuli, but the adaptation is highly specific. An animal that habituates to one type of sound does not, as a consequence of this habituation, become habituated to other sounds. Habituation is distinct from failing to respond to stimulation as a result of fatigue, sensory adaptation, or injury. The effects of habituation are generally long lasting. If an animal is repeatedly exposed to a potentially harmful stimulus (such as to a predator) without being harmed, habituation does not generally occur. Responses to dangerous stimuli often seem to have an inherited resistance to habituation—a mechanism of obvious survival value.

Salivation experiments of Pavlov

Classical conditioning. Classical conditioning was studied early in the 20th century by the Russian physiologist Ivan Pavlov, who observed that dogs salivate when food is placed in their mouths. He gave dogs food and, at the same time, provided another stimulus such as a flashing light or the sound of a bell. After a few such pairings of stimuli, a dog would salivate upon seeing the light or hearing the bell but without the presence of food. The dog had learned to associate the flashing light or the sound of a bell with food. A previously irrelevant stimu-

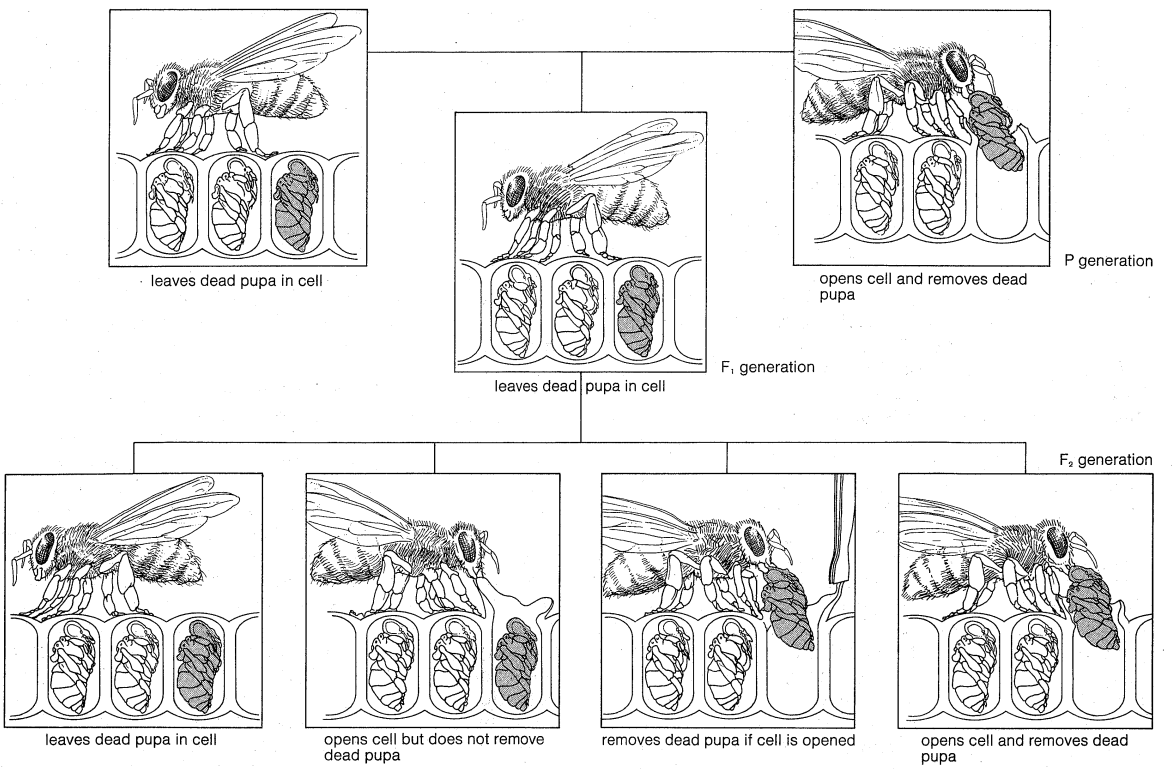


Figure 4: Inheritance of behaviour in bees.

lus that assumes significance as a result of association with a relevant stimulus is called a conditioned stimulus (CS). Salivation in response to such a stimulus is termed a conditioned response (CR). Prior to the learning experience, only the food (unconditioned stimulus) was effective in producing salivation (unconditioned response).

Such conditioned responses have been observed in a wide variety of animals, from lower invertebrates to man. Birds learn to avoid noxious insects in this manner; the distasteful monarch butterfly (*Danaus plexippus*) or a species of stinging wasp provide effective stimuli that quickly become associated with the appearance of such insects. This kind of association together with the conditioned response relevant to it is also the basis for Müllerian mimicry, in which palatable insects and other animals evolve to resemble noxious ones, thus enhancing their chances of survival.

Trial and error. In trial-and-error learning, an animal learns to behave in a particular way by associating something it does with a desired effect. If a dog's foot is lifted by the experimenter and food then given, the dog, after a few such trials, will spontaneously lift its foot in anticipation of food. Both classical conditioning (CR Type I) and trial-and-error learning (CR Type II) are termed associative learning, because in both cases an unconditioned response is associated with a conditioned stimulus. In natural situations an animal probably learns to associate certain spontaneous activity of its own with certain desired results, thus fixing the conditioned stimulus and response. Learning of this type often occurs when animals modify their behaviour during appetitive sequences such as those involving feeding and mating.

Latent learning. Latent learning is the association of indifferent stimuli or situations with one another without reward. The phenomenon is clearly exemplified in exploratory behaviour. Animals finding themselves in unfamiliar environments or among unfamiliar objects, but in familiar surroundings, show exploratory behaviour. The animal uses its sense organs on all that is novel, shows much ambivalence between approach and avoidance, and, finally, as the hesitancy to approach wanes, will test the novelty. A mammal may, at this point, sniff, nudge, or handle a strange object; a bird may peck at it. The first contact often results in abrupt avoidance, but, typically, the object or situation is, at last, thoroughly explored and finally abandoned if neutral. A mouse sniffs and pokes about most of a new environment. A bird, having sight and hearing, rather than smell and touch, as dominant sensory modalities does not have to occupy physically as much of a novel environment. Instead, it places itself successively at several vantage points and then carefully peers about and listens.

Subsequent behaviours of an animal can reveal that it has learned much about its environment during such an exploratory phase. It may learn, for example, the physical features of the environment and their spatial relationships to one another, the location of food and water, and the location of places safe from predators. Animals apparently learn all of these things early in their exposure to an environment, even though the information acquired may become significant only at some time in the future. The learning takes place without being associated with immediate reward (unlike that in conditioned response Types I and II) unless a need to know is postulated as a kind of immediate self-reward.

Insight learning. Insight learning is believed to be an advanced type of learning. Insight involves the spontaneous combination of a number of isolated experiences; the result is a new experience that is effective for gaining a desired result. Humans are able to exercise insight; it is extremely difficult, however, to identify such behaviour in most other animals. Animals are believed to use insight when they solve a problem too rapidly for normal trial and error to occur. It is possible that such an animal is carrying out trials in its brain; this implies reasoning ability. The higher primates are probably capable of insight learning at times, but further down the phylogenetic scale the evidence of such learning becomes progressively less conclusive.

Insight in chimpanzees

Chimpanzees, to get food out of reach, will pile boxes to make a stand for themselves or will fit sticks together to knock the food down. These solutions may come quickly, obviously benefitting from prior experience (latent learning). Much trial-and-error learning is demonstrated, however, when they actually pile boxes or fit sticks together. In humans, insight is probably often aided by latent learning and trial and error.

Imprinting. Imprinting, a learning process observed in young birds and mammals, is the identification of an animal with another animal. Normally, it is a relationship between members of the same species, but it can occur, for example, between a bird and a human. Imprinting can take place only during a particular period of the animal's development—a time span that is specific for each species.

In 1935 the Austrian ethologist Konrad Lorenz first observed the process in ducklings and goslings. After goslings hatch and become dry, they follow their parents. The adults provide warmth, safety, and shelter and bring the goslings food. The more uncomfortable a gosling becomes (*e.g.*, cold, frightened, hungry), the more intensely it follows. If goslings are reared by a human, they become imprinted to humans; thus, they ignore geese.

The development of this response occurs during a sensitive period, before and after which the response cannot be learned; if the response is not acquired during the sensitive period, it will never occur. Zebra finches that are isolated from their own species before they are 35 days old are never able to distinguish males and females of their own species. This is because their sensitive period for imprinting occurs before they are 35 days old.

The duration and time of onset of the sensitive period depend on the species and on the type of behaviour involved. Some animals imprinted to animals of another species will mate with members of their own species but, if given a choice, will prefer the animal to which they have been imprinted. Many species refuse social contact with any animal except the one to which they are imprinted. Male golden pheasants (*Chrysolophus pictus*) imprinted to humans will court females of their own species but immediately transfer this behaviour to a human, should one appear. The same is true for budgerigars (*Melopsittacus undulatus*) and turkeys (*Meleagris gallopavo*). Mallard ducks imprinted to humans, on the other hand, will not associate with members of their own species (conspecifics) and will continue throughout their lives to treat humans as conspecifics. Imprinting is fixed for life, in contrast to other types of learning, in which forgetting is common. Imprinting of motor patterns, such as birdsong, also occurs. Exposure to a particular birdsong may be relatively brief and still be permanently fixed in the bird's memory. Chaffinches (*Fringilla coelebs*) learn their songs during the first 13 months of life, although they do not sing until nearly a year later. A 12-day-old nightingale (*Erithacus megarhynchos*) was kept in the same room with a singing black-capped warbler (*Sylvia atricapilla*) for about one week. The following spring the nightingale sang a typical black-capped warbler song.

Little is known about imprinting in mammals, but hoofed animals, such as sheep and horses, that are imprinted to man express this response by following him about. Dogs from four to six weeks old develop normal social responses to dogs or to another species such as man. Imprinting apparently also occurs in humans. An infant deprived of its mother for a short period during its first year may develop serious mental retardation. A separation of several months—particularly during the seventh to the 12th month—will frequently result in irreparable damage; under such conditions, death may result.

Imprinting in man

Play behaviour and curiosity behaviour. Play and curiosity are exhibited by many mammals and by some birds and figure importantly in the learning of numerous activities. Play is especially characteristic of young animals, but the adults of many species also engage in it. Spontaneous curiosity, in which the animal actively seeks out novel situations for exploration, is exhibited by the young of mammals and some birds; indeed, they seem

to be under the compulsion of some drive to do so. Carnivores and primates exhibit more curiosity than rodents, which gnaw novel objects and may hoard them. Monkeys inspect and manipulate such objects.

The curiosity drive implements the development of new motor skills and ensures the acquisition of new perceptual impressions, thereby resulting in new knowledge. The only reward, however, seems to be the performance of the activities themselves. Rhesus monkeys will learn a puzzle game without any reward except its successful solution. Among rats, it has been observed that the nerve cells in the lateral hypothalamus and preoptic regions of the brain are more active in those rats that explore. Electrical stimulation of these brain areas is rewarding to rats, and they will learn to press a lever that activates this stimulation.

Such curiosity behaviour seems linked to play behaviour. Play is difficult to define; it is usually easy, however, to distinguish a playing animal from one that is seriously occupied. An animal plays only when it is satiated and not preoccupied with other tasks. Play seems not to be dictated by immediate need but is extremely important in behavioral development. Only animals that spontaneously seek new situations on their own initiative play in the true sense. Invertebrates, fish, and amphibians do not seem to play. The taxonomic distribution of play among mammals and birds suggests that play is related to learning. Play involves interactions with the environment; this leads to the acquisition of knowledge about environmental features, including information about conspecifics and the animal's own possibilities of movement. Play behaviour occurs only at particular times; progression to a second play activity takes place only after a certain level of skill has been achieved in the first.

Mock fighting and fleeing

Much play appears to be fighting or fleeing behaviour, and usually it is easily identified as such. An animal that is play escaping or play attacking does not actually escape or attack. A rodent play fleeing into a hole, for example, quickly reappears. If a rodent's flight is truly an effort to escape, it reappears only after a much longer interval. Play-fleeing animals often reverse roles quickly, and the pursuer becomes the pursued. Threat behaviour that is associated with real attack is missing, and there is strong reluctance to bite. Play tends to be highly repetitive. A dog may retrieve a stick many times or play fight until it is exhausted or until a more interesting activity distracts it.

Such play behaviour could mistakenly be postulated as the performance of immature instinctive activities. In many instances, however, this is known not to be the case. Much playful behaviour occurs at a time in an animal's life when it is fully capable of serious activity. Play also involves the use of species-typical patterns of behaviour in various sequences that do not occur in serious activity.

Modification of instinctive behaviour by experience. Behaviours based on both instinct and learning are commonly intercalated into functional wholes. In the peach-faced lovebird, for example, the cutting of nest-material strips is partly instinctive and partly the result of experience. The propensity for cutting is instinctive. This includes punching holes in the sheet of material from which the strips are fashioned and a "knowledge" of the proper width, length, and straightness of strips. The spacing of punch holes in such a manner as to form a strip is learned through experience. It is as if the animal has an instinctive picture in its central nervous system and persistently tries punching holes, in various relationships to one another, until the right pattern is made. The bird tends to repeat punching patterns that most closely approach the ideal and, thus, gradually, through progressively more satisfying feedbacks, approach the definitive functional technique of cutting strips. Idiosyncratic techniques develop—some birds stand on the sheet while cutting, and others stand off the sheet; some cut to the left, others to the right, and still others cut in various combinations of these directions. Birds developing their techniques try all of the directions and places of standing but gradually act with less and less variation. They do not need to observe experienced birds in order to develop a technique. The stimuli necessary for learning response are intrinsic to the birds' individual activities.

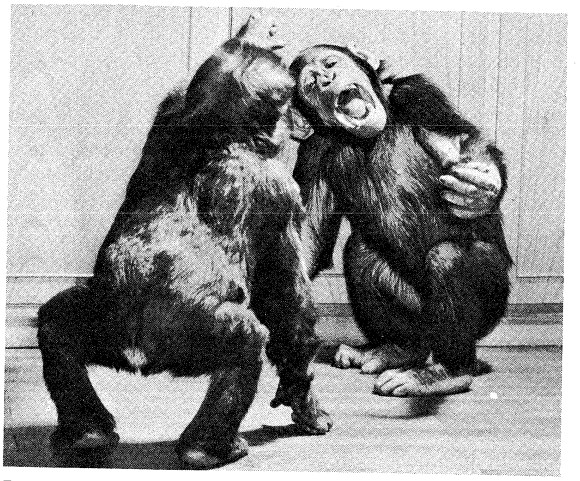


Figure 5: Young chimpanzees exhibiting mock fighting behaviour.
Bernhard Grzimek

Song learning

If a meadowlark (*Sturnella magna*) is exposed to an alien song during its sensitive period for learning song, it will learn the alien song. Meadowlarks, then, learn their species-typical songs rather than inherit the capacity for particular melodies. But, if they are exposed during the sensitive period to alien songs along with meadowlark songs, they will learn only the normal species-typical song. Although the song must be learned, the bird instinctively learns the species-typical song if there is a choice. The bullfinch (*Pyrrhula pyrrhula*) instinctively learns the male parent's song rather than the species-typical song. Bullfinches raised by foster parents of another species will learn the song of the male foster parent, even though the normal bullfinch song is also audible to it during the same period. In both types of song acquisition, learned and instinctive elements combine in the development of the species-typical song. Some species, however, do not learn their songs and do not need the experience of hearing others sing during their development. Different vocalizations in the same species are commonly acquired in more than one way. Some are purely instinctive; others are learned.

The behaviour of an individual animal is the result of a genotype that has developed over millions of years of evolution—a genotype that also permits a certain degree of variability through experience.

HORMONAL AND NERVOUS CONTROL OF BEHAVIOUR

Interaction of endocrine and nervous systems. Physiological changes within an animal are largely the direct or indirect result of nervous and endocrine (hormonal) changes and their interactions. These changing states are also responsible for a changing responsiveness to internal and external environmental stimuli.

Neurosecretion

The endocrine system and the nervous system are probably of equal evolutionary age; the two systems may be evolutionarily linked. Certain nerve cells, or neurons, that are highly modified produce substances that pass through the axons (threadlike extensions of neurons) and into the bloodstream. These neurosecretory cells are sometimes clustered, forming glands that have connections with both the nervous system and the bloodstream. The endocrine glands are similarly distributed. Some may have evolved from clusters of neurosecretory cells. The vertebrate pituitary gland has evolved as a fusion between neural tissue and epithelial tissue (the lining, or covering, of organs) and is intimately associated with the hypothalamus, a ventral portion of the brain. The pituitary may be regarded as a master gland that regulates, with its secretions, all the other endocrine glands. The nervous system, in turn, has a regulatory effect upon the pituitary, which may also be influenced by feedback effects stemming from the secretions of other endocrine

glands. This relationship permits the outside and inside environments to exert influences on the endocrine system. Seasonal changes in day length, for example, influence the nervous system by means of visual stimuli. The endocrine system is then activated through stimulation of the pituitary gland by the hypothalamus; the pituitary, in turn, secretes hormones appropriate to the most adaptive response. The ultimate effect of a seasonal change in day length may be migration or reproduction (see ENDOCRINE SYSTEMS).

The nervous system and the endocrine system interact with and complement each other. The nervous system sends information with great speed but of short duration along its pathways. Its messages can change rapidly. Hormones, secreted by the endocrine system into the bloodstream, travel much more slowly than nervous inpulses. The endocrine system can keep a message constantly available for many months if necessary. The nervous system generally affects only muscles and glands, but hormones can reach every cell in the body. Adrenaline is a hormone that acts with relative speed. It is secreted by the two adrenal glands, which are attached to the kidneys. The adrenals consist of two portions that differ from one another both in origin and in function. The inner portion is the adrenal medulla, and the outer portion is the adrenal cortex. During stress, such as occurs in fighting, mating, and fear, the adrenal medulla is stimulated by the autonomic nervous system to release adrenaline into the bloodstream. Some of the changes that occur throughout the body under the stimulus of adrenaline include hair erection, sweating, and acceleration of heartbeat and breathing; adrenaline also causes the blood to be diverted from the digestive tract to the muscles. All of these changes, and others, help prepare the animal for extreme effort. During brief stressful periods the bloodstream is quickly flushed with adrenaline, but the hormone is quickly dissipated. If the stress situation persists, other events take place—the adrenal cortex becomes involved, and its hormones are released. The cortex, unlike the medulla, is not under direct nervous control but is stimulated by another hormone, the adrenocorticotrophic hormone, or ACTH, produced by the pituitary gland. Prolonged stress stimulates cells in the hypothalamus, which, in turn, stimulates the pituitary gland to produce ACTH. The cortical hormones in turn stimulate various responses to prolonged stress. Some of these responses are concerned with the metabolism of glucose, a sugar that may be associated with the utilization of food reserves. The effects of cortical hormones are, in any case, profound, and continuing stress will result in enlargement of the adrenal cortex, which leads to increased production of the cortical hormones. Chronic stress may cause severe illness and even death. It has been shown that rats confined to the territories of other rats will die as the result of overproduction of cortical hormones in response to the stressful situation. Stress as a result of overcrowding may also cause death in animals. Such stress may, in fact, be the cause of a decline in numbers of mice after they have reached a certain population level in a given area.

ACTH

Sex hormones. The pituitary hormones that have a direct effect upon reproductive behaviour are the follicle-stimulating hormone (FSH), the luteinizing hormone (LH), and prolactin (lactogenic hormone or luteotropic hormone). FSH and LH are called gonadotropins because they stimulate the gonads (ovaries and testes) to produce germ cells and gonadal tissue; gonadal tissue, in turn, secretes other hormones. Prolactin has a variety of effects. In different species of vertebrates, prolactin affects different target organs. In female mammals, for example, it stimulates growth of the mammary glands and the secretion of milk. It also stimulates the corpora lutea —glandular bodies of the ovary—causing them to produce another hormone, progesterone. In pigeons and doves, prolactin causes the characteristic modification of the crop (stomach) associated with the production of so-called pigeon's milk—a soft white substance that is passed from the mouth of the adult pigeon to that of the young. A slow change of colour in some fish is caused by the influence of prolactin on pigment cells. (Rapid colour changes are under nervous control.)

The gonads secrete hormones from special cells when stimulated by FSH and LH from the pituitary gland. Collectively, the female hormones are termed estrogens, and the male hormones are called androgens. Both androgens and estrogens belong to a chemical group known as steroids. All steroids have closely related chemical structures; the different vertebrate groups have slightly different steroid hormones that seem to be largely interchangeable in function. Removal of the testes or ovaries (castration) in vertebrates causes profound changes in behaviour and structure, especially if done early in life. Among many invertebrates castration has no such profound consequences. Apparently, therefore, only in the vertebrates are the gonads important endocrine organs.

Effect of castration

Androgens and estrogens control the development of the secondary sexual characteristics; they also effect the production of eggs and sperm from the gonads. Secondary sexual characteristics tend to be relatively permanent throughout life, but many are temporary, occurring only during the breeding season. Examples of temporary characteristics include the special breeding plumages and songs of many birds, the antlers of deer, colour changes in some fish, and all the behaviour associated with the formation of fertilized eggs (zygotes).

Progesterone, the production of which is stimulated by prolactin, is produced in mammals by the ovary. After an egg is shed, the empty follicle enlarges, forming the corpus luteum, a conspicuous yellowish structure, which secretes progesterone; progesterone, in turn, stimulates changes in the uterus preparatory to its receiving the fertilized egg. Progesterone also inhibits the contraction of uterine muscles. The females of other vertebrate groups produce structures similar to the corpora lutea of mammals. Birds, which have rather inconspicuous corpora lutea, also produce progesterone.

The amount of any hormone normally present in the bloodstream is minute. Artificially high levels often have marked effects; the introduction of massive doses of testosterone into females, for example, causes male sexual behaviour. It will also cause female sexual behaviour in both sexes, in which case testosterone may be acting as a general stimulant.

Sometimes a massive dose of hormone has an effect opposite to the one expected. Female canaries (*Serinus canarius*) given overdoses of estrogen would be expected to show enhanced sensitivity of their brood patches (highly vascularized areas of skin in close contact with the eggs during incubation); instead, overdoses of estrogen may result in some desensitization of these areas.

It is tempting to conclude that the sex hormones have a direct effect on all the structures and behaviours concerned with the formation and nourishment of zygotes. Much is not known, however, about the hormonal control of behaviour. There may be feedback effects after the hormones have initiated a particular response; for instance, estrogen in conjunction with progesterone in some birds causes increased thickening and vascularization of the brood-patch area. The target tissues in this case (skin, blood vessels, and feather follicles) all have a nerve supply, and feedback through them to other parts of the nervous system could initiate further behavioral modifications. Physiological changes resulting from hormonal action may render an animal more or less responsive to its environment and thus modify its behaviour.

There is evidence that hormones directly affect behaviour by acting directly on neurons in the hypothalamus. If estrogen is injected into certain areas in the hypothalamus of castrated female cats, they develop strong estrous behaviour, even though the reproductive system remains underdeveloped.

Behaviour may stimulate hormone production. Female cats, rabbits, and some other mammals are "induced ovulators." In other words, copulation stimulates the hypothalamus via the nervous system, and the pituitary gland is then stimulated to produce luteinizing hormones (LH), which in turn affects the ovaries. A few hours after copulation ovulation occurs at about the time the sperm have

Behavioral stimulation of hormone production

reached the upper levels of the reproductive tract; the germ cells meet, and fertilization takes place. In deer, sheep, and weasels, stimulation of the pituitary gland (resulting in the production of follicle-stimulating hormones [FSH]) is achieved when the animal senses a change in day length. The females of other mammal species (*e.g.*, the house mouse [*Mus musculus*]) seem to have an internally based clock that periodically triggers the release of FSH regardless of environmental changes. They may still be influenced, however, by the presence of a male.

Complex interactions between hormones and behaviour are known to occur in birds. The sight of a courting male stimulates the release of FSH and LH in a female dove (*Streptopelia risoria*). Physical contact between the birds is not necessary for this response to take place.

Nervous system and behaviour. The nervous system receives information about the external environment through sensory receptors and about the internal environment through hormones, internal neural responses, and other physiological events. This information, regardless of the source, is processed in the brain or spinal cord, and appropriate responses are initiated by outgoing (efferent) nerve impulses leading to muscles or glands.

There is much evidence to support the view that the central nervous system (CNS) is hierarchically organized. Its organization is thought to consist of a system of centres, each with the function of collecting stimuli and appropriately redispatching them. Reproduction in the peach-faced lovebird depends first upon the activation in the proper sequence of a number of subordinate centres. These involve formation of a pair bond between a male and female; selection of a nest site; conduction of courtship; laying and incubating of eggs; and caring for the young. Nest building occurs throughout courtship, egg laying, incubation, and care of the young.

Subordinate events such as those mentioned above control, in turn, other neurally controlled events. Nest building, for example, consists of various subordinate activities that have already been described. Each of these subordinate activities also has subordinate activities. Tucking a strip of nesting material consists of several activities: simultaneously turning the head back over the rump, lowering the unfolded wing on the same side, and erecting the rump feathers; pushing the strip into the feathers, performing rapid hooking movements, which seem to function as an anchoring behaviour; releasing the strip; and, finally, simultaneously bringing the head, wing, and rump feathers back to the normal position.

It has been proposed that the smallest irreducible neuromuscular coordinations of an activity be thought of as acts. Each species is capable of performing a finite number of acts, and these are combined in various ways to produce all the behaviours of which an animal is capable. Each act is thought to have an act centre in the CNS. The act centres are subordinate to the behavioral centres, which coordinate them; behavioral centres are, in turn, subordinate to their initiating and coordinating centres and so on. The term centre in each case refers to a functional rather than to an anatomical locus. Portions of the CNS responsible for mediating a given response may be quite diffuse anatomically; typically, there may be much redundancy—a condition that frequently enables an animal to regain normal behaviour following damage to the CNS. Brain tissue, however, does not regenerate.

Act centres in central nervous system

Detailed implementation of various subordinate activities may depend upon the details of environmental feedback, as in the taxis components of many fixed action patterns (see above *Fixed action patterns*). Such detailed implementation may also depend upon long-term changes in behaviour that result from experience. Various releasing mechanisms may also be altered as a result of experience, as may the significance of various environmental stimuli. Experience may therefore exert modifying effects upon the input of information, its mediation in the CNS, and the details of its implementation through muscle coordination and glandular activity. These experiential effects may be of long or brief duration.

For many years the classical reflex was thought to explain adequately the mechanism of various behaviours. Such a reflex consists, in its simplest form, of an afferent neuron carrying information to the CNS, excitation then being carried to an effector (muscle or gland) via an efferent neuron. Intermediate neurons often occur between the afferent and efferent neurons. This entire mechanism is termed a reflex arc. In certain reflexes, the excitations activate the same muscles or glands from which the stimuli originate (monosynaptic reflexes). Most reflexes have intermediate neurons, and the stimulation of a single receptor may activate many effectors; similarly, the stimulation of many receptors may activate but a single effector. Chain reflexes occur as a result of one reflex triggering another. Additional reflex arcs may become established as a result of experience (conditioned reflexes), and some reflexes are thought to have facilitatory or inhibitory effects on others.

It is now known that not all behaviours are the result of afferent impulses. Early in the 20th century it was found that cats' leg muscles with all afferent nerves removed still demonstrate rhythmic movement. Afferent impulses are not necessary for coordinated response; the swimming movements of eels and other fish, the crawling of earthworms, and the flying movements of grasshoppers are some examples. It is now known that spontaneously generated stimulation of the central nervous system initiates and controls much behaviour.

The evolutionarily older portions of the CNS, such as the spinal cord and the medulla and hypothalamus of the brain, seem to be concerned mainly with inborn behaviour such as heartbeat, breathing, and reflexes and with instinctive behaviour. The evolutionarily newer portions of the brain, such as the cerebrum of mammals, seem to be concerned either with new behaviours resulting from experience or with the modification of inborn behaviours. The conspicuous cerebrum of mammals often comprises a major portion of the brain. Although it is generally accepted that mammals, as a class of vertebrates, are the most intelligent of all animals, many birds seem to be capable of greater modification of behaviour through experience than are some mammals. Many other vertebrates, such as certain reptiles and fish, are capable of learning new behaviours rather easily. It was long believed that, because birds and other nonmammalian vertebrates have little or no discernible cerebral tissue, their behaviours were instinctive. It is now known, however, that such animals are often capable of much behavioral modification as a result of experience and that some other portion of the brain must be involved. Different portions of the brain in different animal groups seem to have been selected for the development of learning ability.

BIBLIOGRAPHY. KONRAD LORENZ, *Er redete mit dem Vieh, den Vögeln und den Fischen*, 6–8th ed. (1952; Eng. trans., *King Solomon's Ring: New Light on Animal Ways*, 1952; reduced photographic reprint, 1961), is a highly recommended popular treatment. Semipopular works include KONRAD LORENZ, *Das sogenannte Böse* (1963; Eng. trans., *On Aggression*, 1966); DESMOND MORRIS, *The Naked Ape* (1969) and *The Human Zoo* (1969); NIKOLAAS TINBERGEN, *The Herring Gull's World: A Study of the Social Behaviour of Birds* (1960), *Curious Naturalists* (1958, reprinted 1968), highly recommended to all interested in behaviour, and with the EDITORS OF LIFE, *Animal Behavior* (1965), an excellent résumé of ethology.

The following are technical studies that are mostly understandable to the educated nonspecialist: MARGARET BASTOCK, *Courtship: A Zoological Study* (1967); IRENAUS EIBL-EIBESFELDT, *Ethologie, die Biologie des Verhaltens* (1966; Eng. trans., *Ethology: The Biology of Behavior*, 1970); WILLIAM ETKIN (ed.), *Social Behavior and Organization Among Vertebrates* (1964); ADOLF PORTMANN, *Der Tier als soziales Wesen* (1953; Eng. trans., *Animals As Social Beings*, 1961); ANNE ROE and GEORGE G. SIMPSON (eds.), *Behavior and Evolution* (1958); KENNETH D. ROEDER, *Nerve Cells and Insect Behavior*, rev. ed. (1967); and CLAIRE H. SCHILLER (ed.), *Instinctive Behavior* (1957). More advanced are R.A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology*, 2nd ed. (1970); P.H. KLOPFER, *Behavioral Aspects of Ecology* (1962), and (comp.), *Behavioral Ecology* (1970); KONRAD LORENZ, *Evolution and Modification of Behavior* (1965); and WLADYSLAW SLUCKIN, *Imprinting and Early Learning* (1964).

(W.C.Di.)

Beirut

Beirut, the capital and major port of the Republic of Lebanon, overlooks the Mediterranean Sea at a point known as Jūn Mār Jirjis (St. Georges Bay). Backed by the Lebanon Mountains, it occupies a picturesque site of about 4,200 acres (1,700 hectares), containing a municipal population estimated at 475,000 (1970) and a metropolitan population of close to 950,000.

The name Beirut is derived from either the Aramaic *berotha* ("pine trees") or from Berytus, the Roman name for the city (from Phoenician Canaanite *beroth*, "wells"). Despite its eventful past and important geographic location as an entrepôt and gateway between East and West, the city was until 1860 no more than a small, fortified medieval town.

Sunnī Muslims constitute one-third of the population; other religious groups represented include Armenian Orthodox, Greek Orthodox, Maronites, Shī'ite Muslims, Catholics of Greek and Latin rites, Protestants, and Jews. Ethnic minorities include Armenians, Palestinians, and some 250,000 Syrians. Beirut's climate is Mediterranean; rainfall occurs mainly in winter.

History. One tradition claims that the city, one of the oldest on the coast, was built by the god Illion, who had married a goddess named Beirut. The first recorded reference is in the Tell el-Amarna tablets, in which Ammunira, a Phoenician vassal of Akhenaton (1379–1362 BC), mentions Beirut's strong defenses and prosperity.

Many conquering armies have descended on the city, from the time of Ramses II of Egypt (13th century BC) down to 1941, when British forces took the city. Destroyed by Tryphon, a usurper of the Syrian throne, in 140 BC, it later flourished under Roman rule (64 BC onward), one of its most glorious periods. Marcus Agrippa, Augustus' admiral, had visited it in 15 BC, after which it was granted the status of a Roman colony. The Herodian kings of Palestine embellished Beirut with hippodromes, an amphitheatre, baths, and porticoes, but the city's main fame was intellectual, based on its school of law which flourished between the 3rd and 6th centuries. In AD 551 a tidal wave accompanying one of a series of earthquakes destroyed the city, reducing the population to a few thousand. It was not until the 19th century that Beirut recaptured its past splendour.

Beirut as a Roman city

The Arabs occupied the city from 635 until 1110, when Baldwin I, the Flemish leader of the First Crusade, captured it. In 1151 Beirut was pillaged by the Egyptian fleet; from then until 1187, when it was captured by the Muslim leader Saladin, it was a scene of constant battle.

The city declined under Mamlūk rule (1291–1516), and only after this was ended by the Ottoman conquest did it revive; at this time the Druze emirs, who had won Ottoman favour, became masters of central and southern Lebanon. The most illustrious emir, Fakhr ad-Dīn II (who ruled from 1593 to 1635), made Beirut his winter residence. Trade relations with Western powers were revived; the city was once again fortified, and its famous pine groves were restored.

The Ottoman era

But in 1772 another period of conflict and decline began, following an attack by the Russians. An even fiercer attack by British, Turks, and Austrians in 1841 ended a ten-year occupation by Muḥammad ʿAlī of Egypt but left the city a ruin.

After a religious massacre in 1860, resulting from growing tensions between Maronite Christians and Druzes, an influx of Christians occurred. As a headquarters for French, American, and British religious missions, and as a growing centre of trade and services, Beirut attracted an increasing number of Europeans and Syrians. The burgeoning population spread beyond the medieval walls, already partly demolished, and new harbour facilities were built. The completion of the Damascus railway in 1895 assisted the city's expansion, and by 1900 the population had reached 120,000.

In 1912 the Italians, at war with Turkey, assaulted Beirut by sea. In 1918 the Allies entered the city. During the early years of World War II, Beirut was controlled by the French and British, until, in 1941, it was declared the capital of an independent Lebanon.

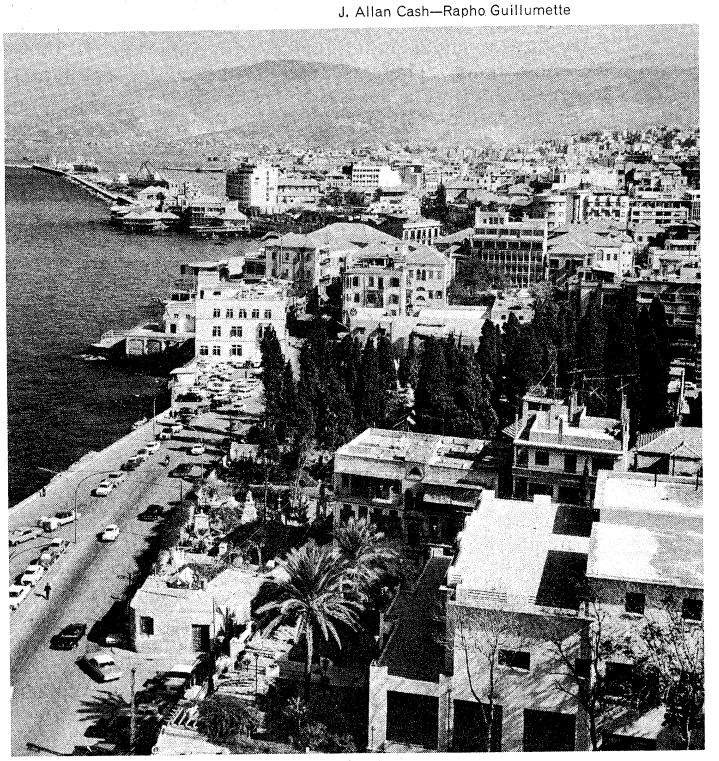
The contemporary city. Beirut is located on a triangular peninsula, bounded by the Mediterranean to the north and west, by a seasonally flowing mountain stream to the east, and by a 200-acre pine forest to the south.

By the early 1970s the population had increased nearly tenfold since the early 1930s, and the city area had grown to three times what it was in 1900 and to about 100 times what it was in 1800. Consequently, few traces of the old city are now to be found. Massacres in Turkey in 1914 resulted in an influx of some 50,000 refugees, who settled in dense slums around the city. Since the early 1950s, high-rise apartments and office buildings have transformed the residential quarters and altered the skyline.

The city's growth

Beirut reveals a curious blend of architectural styles and house types, ranging from flat farmhouses and red-tiled suburban villas to walk-up apartments and modern high-rises. Street plans and block arrangements are not consistent or uniform.

J. Allan Cash—Rapho Guillumette



Beirut meets the Mediterranean Sea at Jūn Mār Jirjis.

Traditional neighbourhoods. Beirut is a mosaic of distinct urban communities. Apart from the multiple agglomerations of village migrants and refugees established in specific settlements, one may identify three broad communities or urban subcultures within Beirut. Al-Ashrafīyah is a homogeneous Christian (mainly Maronite and Greek Orthodox) residential quarter, occupied by middle and upper income groups, and has definite Francophile leanings in both style of life and cultural orientation. Basta is almost entirely Sunnī Muslim and of visibly lower socioeconomic status; in style of life and political orientation the inhabitants are decidedly pro-Arab. Ra's Bayrūt, and Hamra Street in particular, is a fast-growing, fashionable, and cosmopolitan district with almost no distinct or unifying character. Though predominantly middle class and Anglo-Saxon in style of life, the area has a heterogeneous ethnic and religious composition.

Economic life. Modern seaports and airports, as well as Lebanon's free economic and foreign exchange systems, favourable interest rates, and a banking-secrecy law, have established Beirut as a pre-eminent financial and commercial centre of the Middle East. With almost 40 percent of Lebanon's population, the city dominates the national economy. As in its architecture, the economic life of the city continues to display a curious but viable blend of modern and traditional elements. Fashion-

able department stores in new commercial centres adjoin shopping arcades, supermarkets, and the bazaars of the traditional central business district. Numerous street vendors and mobile stalls are to be found on the side streets.

Manufacturing

Industrial establishments—mostly producing textiles, clothes, or shoes or engaged in food processing or printing—are located east and southeast of the capital, where there is easy access to road, rail, and sea transport. Numerous workshops and small cottage industries—mostly making upholstery, furniture, metalwork, or clothing—are scattered throughout the city.

The city's luxurious hotels, bars, nightclubs, restaurants, fashionable sidewalk cafes, sea resorts, and shopping facilities have made Beirut an all-year resort. Between 1960 and 1966, tourism was expanded at the rate of about 20 percent yearly.

Transport. Beirut is the centre of a well-developed transport system. A tree-lined boulevard encircles the city, and four throughways traverse it. Two other major roadways, linking the capital with suburbs, are under construction. There is much traffic, and congestion is great. In 1965 a bus system replaced the trams, which had served the city since 1908.

The Orient Express Railroad links Beirut with Europe but is mainly used for long-distance bulk transport.

The harbour, with three artificial basins and a fourth under construction, is a major port, with a "free zone" for duty-free handling and storage of transit shipments. The docks accommodate at least 20 ships and move about 4,000,000 tons of freight a year.

Beirut International Airport, regularly served by more than 30 international airlines, is one of the busiest in the Middle East. Its two runways can accommodate the largest jet aircraft.

Education and culture. As in Roman times, Beirut is an ebullient centre of learning and culture. The diversity and quality of its educational facilities attract an international student body. Despite the government's efforts to promote the expansion of state schools, the bulk of primary, secondary, and higher education in Beirut remains in private hands.

All of Lebanon's eight institutions of higher learning are centred in Beirut. The four major universities are the American University of Beirut; the Université Saint-Joseph, subsidized by the French government and administered by the Jesuit Order; the Lebanese University; and the Arab University, an affiliate of the University of Alexandria. There are also three museums, 15 libraries, and more than a dozen learned societies and research institutes. An association for the protection of sites and old buildings has also preserved over 60 archaeological and historical sites.

Cultural centres such as the Goethe Institute, the John F. Kennedy Center, the Arab Cultural Club, the Lebanese Cenacle, the Centre Culturel, and others have been instrumental in generating a lively awareness of arts and letters. Many art galleries in Beirut hold weekly exhibits of painting, sculpture, and photography. International opera, ballet, symphony, and drama companies compete to enrich the city's cultural life. Local artists have shown a refreshing readiness to experiment with new, expressive forms, from surrealistic painting to theatre of the absurd. There has also been a revival of national folk arts, particularly of song, dance, poetry, and traditional crafts.

The mass media

Beirut is also the locus of an extensive and well-developed mass media and communication network. The country's various ethnic groups, together with the freedom of expression enjoyed by all elements, have encouraged a proliferation of newspapers and periodicals. In addition to the wide variety of foreign publications, Beirut has a large number of daily and weekly papers.

There are some 45 cinemas, and the city's per capita film attendance has been estimated as the second highest in the world. Exposure to television and radio is also extensive. More than four out of five Beirutis own television sets, and virtually the entire population has access to transistor radios. The city's two private television companies operate four channels, transmitting for a total of 16 hours daily. The government-run broadcasting station (with Arabic, French, English, and Armenian programs) transmits more than 60 hours of programs a day.

Recreation. Despite recent efforts by municipal and voluntary agencies, the city continues to suffer from a critical shortage of public parks, gardens, and recreational open spaces. To some extent, however, this is compensated for by the accessibility of Mediterranean beaches and the nearby mountains.

Altogether, the 20 public gardens and municipal fields cover not more than 25 acres. An ambitious plan is being considered to transform the Beirut pine forest (200 acres) into a major recreational ground with an open-air theatre and facilities for international expositions. In the meantime, those with leisure are able to enjoy the rare scenic delights of the coast or to picnic in the shade of the pine or olive groves on the outskirts of the city's southern fringe.

BIBLIOGRAPHY. SAID CHEHABE ED-DINE, *Géographie humaine de Beyrouth* (1960), though slightly outdated, is the only study that surveys the human geography of Beirut and accounts for its growth patterns. C.W. CHURCHILL, *The City of Beirut* (1954), is a descriptive, socioeconomic survey that provides detailed information on household composition, education, mobility, occupations, housing, income saving, and expenditure. The BEIRUT, EXECUTIVE BOARD OF MAJOR PROJECTS, *Comprehensive Plan Studies for the City of Beirut* (1968), a preliminary but comprehensive survey report, provides detailed information on the physical features, land use patterns, utilities, population, and economic characteristics of the city. See also S. KHALAF and P. KONGSTAD, *Hamra of Beirut: A Case of Rapid Urbanization* (in prep.), an empirical study that explores the ecological transformation and the social structure of one of Beirut's urban communities; and HARVEY PORTER, *The History of Beirut* (1912), a brief but instructive historical sketch of the city from the earliest times to the beginning of the 20th century.

(S.G.K.)

Belgium

One of the smallest and most densely populated nations of Europe, the Kingdom of Belgium has been, since its birth in 1830 and 1831, a hereditary, representative, and constitutional monarchy. Its 11,782 square miles (30,515 square kilometres) had a population of more than 9,700,000 by the early 1970s. The country is bounded on the northwest by the North Sea, on the north and northeast by The Netherlands, on the east by the Federal Republic of Germany (West Germany), on the southeast by the Grand Duchy of Luxembourg, and on the west and southwest by France.

Culturally, Belgium is a heterogeneous nation, a bringing together without union of the ancient Latin and Germanic heritages of western Europe. Aside from a small German-speaking population in the southeast, the nation is divided geographically between the Walloons, a French-speaking people in the four southwestern provinces (Hainaut, Namur, Liège, and Luxembourg), and the Flemings, a Dutch-speaking people in the four northern and northeastern provinces (West-Vlaanderen, Oost-Vlaanderen [West and East Flanders], Antwerp, and Limburg). The central province, Brabant, is itself split, and in it lies the officially bilingual capital, Brussels. This division has had momentous consequences in Belgium's social and economic history, and in the early 1970s it continued to be an issue of major proportions in the national life.

Internal disharmonies aside, Belgium and the political entities that preceded it have been rich with historical and cultural associations, from the Gothic grandeur of its medieval university cities and its small castle-dominated towns on steep-bluffed winding rivers, through its broad traditions in painting and music that marked one of the high points of the northern Renaissance in the 16th century, to its contributions to the arts of the 20th century and its maintenance of the folk cultures of past eras. The Belgian landscape has been a major battleground of Europe for centuries, notably in modern times during the Battle of Waterloo and World Wars I and II. Given its area and population, Belgium today is the most heavily industrialized nation in Europe. It is a member of

the Benelux customs union formed in 1944 with The Netherlands and Luxembourg, and a member of the European Common Market with West Germany, France, Italy, its Benelux allies, and the United Kingdom. (For information on related topics, see the articles LOW COUNTRIES, HISTORY OF; GHENT; BRUGES; BRUSSELS; ANTWERP; and EUROPE.)

THE NATURAL AND HUMAN LANDSCAPE

Topographical character

Belgium generally is a low-lying country, with a broad coastal plain extending from the North Sea and The Netherlands and rising gradually into the Ardennes hills and forests of the southeast, where a maximum height of 2,277 feet (694 metres) is reached at Bottrange.

Physical regions. The main physical regions are the Ardennes and Ardennes foothills; the Anglo-Belgian Basin to the north comprising the Central (Bas) Plateaus, the plain of Vlaanderen (Flanders) and the Kempenland (Campine); and the intrusion of the Paris Basin on the west and southwest known as Belgian Lorraine.

The Ardennes. The Ardennes region is part of the Hercynian Belt reaching from western Ireland into Germany and formed during the late Paleozoic Era well over 225,000,000 years ago. It is a plateau cut deeply by the Meuse River and its tributaries. Its higher points have poor drainage and are more favourable for peat bogs and upland mossy ground than for crops.

The Ardennes foothills. A large depression known east of the Meuse as the Famenne, and as the Fagne west of it, separates the Ardennes from the geologically and topographically complex foothills to the north. The principal part of the area is the Condroz, a plateau more than 980 feet in altitude comprising a succession of valleys hollowed out of the limestone between sandstone crests. Its northern boundary is the Sambre-Meuse Valley, which crosses Belgium from east to west.

Belgian Lorraine. Situated south of the Ardennes and cut off from the rest of the country, the Belgian Lorraine is a series of hills with north-facing scarps. Half of it remains wooded, and in the south is a small region of iron-ore deposits.

The Central Plateaus. A region of sand and clay soils lying between 160 and 650 feet in altitude, the Central Plateaus cover northern Hainaut, southern Brabant, and the Hesbaye plateau region of Liège. It is dissected by several rivers that enter the Schelde (Escaut) River; it is bounded on the north by a series of outlying hills. The Brussels region lies within the Central Plateaus.

The plain of Vlaanderen. Bordering the North Sea from France to the Schelde, the low-lying plain of Vlaanderen has two main sections. Maritime Vlaanderen, extending inland for five to ten miles, is a region of reclaimed land, or polders, protected by a line of dunes and dikes and having largely clay soils. Interior Vlaanderen comprises most of Oost- and West-Vlaanderen and has sand-silt or sand soils. At 70 to 160 feet in altitude, it is drained by the Leie, Schelde, and Dendre rivers flowing northeastward to the Schelde estuary. Several shipping canals interlace the landscape connecting the river systems.

The Kempenland Plateau. Covered by pasturelands and industry and lying between 160 and 330 feet in altitude, the Kempenland forms an irregular watershed between the Schelde and Meuse systems.

Climate. Belgium has a temperate, maritime climate predominantly influenced by air masses from the Atlantic. Rapid and frequent alternation of different air masses separated by fronts gives Belgium considerable variability in weather. Frontal conditions moving from the west produce rainy weather, with rainfall heavy and frequent, averaging around 30 to 40 inches (750 to 1,000 millimetres). Winters are damp and mild with frequent fogs; summers, rather cool. The annual mean temperature is around 50° F (10° C). Brussels, for example, roughly in the middle of the country, has a minimum of 36° F (2.2° C) in December and a maximum of over 62° F (16.6° C) in July.

Regional climatic variations

Regional climatic differences are determined by altitude and distance inland. Farther inland, maritime influences become weaker and the climate becomes more continental, characterized by greater seasonal extremes of temperature. The Ardennes region, the highest and farthest inland, is the coldest. In January, mean temperatures are lower, frost occurs on about 100 days, and snow falls on 30 to 35 days. In summer, the altitude counteracts the effect of distance inland, and July temperatures are the lowest in the country. Because of the topography, the region has the highest rainfall in Belgium. In contrast, the Vlaanderen region enjoys generally higher temperatures throughout the year. There are fewer than 60 days of frost, and fewer than 15 of snow. On the seacoast these figures are reduced to below 50 and ten, respectively. There are a few very hot days, especially on the coast, and the annual rainfall is the lowest in the country.

Vegetation and animal life. Belgium lies within the area of deciduous forestation. The dominant tree is the oak, and other trees include beech, birch, and elm. For about a century, however, coniferous plantings have modified the character of the forests. The two most wooded regions are the Kempenland, where planted coniferous forests predominate, and the Ardennes, with deciduous and coniferous forests. The animal population, greatly reduced by man's activities, is Eurasian. Wild boar, wildcats, and deer roam the forests of the Ardennes.

Patterns of human settlement. The biological resources of the several natural regions and the consequent variations in land use have been major factors in determining patterns of rural settlement. The nature of the urban developments imposed on this landscape is derived mainly from the patterns of mining, manufacturing, commerce, and related enterprise throughout the country.

Rural Belgium. Forest and grassland dominate the landscape south of the Sambre-Meuse Valley: meadows, with orchards, monopolize the land in the Fagne and in the Plateau de Herve, while forest occupies two-thirds of the area along both edges of the Ardennes and in the heart of Lorraine. Forage crops, oats, potatoes, and even wheat are grown everywhere, but especially in Lorraine and in the Condroz. The region is one of striking contrasts: in the Condroz, farms often exceed 125 acres, whereas in the Ardennes half are less than 25 acres. The population is sparse in the Plateau de Herve in the east and western Entre-Sambre-et-Meuse in the southwest, but it is more concentrated throughout the remainder of the region.

Man's uses of the land

The open countryside of Hainaut, Brabant, and Hesbaye is given over to intensive diversified agriculture devoted to pasture, wheat, sugar beet, and oats; local variations include orchards in northern Hesbaye. Settlement is thick and farms, with their closed courts, sometimes exceed 250 acres in area.

The open landscape of maritime Vlaanderen and the lower Schelde, intersected by dikes and canals, is dotted with scattered farms, their buildings loosely grouped around a court. Over one-third of the area of farms between 50 and 125 acres is under pasture, while the remainder is cultivated, with wheat and sugar beet the dominant crops. Inland Vlaanderen is a region of scattered habitation and large market towns, its widespread scrub undergrowth giving it a wooded appearance. It is devoted to grazing. Grasslands are sparse. Intensive cultivation is confined to gardens and small farms, one-third of them under ten acres. Oats, rye, and potatoes are the chief crops, and wheat and such varied industrial crops as sugar beets, chicory, hops, and flax are found in the southwest.

Coniferous forests and grasses dominate the Kempenland countryside, and oats, rye, and potatoes are also important. Farms are small and most of the population lives in villages. Coal seams lie beneath portions of the region.

Urban Belgium. In the Walloon coalfields, roughly in and to the north of the Meuse Valley across south central Belgium, coal mining, glass manufacture, iron production, zinc metallurgy, and the chemical and electrical industries in the 19th and 20th centuries gave rise to heavily populated areas with widely varying urban characteristics. Liège, the regional capital since the Middle

The major cities of Belgium

Ages, is a great administrative and commercial city dominating and coordinating the chief metropolitan area of the Walloon region. Charleroi, the heart of a large metropolitan area, is a newer and smaller city, and is predominantly commercial. La Louvière, founded during the 19th-century industrial development, is a developing metropolitan centre. The Borinage, an area of high population density without a central city, comes under the influence of Mons.

The ancient city of Antwerp and its metropolitan area, the second largest in the nation, extend along the east bank of the Schelde. The port is formed by the base of the estuary and the concave riverbank. The existence of the port has favoured the establishment of important and diverse industries: petroleum refining, chemical and metallurgical industries, food processing, and electronics manufacturing.

Gent (Ghent), a historic university town, is Belgium's second largest port. A centre of the textile industry, Gent is experiencing an industrial regeneration, characterized especially by heavy steel production along the Gent-Terneuzen Canal, connecting the port to the Schelde.

Brugge (in French Bruges, "bridge"), the capital of West Vlaanderen, is a city of medieval aspect, resplendent with cathedrals and ancient homes, often filled with art of great value. As its name implies, the city has many bridges spanning the several canals and the canalized Rei River. Mentioned as early as the 7th century, it became a trading centre for the Hanseatic League and reached its greatest period during the 15th century, when the dukes of Burgundy held court there.

Leuven (in French, Louvain), 16 miles west of Brussels, is the site of the first university founded (1425) in the Low Countries. The institution was damaged severely

MAP INDEX

Political subdivisions

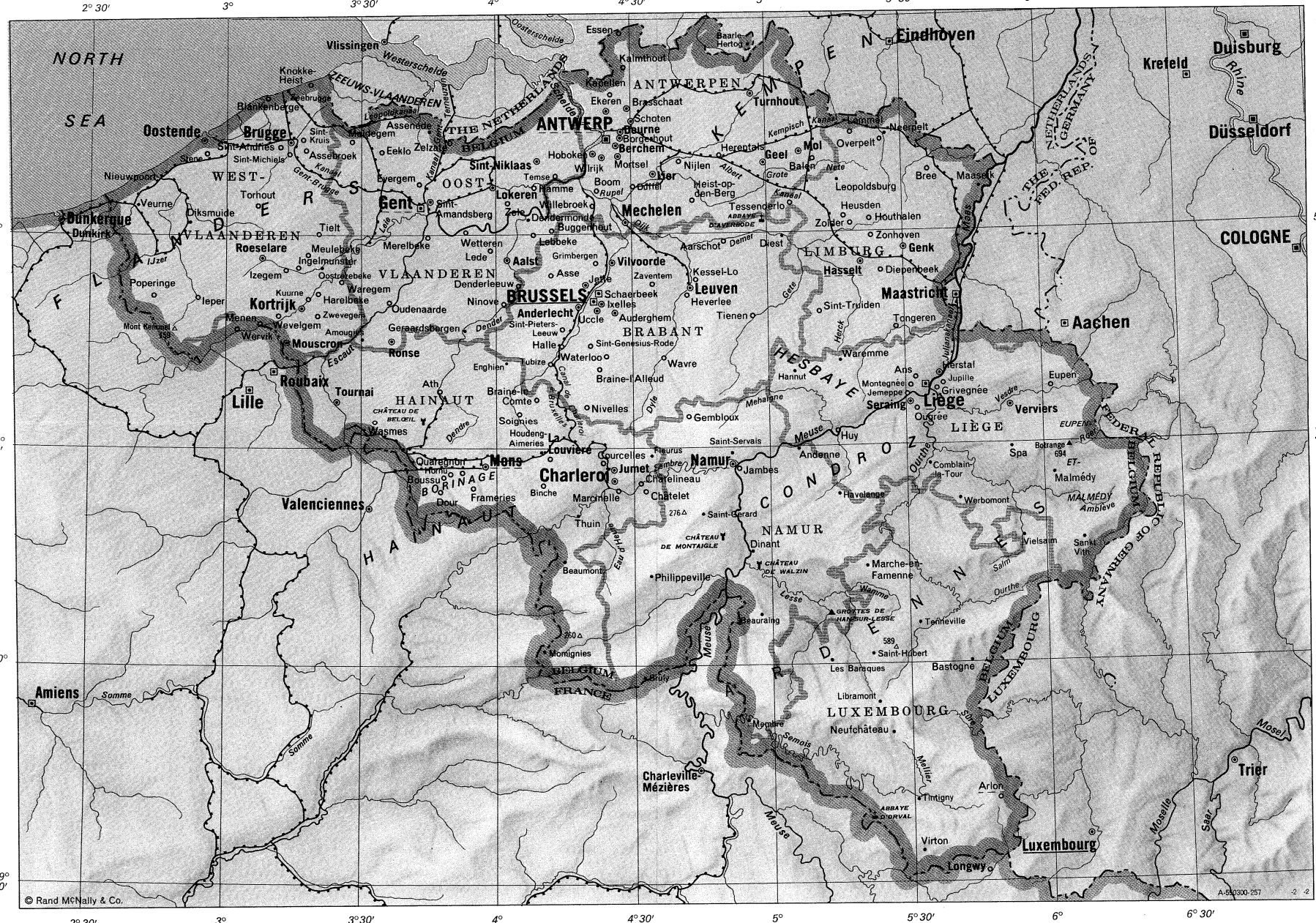
Antwerpen.....51·10n 4·50e
Brabant.....50·45n 4·30e
East Flanders, see Oost-Vlaanderen
Hainaut.....50·30n 3·50e
Liège.....50·30n 5·30e
Limburg.....51·00n 5·30e
Luxembourg.....50·00n 5·30e
Namur.....50·20n 4·50e
Oost-Vlaanderen..51·00n 3·45e
West Flanders, see West-Vlaanderen
West-Vlaanderen..51·00n 2·43e

Cities and towns

Aalst.....50·56n 4·02e
Aarschot.....50·59n 4·50e
Amougies.....50·45n 3·29e
Andenne.....50·29n 5·06e
Anderlecht.....50·50n 4·18e
Ans.....50·39n 5·32e
Antwerp.....51·13n 4·25e
Anvers, see Antwerp
Arlon.....49·41n 5·49e
Asse.....50·55n 4·12e
Assebroek.....51·12n 3·16e
Assenede.....51·14n 3·45e
Ath.....50·38n 3·47e
Auderghem.....50·49n 4·26e
Baarle-Hertog.....51·27n 4·56e
Balen.....51·10n 5·09e
Bastogne.....50·00n 5·43e
Beaumont.....50·14n 4·14e
Beauraing.....50·07n 4·58e
Berchem.....51·12n 4·26e
Bergen, see Mons
Binche.....50·24n 4·10e
Blankenberge.....51·19n 3·08e
Boom.....51·05n 4·22e
Borgerhout.....51·13n 4·26e
Boussu.....50·26n 3·48e
Braine-l'Alleud....50·41n 4·22e
Braine-le-Comte..50·36n 4·08e
Brasschaat.....51·17n 4·27e
Bree.....51·08n 5·36e
Bruges, see Brugge
Brugge.....51·13n 3·14e
Brûly.....49·58n 4·31e
Brussels.....50·50n 4·20e
Bruxelles, see Brussels
Buggenhout.....51·01n 4·12e
Charleroi.....50·25n 4·26e
Châtelet.....50·24n 4·31e
Châtelineau.....50·25n 4·31e
Comblain-la-Tour.....50·27n 5·34e
Courcelles.....50·28n 4·22e
Courtrai, see Kortrijk
Denderleeuw.....50·53n 4·04e
Dendermonde....51·02n 4·07e
Deurne.....51·13n 4·28e
Diepenbeek.....50·54n 5·24e
Diest.....50·59n 5·03e
Diksmuide.....51·02n 2·52e
Dinant.....50·16n 4·55e
Dixmude, see Diksmuide
Dour.....50·24n 3·47e
Duffel.....51·06n 4·31e
Eeklo.....51·11n 3·34e
Ekeren.....51·17n 4·25e
Enghien.....50·42n 4·02e
Essen.....51·28n 4·28e
Eupen.....50·38n 6·02e
Evergem.....51·07n 3·42e
Fleurus.....50·29n 4·33e
Frameries.....50·24n 3·54e
Gand, see Gent
Geel.....51·10n 5·00e
Gembloux.....50·34n 4·41e
Genk.....50·58n 5·30e
Gent.....51·03n 3·43e
Geraardsbergen..50·46n 3·52e
Ghent, see Gent
Grimbergen.....50·56n 4·23e
Grivegnée.....50·37n 5·36e
Halle.....50·44n 4·13e
Hamme.....51·06n 4·08e
Hannut.....50·40n 5·05e
Harelbeke.....50·51n 3·18e
Hasselt.....50·56n 5·20e
Havelange.....50·23n 5·14e
Heist-op-den-Berg.....51·05n 4·43e
Herentals.....51·11n 4·50e
Herstal.....50·40n 5·38e
Heusden.....51·02n 5·16e
Heverlee.....50·52n 4·42e
Hoboken.....51·10n 4·21e
Hornu.....50·26n 3·49e
Houdeng-Aimeries.....50·29n 4·08e
Houthaien.....51·02n 5·22e
Huy.....50·31n 5·14e
Ieper.....50·51n 2·53e
Ingelmunster.....50·55n 3·15e
Ixelles.....50·50n 4·22e
Izegem.....50·55n 3·12e
Jambes.....50·28n 4·52e
Jemeppe.....50·37n 5·30e
Jette.....50·52n 4·20e
Jumet.....50·26n 4·25e
Jupille.....50·39n 5·38e
Kalmthout.....51·23n 4·28e
Kapellen.....51·19n 4·26e
Kessel-Lo.....50·54n 4·45e
Knokke-Heist.....51·21n 3·17e
Kortrijk.....50·50n 3·16e
Kuurne.....50·51n 3·17e
La Louvière.....50·28n 4·11e
Lebbeke.....51·00n 4·08e
Lede.....50·58n 3·58e
Leopoldsburg.....51·07n 5·15e
Les Baraques.....50·01n 5·13e
Leuven (Louvain)..50·53n 4·42e
Libramont.....49·55n 5·23e
Liège.....50·38n 5·34e
Lier.....51·08n 4·34e
Lokeren.....51·06n 4·00e
Lommel.....51·14n 5·18e
Louvain, see Leuven
Luik, see Liège
Maaseik.....51·06n 5·48e
Maldegem.....51·13n 3·27e
Malines, see Mechelen
Malmédy.....50·25n 6·02e
Marche-en-Famenne.....50·12n 5·20e
Mechelen.....51·02n 4·28e
Marcinelle.....50·24n 4·26e
Membre.....49·52n 4·54e
Menen.....50·48n 3·07e
Merelbeke.....51·00n 3·45e
Meulebeke.....50·57n 3·17e
Mol.....51·11n 5·06e
Momignies.....50·02n 4·10e
Mons.....50·27n 3·56e
Montegnée.....50·39n 5·31e
Mortsel.....51·10n 4·28e
Mousoron.....50·44n 3·13e
Namur (Namen)..50·28n 4·52e
Neerpelt.....51·13n 5·25e
Neufchâteau.....49·50n 5·26e
Nieuwpoort.....51·08n 2·45e
Nijlen.....51·10n 4·39e
Ninove.....50·50n 4·01e
Nivelles.....50·36n 4·20e
Oostende.....51·13n 2·55e
Oostrozebeke....50·55n 3·20e
Ostend, see Oostende
Oudenaarde.....50·51n 3·36e
Ougrée.....50·36n 5·32e
Overpelt.....51·13n 5·25e
Philippeville.....50·12n 4·32e
Poperinge.....50·51n 2·43e
Quaregnon.....50·26n 3·51e
Roeselare.....50·57n 3·08e
Ronse.....50·45n 3·36e
Saint, see also under Sint, Sankt
Saint-Gérard.....50·21n 4·45e
Saint-Hubert.....50·01n 5·23e
Saint-Servais.....50·28n 4·50e
Sankt Vith.....50·17n 6·08e
Schaerbeek.....50·51n 4·23e
Schoten.....51·15n 4·30e
Seraing.....50·36n 5·29e
Sint-Amandsberg.51·04n 3·45e
Sint-Andries.....51·12n 3·10e
Sint-Genesius-Rode.....50·45n 4·21e
Sint-Kruis.....51·13n 3·15e
Sint-Michiels....51·11n 3·12e
Sint-Niklaas.....51·10n 4·08e
Sint-Pieters-Leeuw.....50·47n 4·14e
Sint-Truiden.....50·48n 5·12e
Soignies.....50·35n 4·04e
Spa.....50·30n 5·52e
Stene.....51·12n 2·55e
Temse.....51·08n 4·13e
Tenneville.....50·06n 5·32e
Tessenderlo.....51·04n 5·05e
Thuin.....50·20n 4·17e
Tielt.....51·00n 3·19e
Tienen.....50·48n 4·57e
Tintigny.....49·41n 5·31e
Tongeren.....50·47n 5·28e
Tournai.....50·36n 3·23e
Torhout.....51·04n 3·06e
Tubize.....50·41n 4·12e
Turnhout.....51·19n 4·57e
Uccle.....50·48n 4·19e
Verviers.....50·35n 5·52e
Veurne.....51·04n 2·40e
Vielsalm.....50·17n 5·55e
Vilvoorde.....50·56n 4·26e
Virton.....49·34n 5·32e
Waregem.....50·53n 3·25e
Waremme.....50·41n 5·15e
Wasmes.....50·33n 3·32e
Waterloo.....50·43n 4·23e
Wavre.....50·43n 4·37e
Werbomont.....50·23n 5·41e
Wervik.....50·47n 3·02e
Wetteren.....51·00n 3·53e
Wevelgem.....50·48n 3·10e
Willebroek.....51·04n 4·22e
Wilrijk.....51·10n 4·24e
Ypres, see Ieper
Zaventem.....50·53n 4·28e
Zeebrugge.....51·20n 3·12e
Zele.....51·04n 4·02e
Zelzate.....51·12n 3·49e
Zinnik, see Soignies
Zolder.....51·01n 5·18e
Zonhoven.....50·59n 5·21e
Zwevegem.....50·48n 3·20e

Physical features and points of interest

Albert Kanaal, *canal*.....50·39n 5·37e
Amblève, *river*....50·28n 5·36e
Averbode, Abbaye d', *abbey*.51·02n 4·59e
Ardennes, *physical region*...50·10n 5·45e
Belœil, Château de, *palace*.....50·35n 3·44e
Borinage, *historic region*....50·25n 3·50e
Botrange, *mountain*.....50·30n 6·08e
Charleroi à Bruxelles, Canal de.....50·51n 4·19e
Condroz, *historic region*....50·20n 5·00e
Demer, *river*.....50·58n 4·42e
Dender (Dendre), *river*.....51·02n 4·06e
Dijk, *river*.....51·04n 4·25e
Dyle, *river*.....51·04n 4·25e
Eau d'Heure, *river*.....50·18n 4·24e
Escaut (Schelde), *river*.....51·20n 4·15e
Eupen-et-Malmédy, *historic region*....50·30n 6·05e
Flanders, *historic region*....51·00n 3·00e
Gent-Brugge, Kanaal, *canal*....51·03n 3·43e
Gent-Terneuzen Kanaal, *canal*.....51·04n 3·44e
Gete, *river*.....50·57n 5·07e
Grote Nete, *river*..51·07n 4·34e
Hainaut, *historic region*....50·30n 3·50e
Han-sur-Lesse, Grottes de, *cave*..50·08n 5·15e
Herck, *river*.....50·58n 5·07e
Hesbaye, *historic region*....50·35n 5·15e
IJzer, *river*.....51·09n 2·43e
Julianakanaal, *canal*.....51·05n 5·50e
Kemmel, Mont, *hill*.....50·47n 2·50e
Kempen, *historic region*....51·10n 5·20e
Kempisch Kanaal, *canal*....51·10n 4·49e
Leie, *river*.....51·03n 3·43e
Leopoldkanaal, *canal*.....51·14n 3·46e
Lesse, *river*.....50·14n 4·54e
Maas (Meuse), *river*.....51·10n 5·52e
Mehaigne, *river*...50·32n 5·13e
Mellier, *river*.....49·43n 5·32e
Meuse (Maas), *river*.....51·10n 5·52e
Montaigle, Château de, *castle*.....50·18n 4·49e
North Sea.....51·20n 2·30e
Orval, Abbaye d', *abbey*.....49·38n 5·22e
Ourthe, *river*.....50·38n 5·35e
Roer, *river*.....50·32n 6·13e
Rupel, *river*.....51·07n 4·19e
Sambre, *river*.....50·28n 4·52e
Salm, *river*.....50·22n 5·52e
Schelde (Escaut), *river*.....51·20n 4·15e
Semois, *river*.....49·52n 4·52e
Sûre, *river*.....49·50n 5·45e
Vesdre, *river*.....50·37n 5·37e
Walzin, Château de, *castle*.....50·13n 4·55e
Wamme, *river*.....50·10n 5·16e


2° 30′
3°
3° 30′
4°
4° 30′
5°
5° 30′
6°
6° 30′
7°
51°
50° 30′
50°
49° 30′
NORTH SEA
Vlissingen
Oostende
Brugge
Knokke-Heist
Blankenberge
Zeebrugge
Sint-Andries
Sint-Michiels
Assebroek
Sint-Kruis
Oostkerke
Assenede
Eeklo
Zelzate
Nieuwpoort
Veurne
Diksmuide
Dunkerque
Dunkirk
Torhout
Tielt
Roeselare
Meulebeke
Ingelmunster
Izegem
Poperinge
Ieper
Kortrijk
Harelbeke
Waregem
Oudenaarde
Menen
Wevelgem
Zwevegem
Wervik
Mouscron
Roubaix
Lille
Tournai
Ronse
Geraardsbergen
Amougies
Mont Kemmel 159
Gent
Evergem
Sint-Niklaas
Temse
Lokeren
Zele
Wetteren
Lede
Aalst
Denderleeuw
Ninove
Dendermonde
Buggenhout
Lebbeke
Grimbergen
Asse
Merelbeke
Wilrijk
Hoboken
Mortsel
Boom
Rupel
Willebroek
Mechelen
Aarschot
Diest
Zaventem
Kessel-Lo
Leuven
Heverlee
Tienen
Jette
Schaerbeek
Ixelles
Uccle
Auderghem
Anderlecht
BRUSSELS
Halle
Sint-Genesius-Rode
Waterloo
Wavre
Braine-l'Alleud
Enghien
Tubize
Braine-le-Comte
Soignies
Nivelles
Gembloux
Ath
Leuze
Dendre
Dyle
Senne
Escaut
Lys
IJzer
Yser
Meuse
Sambre
Ourthe
Amblève
Vesdre
Hesbaye
Mehaigne
Geete
Herck
Albert Canal
Kempisch Kanaal
Zuid-Willemsvaart
Leopoldkanaal
Gent-Brugge Kanaal
Gent-Terneuzen Kanaal
Schelde
Oosterschelde
Westerschelde
Rupel
Dijle
Demer
Grote Nete
Kleine Nete
Essen
Kalmthout
Kapellen
Brasschaat
Ekeren
Schoten
Borgerhout
Berchem
ANTWERP
Baarle-Hertog
Turnhout
Herentals
Geel
Mol
Balen
Overpelt
Neerpelt
Lommel
Eindhoven
Krefeld
Duisburg
Rhine
Düsseldorf
COLOGNE
Bree
Maaseik
Leopoldsburg
Heusden
Houthalen
Zolder
Zonhoven
Genk
Hasselt
Diepenbeek
Maastricht
Aachen
Sint-Truiden
Tongeren
Heist-op-den-Berg
Tessenderlo
Beringen
Lommel
Nijlen
Bornem
Puurs
Willebroek
Boom
Mechelen
Dijle
Demer
Aarschot
Diest
Zaventem
Kessel-Lo
Leuven
Heverlee
Tienen
Jette
Schaerbeek
Ixelles
Uccle
Auderghem
Anderlecht
BRUSSELS
Halle
Sint-Genesius-Rode
Waterloo
Wavre
Braine-l'Alleud
Enghien
Tubize
Braine-le-Comte
Soignies
Nivelles
Gembloux
Ath
Leuze
Dendre
Dyle
Senne
Escaut
Lys
IJzer
Yser
Meuse
Sambre
Ourthe
Amblève
Vesdre
Hesbaye
Mehaigne
Geete
Herck
Albert Canal
Kempisch Kanaal
Zuid-Willemsvaart
Leopoldkanaal
Gent-Brugge Kanaal
Gent-Terneuzen Kanaal
Schelde
Oosterschelde
Westerschelde
Rupel
Dijle
Demer
Grote Nete
Kleine Nete
Essen
Kalmthout
Kapellen
Brasschaat
Ekeren
Schoten
Borgerhout
Berchem
ANTWERP
Baarle-Hertog
Turnhout
Herentals
Geel
Mol
Balen
Overpelt
Neerpelt
Lommel
Eindhoven
Krefeld
Duisburg
Rhine
Düsseldorf
COLOGNE
Bree
Maaseik
Leopoldsburg
Heusden
Houthalen
Zolder
Zonhoven
Genk
Hasselt
Diepenbeek
Maastricht
Aachen
Sint-Truiden
Tongeren
Heist-op-den-Berg
Tessenderlo
Beringen
Lommel
Nijlen
Bornem
Puurs
Willebroek
Boom
Mechelen

during both world wars, but it was rebuilt and its libraries restocked by aid from many nations.

Brussels, the capital and the only metropolitan area with a population exceeding 1,000,000, has suburbs that spread over more than half of Brabant. It is the centre of commerce, industry, and intellectual life in Belgium.

THE BELGIAN PEOPLE

Linguistic groups. The population of Belgium is divided into three linguistic communities. The Dutch-speaking Flemings in the north comprise about 56 percent of the population and are increasing relative to the French-speaking Walloons in the south and west. The German-language region, containing less than 1 percent of the Belgians, in eastern Liège Province consists of 25 communes around Eupen and St. Vith. Brussels comprises 19 bilingual and seven Flemish communes. The French-speaking population is by far the larger, however, in the capital region.

Consequences of linguistic diversity

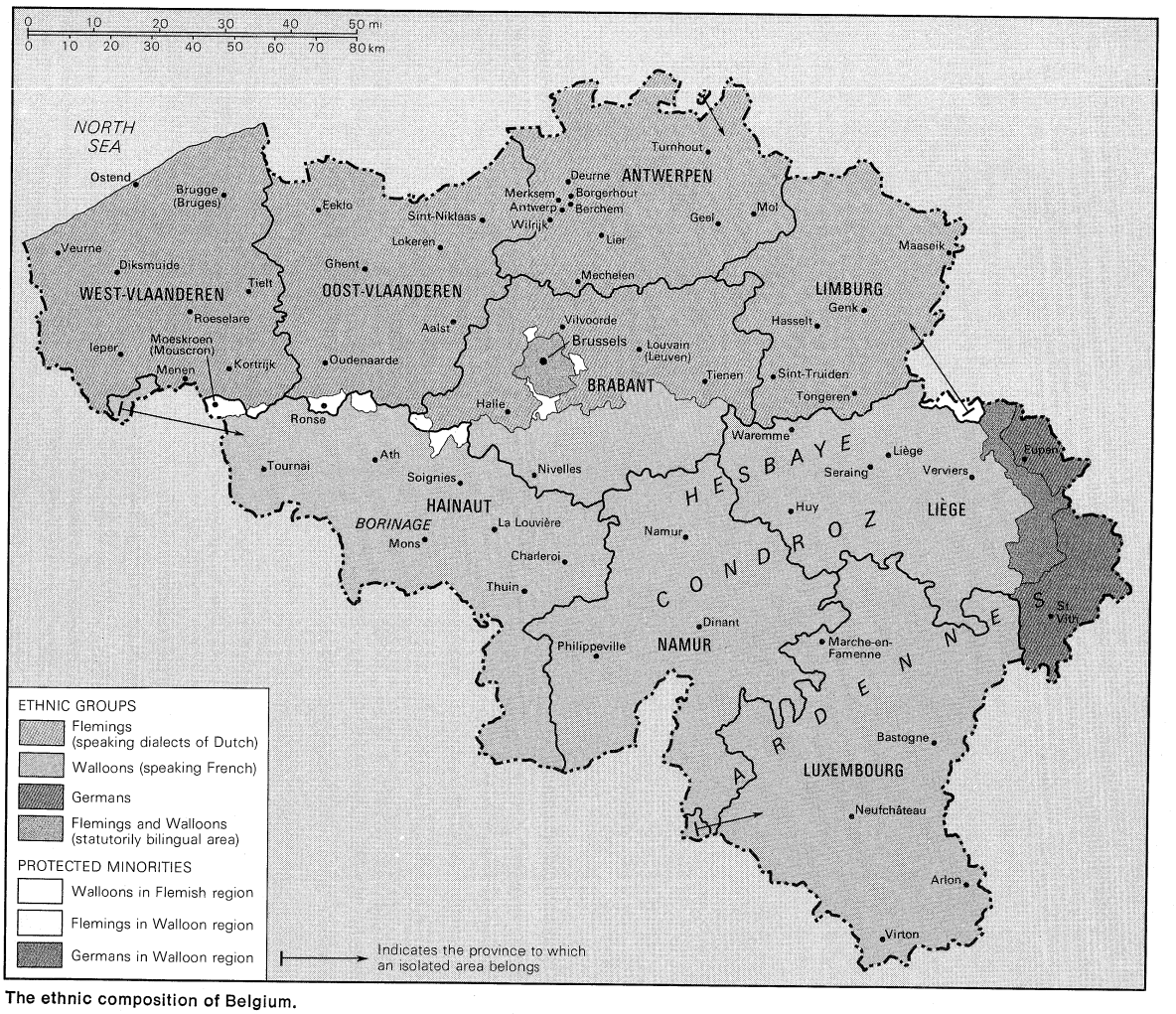
Historically, Belgium's linguistic division has created a social and economic fabric in which the managerial, professional, and administrative ranks were filled almost entirely by the French-speaking segment of the population, which at one time was also numerically superior. The Flemings long protested against what they felt was the exclusion of the average, non-bilingual Fleming from effective participation even in everyday dealings concerning law, medicine, and industrial employment. Along with their gradually increasing numerical strength, the Flemings have made gains in many areas, and since 1898 Belgium has been officially a bilingual nation. In the early 1970s many disputes and much rancour remained, however, and the future direction of the controversy was uncertain.

Religious groups. The majority of Belgians are Roman Catholic, but regular attendance at religious services is variable. Although it is marked in the Flemish region, regular attendance at church has decreased in the Walloon industrial region and in Brussels. The Protestant Churches number around 75,000 adherents, who live mostly in Hainaut, particularly in the Borinage, and in Brabant. The Jewish population of about 40,000 is concentrated in the general areas of Brussels and Antwerp.

Demography. *Population growth.* The annual growth rate of the Belgian population is very low, for overall birth rates and immigration exceed death rates and emigration only slightly. Marked differences exist between the Dutch-language region and the remainder of the country. Its natural increase—*i.e.*, preponderance of births over deaths—and net immigration give Flemish Belgium an annual growth rate of 3.2 per 1,000 inhabitants. In contrast, the French- and German-language regions together have a negative rate of natural increase that would reduce the population progressively were it not for a very small net immigration that raises these regions' total rate of population increase to 0.93 per 1,000. Brussels, moreover, has both a negative rate of natural increase and a net emigration that reduce its population by 1.8 persons per 1,000 annually. The nation's relatively low birth rates and high death rates are phenomena reflecting one of the most aged populations in the world: more than 19 percent are 60 years of age and over.

Linguistic regions and demographic phenomena

Migration and distribution of population. Since World War II, the foreign-born population has increased continuously at a rate higher than that of Belgian nationals: in addition to its continuing inflow, it has also had a high birth rate. By 1970 the foreign-born and their children represented more than 7 percent of the inhabitants. The largest concentrations of foreigners are found in the Walloon mining and industrial areas; foreigners, largely of



The ethnic composition of Belgium.

Mediterranean origin, with a preponderance of Italians, account for 16 to 19 percent of the population of the Liège region basin and in the Hainaut regions of Charleroi, La Louvière, and Mons. Between 1961 and 1970, the foreign population of Brussels more than doubled, from about 69,000 to about 170,000, one-half of them Spaniards and Italians, about 15 percent of the capital's inhabitants.

Internal migrations reveal the attraction that Brussels and its environs exert, particularly over Hainaut. Between 1964 and 1968, more than 13,000 people left Hainaut for the Brussels area. In the same period some 33,000 persons emigrated from Brussels to the outlying Flemish areas and Walloon northern Brabant, accounting for the net emigration from Brussels and resulting in a geographic expansion of the metropolitan area.

The Antwerp metropolitan area exerts an attraction over its own province and over the eastern part of Oost-Vlaanderen. Liège and environs extends an influence over its own province and the Ardennes. The Ardennes plateau is a rural area of continuous emigration, as are also central and western Vlaanderen.

Most of Flemish Belgium is characterized by population densities greater than 250 persons per square mile, although in the reclaimed polder region behind the seacoast densities generally are about 130. In Antwerp and Ghent, densities exceed 2,500, often even 7,500. The towns and some of the summer resorts on the coast, with densities of more than 2,500, constitute other high-density Flemish areas.

Belgium, Area and Population

	area		population*	
	sq mi	sq km	1961 census	1971 estimate
Provinces				
Antwerpen	1,104	2,861	1,443,000	1,536,000
Brabant	1,302	3,372	2,009,000	2,178,000
Hainaut	1,463	3,790	1,317,000	1,331,000
Liège	1,497	3,876	992,000	1,015,000
Limburg	935	2,422	572,000	656,000
Luxembourg	1,706	4,418	217,000	219,000
Namur	1,413	3,660	369,000	385,000
Oost-Vlaanderen	1,151	2,982	1,272,000	1,314,000
West-Vlaanderen	1,210	3,134	998,000	1,057,000
Total Belgium	11,782	30,515	9,190,000†	9,691,000†

*De jure population. †Figures do not add to total given because of rounding.
Source: Official government figures.

The Walloon region, in contrast, is on the whole an area in which densities fall below 250. To the north of the centrally located coalfield—in the low plateaus of rural Hainaut, Brabant, and Hesbaye—densities are greater than 130, whereas to the south, in the Ardennes and the Condroz, densities usually drop below this figure. In the industrial–mining region, Liège and Charleroi have densities comparable to those of Antwerp and Ghent. Such other urban Walloon areas as Mons and the Borinage, La Louvière and the Centre, Namur and the Basse-Sambre, and Verviers have densities above 2,500. Brussels remains the most densely populated area of the country. In the early 1970s, its density was over 22,000, and certain communes had densities of from 25,000 to 50,000.

Trends. According to the forecasts of the National Institute of Statistics (Institut National de Statistique), the Belgian population should increase to nearly 10,200,000 by 1980. The projections also indicate a slight age reduction in the population. Those 60 and over would fall to 17.3 percent of the population, while those under 20 would rise to nearly 32 percent.

THE NATION'S ECONOMY

Less than 5 percent of Belgium's active population is engaged in agriculture, suggesting the great role of industry and commerce in the national economy. Drastic monetary reform after World War II aided postwar recovery and expansion, and Belgium was one of the first European countries to re-establish a favourable balance of trade. Although from 1958 to 1968 the average annual rate of growth of its gross national product (GNP) was slightly lower than that of the European Economic Community (EEC), or Common Market, as a whole, Belgium rates slightly above the EEC average in standard-of-living criteria except number of automobiles and television sets. National prosperity is dependent mainly on Belgium's role as a fabricator and processor of imported raw materials and the subsequent export of finished goods.

General picture of the economy

With an economically active population of over 3,900,000 persons, Belgium's employment rate is slightly lower than the Common Market average. Services employ over half of the active population, one of the highest rates in western Europe.

A geographic analysis of the gross domestic product (GDP) by province reveals that the highest levels of GDP per capita are in the Brabant and Antwerp areas. The analysis of the GDP according to economic activities confirms the importance of services and the weakness of agriculture. The provinces of Luxembourg, Namur, and West-Vlaanderen have the most highly developed agriculture, whereas industry is the major economic activity of Hainaut, Limburg, Liège, and Oost-Vlaanderen. The Antwerp region dominates in transportation.

Resources and economic activity. *Mining.* Coal long has been Belgium's only important mineral resource, but the exploitation of petroleum deposits under the North Sea is under way. The iron-ore deposits of the Sambre-Meuse belt, in decline since the latter part of the 19th century, have now been worked out, and only small deposits of nonferrous ores exist.

The Borinage coal-mining area, in the Sambre-Meuse Valley, lies in a narrow band across the centre of Belgium from the French border through Mons, Charleroi, Namur, and Liège. Although many mines in the Kempenland field are worked out or abandoned as uneconomic, this second region provides most of the coking and slow-burning coal for domestic industry. Difficult mining conditions add to costs, and the yield per miner is the lowest of any Common Market country. Although the industry operates as a private enterprise, it is a considerable drain on public finances, for the government subsidizes the industry and assists new industries in areas hit by mine closings.

Agriculture, fishing, and forestry. Agricultural production accounts for less than 5 percent of Belgium's GNP and continues to decline in importance. The principal crops are cereal grains, potatoes, and sugar beets, with livestock the major agricultural export. Horticulture is being developed slowly. Ostend, the main fishing port, sends a modest fleet of trawlers to the North Sea fishing grounds. Reforestation is extensive, and the lumbering industry, aided by mechanization, has increased its output. A special forestry fund helps to ameliorate social conditions among forestry workers.

Manufacturing. The manufacturing sector accounts for nearly one-third of national income. The steel, metallurgical, and chemical sectors are the fastest growing in the Belgian industry. The steel industry produces more than 10 percent of the Common Market output, and during the 1960s the exports of the metal manufactures increased by 400 percent, while the chemical industry expanded its production capacity considerably. Belgium leads the world in the processing of cobalt, germanium, and radium. Refineries, located principally in the Antwerp area, process more than 20,000,000 tons of crude petroleum a year. The lace for which Belgium has been known around the world long was part of a slowly dying industry that depended heavily on the handwork of elderly women. The authorities and individuals in Brugge, the centre of the industry, gave new impetus to the tradition, however, establishing specialized schools for training younger workers and arranging expositions in various countries.

The steel, chemical, and metallurgical industries

Financial services. The economic importance of the financial sector has increased significantly in recent years; it accounts for about 4 percent of the GNP. The General Savings and Pensions Fund is the major financial organi-

zation for the collection of savings, while the Société Générale de Belgique dominates the banking sector. The National Bank is charged with issuing currency, and has considerable autonomy.

Foreign trade. Trade, representing over 40 percent of the GNP, is more important than in any other economy in the world. The main imports are raw materials and food products. The greatest increases in export rates have occurred in the metallurgical, automobile, and chemical industries. The principal trade partners of Belgium are the member nations of the Common Market, which account for almost 70 percent of Belgium's foreign trade.

Management of the economy. *The private and public sectors.* Free enterprise, which dominates the entire economy with the exception of the air and rail transportation sectors and broadcasting, faces scarcely any government competition. The rate of increase in public expenditure during the 1960s was the highest in the Common Market. At the beginning of the 1970s, public expenditure accounted for 14 percent of the GNP. Government activity is aimed primarily at maintaining overall economic equilibrium. Ordinary expenditures are covered by tax revenues, while such extraordinary expenditures as public-works projects are financed through loans. The level of public debt is high, and the interest on it is a significant item in the annual budget.

Taxation. Indirect taxes account for almost 60 percent of tax revenues, direct corporate taxes for less than 10 percent, and direct taxation of private citizens for over 30 percent.

In order to coordinate its fiscal policy with those of the other Common Market countries, Belgium in 1971 instituted a system of value-added taxation in which the tax is imposed at the point of manufacture and included in the price charged.

Labour and management. Some 2,700,000 workers are represented by three trade-union organizations: the General Federation of Labour, the Federation of Christian Trade Unions, and the Belgium General Federation of Liberal Trade Unions. The first two enjoy a virtual monopoly of trade-union membership, accounting for nearly 95 percent of the organized workers. The Federation of Belgian Industries and the Federation of Non-Industrial Enterprises of Belgium represent employers. During the 1960s, the nature of labour-management relations changed markedly. To the traditional demands for higher wages were added demands for improved working conditions and greater participation in the economic life of businesses.

Economic problems and policies. Since the end of World War II, government has intervened only indirectly in the economy, except under unusual circumstances. Two facts point to issues of concern. First, for many years Belgium's GNP has climbed more slowly than those of its European partners; second, the rate of population increase remains low. Further, since Belgium is very dependent on international trade, its future prospects will reflect the growth and terms of this trade.

Governmental intervention in the economy

In connection with these issues, the government seeks to encourage growth through the expansion of certain key industries. It offers contracts that guarantee certain public markets to private firms and requires, in return, that the firms carry into effect large investment programs. On the international level, Belgium has consistently advocated lowering customs tariffs, and, to help maintain a healthy balance of payments, government activity is always aimed at maintaining stability of prices, improving productivity, and diversifying markets.

Economic prospects appear favourable: once the conversion of the coal industry has been successfully completed, there should be no serious economic problems provided that the high level of investment in manufacturing is maintained.

Transportation. *Highways.* The extensive system of main roads in Belgium is being supplemented by the construction of numerous expressways. These modern roadways extend from Brussels to Ostend by way of Ghent and Brugge, and from Antwerp to Aachen, West Germany, by way of Hasselt and Liège. Other expressways under construction are those from Antwerp to Kortrijk by way of Ghent, from Brussels to the Ardennes by way of Namur, and from Brussels to Paris through Mons and Charleroi. A plan is afoot to transform the Brussels–Antwerp route into an expressway.

Railways. The railway network, a state enterprise, is denser than that of any other country, with a total length of about 7,000 miles. Brussels is the heart of the system, the centre of a series of lines that radiate outward and link the capital to other cities both inside and outside the country. At the beginning of the 1970s, the railways carried about 247,000,000 passengers and about 71,000,000 tons of freight annually. The heaviest traffic is between Brussels and Antwerp.

Waterways. More than 50,000,000 tons of freight were loaded annually at Belgian ports in the early 1970s, but over 66,000,000 tons were unloaded. Antwerp handles three-quarters of the total tonnage of Belgian ports. Other important ports were Brugge-Zeebrugge, Ostend, Ghent, and Brussels. Navigable inland waterways include the Meuse and Schelde, which are navigable throughout their length in Belgium. The Albert Canal, which can take barges up to 2,000 tons, links Antwerp with the Liège region. A canal from Charleroi to Brussels, carrying barges up to 1,350 tons, links the basins of the two main rivers through Ronquieres lock. A maritime canal connects Brugge and Zeebrugge, another connects Ghent and Terneuzen, on the Schelde estuary, and a third links Brussels and Rupel.

Airways. The Brussels airport is the centre of Belgian air traffic, while smaller facilities are maintained at Antwerp and Liège. SABENA, the national airline, was created in 1923; although 90 percent government owned, it is incorporated as a private corporation. It serves 51 nations on four continents and has domestic flights from Liège and Antwerp.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of national and local government. Under the terms of the Belgian constitution, ministers are appointed and dismissed by the king. Since the king's person is inviolable, his acts must be countersigned by a minister, who in turn becomes responsible for them to Parliament. Legislative power is shared by the king with the Chamber of Representatives and the Senate; the members of both these bodies are elected for four-year terms. Except for a few points that were clarified by the constitutional revision of 1970, the constitution is silent on the subject of governmental structure. The structure thus has been created by custom, and has varied through the course of time.

Ministerial and legislative bodies

In 1831 there were only five ministers, whereas today the number has risen to 29. Apart from the secretaries of state, who were introduced in the revision of 1970 and who are not part of the Council of Ministers, there has never been a constitutionally established hierarchy among the ministers. The position of prime minister was created in 1919, and that of vice prime minister in 1961. Before linguistic equality was formally established, an unwritten rule prevailed that, except for the prime minister, the government must include as many Dutch- as French-speaking ministers. The provinces, divided into 44 administrative districts (*arrondissements*) and 2,373 *communes*, are each under the authority of a governor, with legislative power exercised by the provincial council. The Députation Permanente (Permanent Deputation), elected from the members of the provincial council, provides for daily provincial administration.

Each *commune* is headed by a burgomaster, and the communal council elects the deputy mayors (*échevins*). The amalgamation of *communes* into federations is permitted under the constitution.

The political process. *Elections.* All male citizens have enjoyed suffrage since 1921, and all female citizens since 1932 for communal elections, since 1948 for provincial and national elections. The age of suffrage was lowered from 21 to 18 in October 1970.

Communal elections take place every four years, while provincial and national elections are held every six years.

Deputies and communal representatives are elected directly, as are certain senators, while other senators are either designated by the provincial councils or are selected by the elected senators and the provincial senators.

Parties. The traditional parties are, in order of importance, the Christian Social Party (PSC), the Belgian Socialist Party (PSB), the Liberty and Progress Party (PLP), and the Communist Party (PC). In addition, a few new parties have emerged, notably the People's Union (VU) in Flanders, the Walloon Assembly (RW) in the French-speaking area, and the French Democratic Defense Front (FDF) in Brussels. Certain tendencies seemed to hold for the country as a whole in the early 1970s: the PSC, the PSB, and the VU were maintaining their positions; the PLP and the PC were in retreat; the Walloon parties were improving their position; and the FDF was clearly progressing. The representatives of the linguistic political groups, who are in favour of a federal Belgium, have taken a stand against the recent constitutional modifications, which they regard as ineffective.

Participation of the citizen. A referendum on specific issues is very unusual. Participation in decision making is exercised on various levels by the representatives, party members, and the electorate, in the latter case through the compulsory vote. Citizens are informed of political events through the press, but with press ownership increasingly concentrated in fewer hands, many persons consider the medium to be unamenable to the expression of a wide range of opinion. Radio and television often organize debates and discussions that provide political information. In spite of these efforts, a marked disaffection exists among the citizens with regard to politics.

Justice. Judges are appointed for life by the king; they cannot be removed. At the cantonal, or lowest, judicial level, justices of the peace decide civil and commercial cases, and police tribunals decide criminal cases. At the district level, judicial powers are divided among the tribunals of first instance, which are subdivided into civil, criminal, and juvenile courts, and commercial and labour tribunals. At the appeals level, the courts of appeal include civil, criminal, and juvenile divisions that are supplemented by labour courts. Courts of assizes sit in each province to judge crimes and political and press offenses. These are composed of three judges and 12 citizens chosen by lot.

Structure of the judiciary

The Supreme Court of Appeal is composed of three chambers: civil and commercial, criminal, and one for matters of social and fiscal law and the militia. The last court does not deal with cases in depth, but regulates the application of the law throughout all jurisdictions. The military jurisdictions judge all cases concerning offenders responsible to the army and, in time of war, those concerning persons accused of treason. The State Council arbitrates in disputed administrative matters and gives advice on all bills and decrees.

National defense. Under the constitution, the king is required to "maintain national independence and territorial integrity," is charged with "the command of land and sea forces" and the making of treaties of alliance and of peace, and is given the power to confer ranks in the army. Belgium's defense policy is, in fact, determined by the Council of Ministers, although the constitution limits the government's powers in military matters by requiring the intervention of the legislature to determine the method of recruitment into the army, the size of the annual conscription draft, and the promotion, rights, and obligations of military personnel. The legislature is also required to decide military pensions and retirement ages, as well as on the deprivation of military ranks, honours, and pensions.

The country's armed forces comprise land, air, and naval forces of some 96,000 men, about 51,000 of whom are career military personnel. Belgium is a member of the North Atlantic Treaty Organization (NATO). The intervention force attached to NATO is equipped with Leopard tanks, while the air defense has various missile capabilities.

The social milieu. *Wages and the cost of living.* Wages are tied to the consumer price index and increase each time the price level rises by 2 to 3 percent. The social-planning agreements that the public services as well as several branches of industry provide for wage increases are linked with increased productivity. These arrangements partially explain why there have been relatively few industrial disputes since the nationwide strike of 1960.

Productivity agreements

Health and welfare. The great improvement in health conditions during the middle of the 20th century has been due as much to the programs of social insurance, covering about 95 percent of the population, as to advances in medical science. Compulsory medical examinations for workers and children were legislated during the 1960s. Overall, the kingdom has more than 15 doctors per 10,000 citizens and more than 350 hospitals, though the distribution of facilities and personnel in all provinces is uneven. In addition, hundreds of centres offered specialized help in medical, psychological, and geriatric areas as well as in rehabilitation of the handicapped. Under a 1925 statute, each commune has a commission of public assistance that is represented on the communal council; it acts to provide aid to the indigent.

Housing. Building is encouraged in a number of ways, including government-guaranteed mortgage loans that have lower interest rates. The National Housing Society oversees public housing construction for low-income families, and the state also is in charge of eliminating slum conditions in urban areas or where immigrant workers are forced to seek housing.

Education. Freedom of education is a constitutional guarantee in Belgium, but conflicts between public confessional (*i.e.*, Roman Catholic) schools date almost to the founding of the kingdom and remain among the most delicate problems within the entire social fabric. Currently, a dual system of state and "free" schools exists on the primary and secondary levels, the latter subsidized by the state to compensate for the abolition of fees in 1958. The language of instruction is either French or Dutch, depending on the region. Secondary schools are graded into a lower category, staffed by graduates from teachers' colleges, and institutions staffed by university graduates and offering either a classical or modern curriculum.

In addition to numerous specialized institutions for advanced training, Belgium has four universities. Both the Catholic University of Louvain (1425) and the Free Brussels University (1834) were bilingual, whereas Liège (1817) teaches in French, Ghent (1817) in Dutch.

CULTURAL LIFE AND INSTITUTIONS

The cultural and artistic heritages of the Belgian people are long and rich, reaching high points in the painting of Pieter Bruegel the Elder, the music of Josquin des Préz, Orlando di Lasso, and César Franck, the dramas of Maurice Maeterlinck and Michel de Ghelderode, and in the many palaces, castles, and cathedrals of the Belgian cities and countryside.

A basic cultural diversity exists between the Flemish and Walloon sectors. In general, the Walloons are attuned to the vibrations of Paris, whereas the Flemings are more concerned with maintaining and broadening their own cultural identity without strong reference to The Netherlands. Cultural unity is stressed throughout the country by means of the institutions known as "houses of culture," a use to which the former royal palace in Antwerp, for example, has been converted. The Walloons have emphasized half-cultural, half-touristic activities of the "poetry festival" type.

The cultural milieu

The arts. A variety of trends has resulted from this cultural diversity. Various schools of art have arisen, particularly in Vlaanderen. In music, avant-garde tendencies have become influential in Brussels, Liège, Ghent, and Antwerp; while Hainaut remains the centre of the classical and popular traditions. Literary works produced in Vlaanderen have a style peculiar to the region; whereas in the Walloon area and in Brussels, most authors are trying to write for a readership that is more inclined to French, and especially Parisian, works. More-

over, some of these same French works are by Belgian authors living in France, and still others are by writers living in Belgium but considered French.

Political decentralization and its corollary, regional autonomy, both of which characterize Belgium, have engendered their cultural counterpart. Brussels has lost its former dominance in the arts to certain provincial capitals, which have become cradles of contemporary artistic creation, although it must be added that the Brussels Philharmonic remains the most important Belgian orchestra. Even so, the provinces have their own music festivals, such as the Flanders, Walloon, Mouscron, and Saint-Hubert festivals.

The renowned Queen Elizabeth of Belgium International Music Contest attracts talented young artists from all over the world. The international String Quartet Competition in Liège and the Young Musicians' Competition were first held in Belgium.

Among the most famous of Belgium's contributions to the contemporary arts is the Ballet of the 20th Century, founded in 1960 by Maurice Béjart. It has gained a world reputation for its combination of brilliant classical technique and flamboyant innovation in the dance, and it regularly tours the major cities of Europe and the Americas.

The traditional dynamism of Belgium's visual arts has found expression in a number of directions. The Surrealist René Magritte is probably the nation's best known painter of the 20th century. Important work has been done in abstractions by a group of young painters and also in neorealism. (See also the appropriate sections of such articles as LITERATURE, WESTERN; VISUAL ARTS, WESTERN; and THEATRE, WESTERN.)

Traditional and popular culture. The traditional culture is based on the celebration of the seasons. In the Walloon areas there are joyous spring festivals such as the carnivals of Binche and Stavelot; summer festivals such as the procession of giants at Ath and the dragon battle in Mons; fall festivals such as the Toussant family mortuary celebrations or collective ones like that of November 11; and the winter festivals of St. Nicholas, Christmas, and New Year. In Vlaanderen, these festivals have become folkloric celebrations with a religious or historic character. Notable festivals include the Festival of Cats in Ieper, and the Procession of the Holy Blood in Brugge. Finally, marionette shows can be found in towns from Toone to Brussels. This traditional folk culture is in strong contrast to modern forms of popular culture, which as everywhere in the West are dominated by television and cinema and exemplified by the pop-music festival held first at Comblain-La-Tour and then at Amougies.

The artistic heritage is preserved in the many arts museums located throughout the country and in the very structure of countless buildings and parks. The Musée Instrumental du Conservatoire Royal is the only one of its kind in Europe. The framework for cultural and artistic activities is provided by fine arts palaces, operas, the "houses of culture," and other cultural centres in Brussels and the provinces, and by private and subsidized theatres and the five royal conservatories of music, in Antwerp, Brussels, Ghent, Liège, and Mons.

Press and broadcasting. About 35 daily newspapers are published in Belgium, with a circulation of more than 2,500,000 copies. This apparent diversity is actually controlled by only four large groups: Groep een (Group one), which is Catholic and Flemish; the Belgian Newspaper Union, a politically and linguistically diversified group centring around the Brussels daily *Le Soir;* Tiercé (the Third), a group controlling the Catholic *Libre Belgique* and the liberal *Dernière Heure* and a chain; and the Socialist Newspaper Union. Each paper or chain within these larger groups preserves its autonomy, with cooperation limited to technical and advertising matters. A German-language daily is published in Eupen. The majority of newspapers have some political affiliation, but only those of the Socialist press are linked to a political party. Belgium has several periodicals, but these face strong foreign competition. Nevertheless, the women's and the children's press are more widely distributed abroad than in Belgium.

Newspaper groups

Radio broadcasting was born in Belgium. As early as 1913, weekly musical broadcasts were given from the Laeken Royal Park. Radio-Belgium, founded in 1923, was broadcasting the equivalent of a spoken newspaper as early as 1926.

Belgian Television Broadcasting (RTB), which broadcasts in French, and Belgian Radio and Television (BRT), in Dutch, were created as public services. They are both autonomous and managed by an administrative council of ten members, of whom eight are named by Parliament and, in fact, represent the traditional parties. The two systems share a service institute, which assures administrative and financial coordination, and are financed by a tax on receivers. Unlike neighbouring countries, Belgium forbids advertising on radio and television. The country's small size and linguistic divisions limit the resources available to each of the systems. Television is available only in the evening, and colour broadcasting had not become available by the early 1970s. The BRT and RTB provide three types of radio: general information and entertainment; programs for specific audiences, particularly regional ones; and cultural programs and stereophonic music. Television programs do not differ greatly from those in other European countries. The BRT emphasizes adult education, whereas the RTB attempts to make information and culture accessible to the widest possible audience.

PROSPECTS

Although the demographic structure is an aged one, Belgium is a modern country. Since the 1960s, the country and the capital have been transformed both intellectually and physically by the construction of expressways, hotels, and office buildings of daring design. The cultural influence of Belgium's contemporary writers, painters, and sculptors, and of its dance companies, spreads far beyond its borders.

Belgium's international influence is enhanced by its role as the seat of NATO, the Common Market, and SHAPE (Supreme Headquarters Allied Powers Europe), and by the government's concern to adapt the country and its capital to a new European solidarity. In spite of these roles, which give Belgium a place in the community of Western nations that its size would not suggest, the kingdom remains prey to instabilities within its social structure and to intergroup hostilities that trace their genesis well into the past. Close observers have felt that the internal tensions might erupt at any time and tear the nation asunder, especially over differences emerging from the linguistic, religious, and educational problems. Throughout its history, however, Belgium has known this turmoil, and much of its public energy has gone into discovering the paths to compromise. With the increasing efforts by the government to provide each sector of the society with a freer hand in arranging its own destiny and with a greater sense of its distinctive cultural past and present, Belgium may well be able to maintain the tenuous balance that has characterized its existence as a political entity.

BIBLIOGRAPHY

General Works: VERNON MALLINSON, *Belgium* (1969), detailed, well documented, with bibliography; FRANK E. HUGGETT, *Modern Belgium* (1969), thorough and discerning; G.M. ASHBY, *Belgium* (1955), a brief, light account; *Fodor's Belgium and Luxembourg* (annual), a useful travel guide; JAN A. GORIS (ed.), *Belgium* (1945), articles by a number of American and Belgian scholars on all aspects of national life; THERESA HENROT, *Belgique* (1958; Eng. trans., 1961), a popular, informative work about daily life; TUDOR EDWARDS, *Belgium and Luxembourg* (1951), interesting on architectural features; DOROTHY LODER, *Belgium and Her People,* rev. ed. (1967), a popular account; F.J. MONKHOUSE, *A Regional Geography of Western Europe,* 2nd ed. (1964), good coverage on Belgium; JULES TARLIER and ALPHONSE WAUTERS, *La Belgique ancienne et moderne: Géographie et histoire des communes belges,* 4 vol. (1859–87, reprinted 1963).

History: ADRIEN DE MEEUS, *Histoire des Belges* (1956; Eng. trans., 1962), from prehistory to the aftermath of World War

II; HENRI PIRENNE, *Histoire de Belgique*, 7 vol. (1900–32), the standard work; and *Les Anciennes Démocraties des Pays-Bas* (1910; Eng. trans., *Early Democracies in the Low Countries*, 1963), on urban society and political conflict in the Middle Ages and the Renaissance; F.G. EYCK, *The Benelux Countries: An Historical Survey* (1959), includes readings in original sources; MARC SCHREIBER, *Belgium* (1945), a concise history with a short bibliography; B.D. GOOCH, *Belgium and the February Revolution* (1963), a study of the impact on Belgium of the 1848 revolution in France; EMILE CAMMAERTS, *The Keystone of Europe* (1939), the foundation and development of independent Belgium to 1939; SHEPARD B. CLOUGH, *A History of the Flemish Movement in Belgium* (1930, reprinted 1968), scholarly; HERMAN VAN DER WEE, *The Growth of the Antwerp Market and the European Economy (Fourteenth–Sixteenth Centuries)*, 3 vol. (1963); FRANS VAN KALKEN, *Histoire de la Belgique et de son expansion coloniale* (1954).

Culture: CENTRE BELGE DE DOCUMENTATION MUSICALE, *Music in Belgium: Contemporary Belgian Composers* (1964), biographies with brief lists of recordings; VERNON MALLINSON, *Modern Belgian Literature, 1830–1960* (1966); BELGIAN INFORMATION AND DOCUMENTATION INSTITUTE, *The Language Problem in Belgium* (1967).

Specialized aspects: S.J. DE LAET, *The Low Countries* (1958), on prehistoric Belgium; L. MORISSENS, "Economic Policy in Belgium," in E.S. KIRSCHEN *et al.*, *Economic Policy in Our Time*, vol. 3 (1964), traces economic developments from 1949 to 1961; VERNON MALLINSON, *Power and Politics in Belgian Education, 1815 to 1961* (1963), a detailed examination.

Photographic views: JEAN ROUBIER, *Benelux: Holland, Belgium, Luxembourg* (1958); CAS OORTHUYS, *This Is Belgium* (1955).

(Ar.D)

Belgrade

Belgrade (Beograd, meaning White Fortress), the capital of the Socialist Federal Republic of Yugoslavia as well as of the component republic of Serbia, is situated on the Danube River at the point where it is joined by the Sava. The population of the city proper in 1971 was about 746,000, that of Greater Belgrade a little over 1,200,000.

Location

Belgrade's location is exceptionally favourable, at the convergence of three historically important arteries of travel: an east–west route along the Danube Valley from Vienna to the coast of the Black Sea; another that goes westward along the valley of the Sava toward Trieste and northern Italy; and a third running southeast along the valleys of the Morava and Vardar rivers to the Aegean Sea. To the north and west of the city lies the Pannonian Basin, which includes the great grain-growing region of Vojvodina. To the south is the undulating hill country of the Šumadija, with its diversified farming.

There are evidences of Stone Age settlements in the area. The city grew up around an ancient fortress built on the Kalemegdan headland, encompassed on three sides by the Sava and the Danube. The first fortress was built by the Celts in the 4th century BC and was known by the Romans as Singidunum. They built roads to the fortress and a bridge across the Sava, establishing a city with paved streets, temples, an amphitheatre, and other public buildings. It was destroyed by the Huns in 442, and after that it changed hands among the Sarmatians, Goths, and Gepidae, was recaptured by the emperor Justinian, was held by the Franks and the Bulgars, and in the 11th century became a frontier town of Byzantium. In the year 1284 it came under Serbian rule, and in 1402 Stephen Lazarević made it the capital of Serbia. The Turks besieged the city in 1440, and from 1521 onward it was in their hands except for three periods of occupation by the Austrians (1688–90, 1717–39, and 1789–91).

During the Turkish period Belgrade was a city of narrow, winding streets and walled houses, a lively commercial centre where goods were traded from various parts of the Ottoman Empire. In the year 1661 it had 98,000 inhabitants. After the first Serbian uprising under Karageorge in 1804, Belgrade became the Serbian capital during 1807–13, but the Turks recaptured it. The Serbs were given control of the citadel in 1867, when Belgrade once more became the capital of Serbia. An exodus of the Turkish inhabitants followed, reducing the population to a level far below that of two centuries earlier.

Belgrade's population in the year 1900 was around 70,000. By 1921, when it became the capital of Yugoslavia, it had about 110,000 inhabitants. After World War II it expanded from around 550,000 in 1948 to about 746,000 in 1971. The rapid growth since World War II can be attributed primarily to the migration from rural areas of Serbia, as a consequence of industrialization. In 1971, 68 percent of the population of Greater Belgrade had been born elsewhere. Most of the inhabitants are Serbs; according to the census of 1971, only 15.8 percent were of other nationalities, the largest non-Serb groups being Croats and Montenegrins.

Industrial and commercial importance

Since World War II Belgrade has become an industrial city, accounting in 1973 for around 7 percent of Yugoslavia's industrial output. It produces motors, tractors and combines, machine tools, electrical equipment, chemicals, textiles, and building materials. It is the largest commercial centre in Yugoslavia, handling more than half the country's foreign trade. Three international railroad lines pass through Belgrade, which is also served by about 10 paved highways and by river vessels. The growth of the city has exceeded the capacity of its narrow, hilly streets, where passenger cars must compete with trucks, buses, and streetcars. The airport is 10 miles west of the city at Surčin.

Belgrade is the seat of numerous government bodies as well as of political, cultural, and economic organizations. The University of Belgrade, founded in 1863, and other institutions of higher education enroll about 53,000 students. There are about 310 schools, with more than 164,000 students. Belgrade has 30 museums and galleries, of which the oldest, the National Museum (Narodni Muzej), was founded in 1844.

Toni Schneiders



Old Belgrade on the Sava, with the Orthodox cathedral in the background.

Recreation areas are situated along the rivers and in the nearby hills. Many people have weekend cottages on the Danube and Sava. At Avala, about 12 miles (19 kilometres) south of Belgrade, is Ivan Meštrović's monument to the unknown soldier and also a television tower with a restaurant and a viewing platform, hotels, and hiking trails. Košutnjak is a large park and forest preserve. The Lipovačka Šuma, about 15 miles (24 kilometres) southwest of the city, is another forested public park. To the north are hunting areas along the Danube and the lakes near Bela Crkva.

In the course of its growth, Belgrade spread southward and southeastward over a hilly terrain that varies from 218 to 771 feet (66 to 235 metres) above sea level. The old fortress of Kalemegdan is now a historical monument, its former glacis having been rebuilt as a garden, from which is seen a famous view of the plain across the Sava and the Danube. Since World War II a district called New Belgrade (Novi Beograd) has been built on the plain to the north between the Sava and the Danube. By 1971 dwellings for around 100,000 people had been constructed, along with government buildings, offices, parks, and commercial and industrial facilities. With the completion of the Iron Gate hydroelectric and navigation scheme in 1972, the Danube became navigable for ships of up to 5,000 tons, making it possible for oceangoing ships to reach Belgrade from the Black Sea. When the Rhine–Main–Danube canal is completed it will also be possible for ships from Rotterdam to sail through central Europe to Belgrade.

BIBLIOGRAPHY. VOJISLAV RADOVANOVIC, *Geografski položaj našeg glavnog grada* (1960); and MILORAD VASOVIC, *Osobenosti geografskog položaja Beograda* (1969), discuss the location of Belgrade in terms of communications, access to raw materials, and other economic, political, and administrative factors. SIMA MILOJEVIC, *Reljef beogradskog zemljišta* (1930); TOMISLAV RAKICEVIC, *Klima Beograda* (1960); and DUSAN DUKIC, *Reke Beograda i njegove okoline* (1960), analyze in detail the physical and geographic characteristics of the Belgrade region. MILORAD VASOVIC, *Industrija Beograda* (1962); LJUBINKO SRETENOVIC and JOVAN ILIC, *Saobraćajni problemi Beograda* (1962); and ZIVADIN JOVICIC, *Beograd: Turističko-geografska monografija* (1968), treat the chief economic functions of contemporary Belgrade. *Statistički godišnjak Beograda*, the Belgrade statistical yearbook, gives information on population and economic, educational, and cultural activities in Greater Belgrade. SERBIAN ACADEMY OF ARTS AND SCIENCES, *Istorija Beograda*, 3 vol. (1974), covers its history from pre-Roman times to the present.

(M.V./F.B.S.)

Belisarius

One of history's great generals, Belisarius was the leading military figure in the age of the later Roman (Byzantine) emperor Justinian I (527–565). As one of the last important figures in the Roman military tradition, he helped restore North Africa and Italy to the empire.

Little is known of Belisarius' early years. Like Justinian, he was born in the Balkans. Some traditions assign him an unlikely Slavic background, but his exact origins and the precise date of his birth (*c.* 505) are undocumented. As a member of Justinian's bodyguard, he came to the Emperor's attention, and he was appointed to a command at about the age of 25. His public career thereafter is thoroughly described by the historian Procopius, who was a member of his personal staff for the first 15 years of his campaigns and who observed the general's activities personally.

Belisarius won his first laurels as commander on the Mesopotamian front against the empire's eastern neighbour and rival, Sāsānian Persia. He won a brilliant victory at Dara in 530; despite a subsequent defeat the following year at Sura (Callinicum), he emerged as the hero of the war by the time Justinian negotiated its end. Belisarius was in Constantinople, the capital, when the Nika Insurrection broke out there in January 532, and he further gained the Emperor's confidence by commanding the troops that ended the episode by massacring the rioters. About this time, meanwhile, Belisarius married the widowed Antonina, who, as an old friend to the empress Theodora, had influence at court that was later to be of great importance to him.

Justinian next chose Belisarius to begin the reconquest of the western Roman territories occupied by Germanic peoples. In 533 he was sent with a small force to attack the Vandals in North Africa. In two stunning victories he shattered the Vandal kingdom within a few months. Returning to Constantinople, he was granted a triumphal celebration. The recovery of Italy from the Ostrogoths began in 535. Belisarius quickly took Sicily and moved steadily northward on the mainland, seizing Naples by storm and occupying Rome. Revitalized under their new king, Witigis, the Goths besieged Rome in 537–538, but Belisarius held out there brilliantly. Hampered by conflicts within his command, his advance further northward was delayed, but by 540 the Goths, hard-pressed, offered to surrender if Belisarius would rule over them as emperor. Justinian had already come to fear that so popular a commander might win sufficient prestige to aim at his throne. Dissembling, Belisarius accepted the Goths' capitulation and then refused the title, thus antagonizing the Goths without relieving Justinian's suspicions.

First Italian campaign

The Emperor recalled him from Italy in temporary disfavour but sent him in the following year to fight again in Mesopotamia against the Sāsānians. Despite some successes, Belisarius had difficulties with his unruly soldiers, and then he was stripped of his command on charges of disloyalty. Only Theodora's intervention, out of friendship for Antonina, prevented his disgrace and ruin. Imperial rule had broken down in Italy under Belisarius' incompetent successors. He was reassigned there in 544, but Justinian, more suspicious and niggardly than ever, would not back him with sufficient men and money. Belisarius operated insecurely around the Italian coasts for the next few years, even briefly holding Rome once more, but effective opposition to the Ostrogoths was impossible. Theodora died in 548, and he was soon recalled. The Italian wars were left to be completed by other generals, notably the eunuch Narses, who would receive Justinian's fuller support.

Returning to Constantinople, Belisarius was allowed to retain his wealth and large household bodyguard. When marauding Hun tribes menaced the city in 559, the Emperor summoned Belisarius back into service. Adding what men he could find to his private retinue, he frightened the Huns away by clever stratagems and then resumed his retirement. Three years later he was accused of involvement in a plot against Justinian's life and, though probably innocent, was disgraced. Partially restored to favour in 563, he was left in peace until his death, in March 565, a few months before the death of the ungrateful emperor he had served so well.

Final disgrace

Belisarius' character is elusive. Two primary impulses guided his life: loyalty to Justinian and passion for his wife, Antonina. Despite the treatment he often received from Justinian, Belisarius never wavered in his obedience, contributing one of the nobler dimensions to Justinian's era. Antonina seems to have utterly captivated him, but her reckless and immoral behaviour brought him embarrassment and humiliation.

In Procopius' *Secret History* (*Historia arcana*), Belisarius is given the least unfavourable treatment of the age's leading personalities. His reputation endured for centuries, and later legends, often mixed with stories about others, developed about him. The most famous had him actually blinded by Justinian and forced to beg in the streets in his old age. The 18th-century French writer Jean-François Marmontel used the story of Belisarius as a vehicle for an oblique attack on Louis XV and for a plea for tolerance and justice, in his philosophical novel *Bélisaire* (1767). Robert Graves's vivid novel *Count Belisarius* (1938) is the best fictionalized treatment of the General's life.

BIBLIOGRAPHY. The chief contemporary record of Belisarius' life is in the *History of the Wars* and *Secret History* by PROCOPIUS OF CAESAREA. The only recent biography is L.M. CHASSIN, *Bélisaire, généralissime byzantin (504–565)* (1957); the most detailed among general narratives is still that in

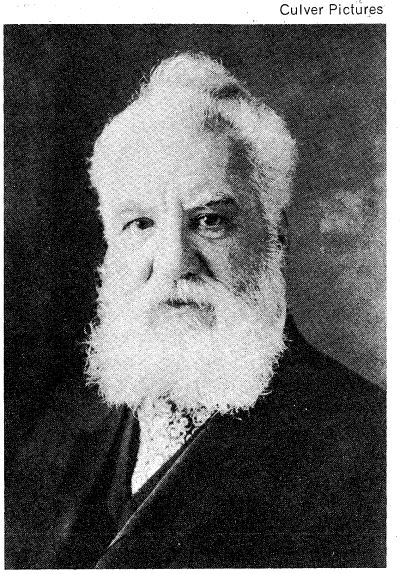
T. HODGKIN'S old classic, *Italy and Her Invaders*, vol. 3 and 4 (1885).

(J.W.Ba.)

Bell, Alexander Graham

Versatility and an insatiable curiosity describe the career of Alexander Graham Bell, who is best remembered as the inventor of the telephone. For two generations his family had been recognized as leading authorities in elocution and speech correction, with Alexander Melville Bell's *Standard Elocutionist* passing through nearly 200 editions in English. Young Bell and his two brothers were trained to continue the family profession. His early achievements on behalf of the deaf and his invention of the telephone before his 30th birthday bear testimony to the thoroughness of his training.

Culver Pictures



Bell.

Born March 3, 1847, in Edinburgh, Alexander ("Graham" was not added until he was 11) was the second of the three sons of Alexander Melville Bell and Eliza Grace Symonds Bell. Apart from one year at a private school, two years at Edinburgh's Royal High School (from which he graduated at 14), and attendance at a few lectures at Edinburgh University and at University College in London, Bell was largely family trained and self-taught. His first professional post was at Mr. Skinner's school in Elgin, County Moray, where he instructed the children in both music and elocution. In 1864 he became a resident master in Elgin's Weston House Academy, where he conducted his first studies in sound. Appropriately, Bell had begun professionally as he would continue through life—as a teacher-scientist.

Career as teacher-scientist

In 1868 he became his father's assistant in London and assumed full charge while the senior Bell lectured in America. The shock of the sudden death of his older brother from tuberculosis, which had also struck down his younger brother, and the strain of his professional duties soon took their toll on young Bell. Concern for their only surviving son prompted the family's move to Canada in August 1870, where, after settling near Brantford, Ontario, Bell's health rapidly improved.

In 1871 Bell spent several weeks in Boston, lecturing and demonstrating the system of his father's *Visible Speech*, published in 1866, as a means of teaching speech to the deaf. Each phonetic symbol indicated a definite position of the organs of speech such as lips, tongue, and soft palate, and could be used by the deaf to imitate the sounds of speech in the usual way. Young A. Graham Bell, as he now preferred to be known, showed, using his father's system, that speech could be taught to the deaf. His astounding results soon led to further invitations to lecture.

Even while vacationing at his parents' home Bell continued his experiments with sound. In 1872 he opened his own school in Boston for training teachers of the deaf, edited his pamphlet *Visible Speech Pioneer*, and continued to study and tutor; in 1873 he became professor of vocal physiology at Boston University.

Never adept with his hands, Bell had the good fortune to discover and inspire Thomas Watson, a young repair mechanic and model maker, who assisted him enthusiastically in devising an apparatus for transmitting sound by electricity. Their long nightly sessions began to produce tangible results. The fathers of George Sanders and Mabel Hubbard, two deaf students whom he helped, were sufficiently impressed with the young teacher to assist him financially in his scientific pursuits. Nevertheless, during normal working hours Bell and Watson were still obliged to fulfill a busy schedule of professional demands. It is scarcely surprising that Bell's health again suffered. On April 6, 1875, he was granted the patent for his multiple telegraph; but after another exhausting six months of long nightly sessions in the workshop, while maintaining his daily professional schedule, Bell had to return to his parents' home in Canada to recuperate. In September 1875 he began to write the specifications for the telephone. On March 7, 1876, the United States Patent Office granted him Bell Patent Number 174,465 covering "The method of, and apparatus for, transmitting vocal or other sounds telegraphically . . . by causing electrical undulations, similar in form to the vibrations of the air accompanying the said vocal or other sounds."

Invention of the telephone

Within a year followed the commercial application and, a few months later, the first of hundreds of legal suits. Ironically, the telephone—until now all too often regarded as a joke and its creator-prophet as, at best, an eccentric—was the subject of the most involved patent litigation in history. The two most celebrated of the early actions were the Dowd and Drawbaugh cases wherein the fledgling Bell Telephone Company successfully challenged two subsidiaries of the giant Western Union Telegraph Company for patent infringement. The charges and accusations were especially painful to Bell's Scottish integrity, but the outcome of all the litigation, which persisted throughout the life of his patents, was that Bell's claims were upheld as the first to conceive and apply the undulatory current. In 1877 Bell married Mabel Hubbard, ten years his junior.

The Bell story does not end with the invention of the telephone; indeed, in many ways it was a beginning. A resident of Washington, D.C., Bell continued his experiments in communication, which culminated in the invention of the photophone—transmission of sound on a beam of light; in medical research; and in techniques for teaching speech to the deaf.

In 1880 France honoured Bell with the Volta Prize; and the 50,000 francs (then roughly equivalent to U.S. $10,000) financed the Volta Laboratory, where, in association with Charles Sumner Tainter and his cousin, Dr. Chichester A. Bell, the Graphophone was invented. Employing an engraving stylus, controllable speeds, and wax cylinders and disks, the Graphophone presented a practical approach to sound recording. Bell's share of the royalties financed the Volta Bureau and the American Association to Promote the Teaching of Speech to the Deaf (since 1956 the Alexander Graham Bell Association for the Deaf). May 8, 1893 was one of Bell's happiest days; his 13-year-old prodigy, Helen Keller, participated in the ground-breaking ceremonies for the new Volta Bureau building—today an international information centre relating to the oral education of the deaf.

In 1885 Bell acquired land on Cape Breton Island in Nova Scotia. There, in surroundings reminiscent of his early years in Scotland, he established a summer home, Beinn Bhreagh, complete with research laboratories.

In 1898 Bell succeeded his father-in-law as president of the National Geographic Society. Convinced that geography could be taught through pictures, he sought to promote an understanding of life in distant lands in an age when travel was limited to a privileged few. Again he found the proper hands, Gilbert Grosvenor, his future son-in-law, who transformed a modest pamphlet into

a unique educational journal reaching millions throughout the world.

As interest in the possibility of flight increased after the turn of the century, he experimented with giant man-carrying kites. Characteristically, Bell again found a group of four willing young enthusiasts to execute his theories. Always an inspiration, Mabel Hubbard Bell, wishing to maintain the stimulating influence of the group, soon founded the Aerial Experiment Association, the first research organization established and endowed by a woman. Deafness was no handicap to the wife of Professor Bell. At Beinn Bhreagh, Bell entered new subjects of investigation, such as sonar detection, solar distillation, the tetrahedron as a structural unit, and hydrofoil craft, one of which weighed more than 10,000 pounds and attained a speed record of 70 miles per hour in 1919.

Other inventions

Apart from his lifelong association with the cause of the deaf, Bell never lingered on one project. His research interests centred on basic principles rather than on refinements. The most cursory examination of his many notebooks shows marginal memos and jottings, often totally unrelated to the subject at hand—reminders of questions and ideas he wanted to investigate. It was impossible for him to carry each of his creative ideas through to a practical end. Many of his conceptions are only today seeing fruition; indeed, some undoubtedly have yet to be developed. The range of his inventive genius is represented only in part by the 18 patents granted in his name alone and the 12 he shared with his collaborators. These included 14 for the telephone and telegraph, four for the photophone, one for the phonograph, five for aerial vehicles, four for hydroairplanes, and two for a selenium cell.

Until a few days before his death at Beinn Bhreagh, August 2, 1922, Bell continued to make entries in his journal. During his last dictation he was reassured with "Don't hurry," to which he replied, "I have to."

BIBLIOGRAPHY. Two general biographies are R.V. BRUCE, *Alexander Graham Bell and the Conquest of Solitude* (1973), the best account of Bell and his work; and CATHERINE MACKENZIE, *Alexander Graham Bell* (1928). HELEN E. WAITE, *Make A Joyful Sound: The Romance of Mabel Hubbard and Alexander Graham Bell* (1961); and FRED DELAND, *Dumb No Longer* (1908), provide sympathetic accounts of his work with the deaf and the story of the telephone. The most detailed published account of Bell's telephonic work is his own last deposition (1892), published in 1908 by AMERICAN BELL TELEPHONE, *The Deposition of Alexander Graham Bell;* the original patents of 1876 and 1877 are printed in full. THOMAS B. COSTAIN, *Chord of Steel* (1960), is limited largely to the telephone.

(T.K.B.)

Bellini, Giovanni

Giovanni Bellini (*c.* 1430–1516), the founder of the Venetian school of painting, brought to painting a new degree of realism, a new wealth of subject matter, and a new sensuousness in form and colour. The teacher of Giorgione and Titian, he raised Venice to a centre of Renaissance art comparable to Florence and Rome. To be sure, more than a century before him a great Venetian painter, Paolo Veneziano, had established a considerable tradition, but it was short-lived. It had depended largely, as Venice itself had once done, on the Byzantine tradition, but Venice had now begun to be Westernized.

Little is known about Giovanni Bellini's family. His father, Jacopo, a painter, was a pupil of Gentile da Fabriano, one of the leading painters of the 15th-century Gothic revival, and may have followed him to Florence. In any case, Jacopo introduced the principles of the Florentine Renaissance to Venice before either of his sons. Apart from his sons Gentile and Giovanni, he had at least one daughter, Niccolosa, who married the painter Andrea Mantegna in 1453. The exact date of Giovanni's birth is unknown, but it is generally assumed that he was the younger brother, born about 1430. Both sons probably began as assistants in their father's workshop.

Early work

Giovanni's earliest independent paintings were more strongly influenced by the severe manner of the Paduan school, and especially of his brother-in-law, Mantegna, than they were by the graceful style of Jacopo. This influence is evident even after Mantegna left for the court of Mantua in 1460. Giovanni's earliest works date from before this period. They include a "Crucifixion," a "Transfiguration," and a "Dead Christ Supported by Angels." Several pictures of the same or earlier date are in the United States and others at the Museo Civico Correr in Venice. Four triptychs, a set of three panels used as an altarpiece, are still in the Venice Accademia, and two "Pietàs," both in Milan, are all from this early period. His early work is well exemplified in two beautiful paintings now in the National Gallery of London, "The Blood of the Redeemer" and "The Agony in the Garden."

In all his early pictures he worked with tempera, combining the severity and rigidity of the Paduan school with a depth of religious feeling and human pathos all his own. His early madonnas, following in his father's tradition, are mostly sweet in expression, but he substituted for a mainly decorative richness one drawn more from a sensuous observation of nature. Although the pronounced linear element—*i.e.*, the dominance of line rather than mass as a means of defining form, derived from the Florentine tradition and from the precocious Mantegna—is evident in the paintings, the line is less self-conscious than Mantegna's work, and, from the first, broadly sculptured planes offer their surfaces to the light from a dramatically brilliant sky. From the beginning Giovanni Bellini was a painter of natural light, as were Masaccio, the founder of Renaissance painting, and Piero della Francesca, its greatest living practitioner. In these earliest pictures the sky is apt to be reflected behind the figures in streaks of water making horizontal lines in a mere strip of landscape. In "The Agony in the Garden," the horizon moves up, and a deep, wide landscape encloses the figures, to play an equal part in expressing the drama of the scene. As with the *dramatis personae*, the elaborately linear structure of the landscape provides much of the expression, but an even greater part is played by the colours of the dawn, in their full brilliance and in the reflected light within the shadow. This is the first of a great series of Venetian landscape scenes that was to develop continuously for a century or more. To a city surrounded by water, the emotional value of landscape was now fully revealed.

Use of natural light

The great composite altarpiece with St. Vincent Ferrer, which is still in the church of SS. Giovanni e Paolo in Venice, was painted perhaps ten years later, toward the mid-1470s. But the principles of composition and the method of painting had not yet changed essentially; they had merely grown stronger in expression. It seems to have been during a voyage down the Adriatic coast, made probably not long afterward, that Bellini encountered the influence that must have helped him most toward his full development: that of Piero della Francesca. Bellini's great "Coronation of the Virgin" at Pesaro, the first Venetian picture in the full style of the Renaissance, probably reflects and carries still further in composition the ideas expressed by Piero in an unrecorded "Coronation of the Virgin," the lost centrepiece of a polyptych originally in the church of S. Agostino at Borgo Sansepolcro. Christ's crowning of his Mother beneath the effulgence of the Holy Ghost is a solemn act of consecration and the four saints who stand witness beside the throne are characterized by their deep humanity. Every quality of their forms is fully realized: every aspect of their bodies, the textures of their garments, and the objects that they hold. As with work by Masaccio and Piero della Francesca, the perspective and the polychrome of pavement and throne help to establish the group in space, and the space is enlarged by the great hills behind and rendered infinite by the luminosity of the sky, which envelops the scene and gathers all the forms together into one. Harmony is the aim of all art, but the significance of the harmony depends upon the significance of its parts, as well as upon the degree of its intensity. Here, Bellini has provided humanity with the full grandeur of nature, and it is nature endowed with all that is religious in man. The unity achieved has an emotional warmth that is uniquely his, a sensuousness peculiarly Venetian.

A new degree of technical achievement is implied. The fact that at this point Giovanni painted mainly in oil does not completely explain his greatness. Piero was one of many Italian painters who were already using the oil medium. A legend that Giovanni ceased to paint in tempera only after he was introduced to oils by Antonello da Messina, who was in Venice in 1475/76, is without point, for much the same effects can be produced in either medium.

Bellini's mastery of oils

It is the way of using the medium that makes the difference—and that depends upon the painter's intentions and upon his vision. It was Bellini's richer and wider vision that determined his future development. Oil paint is inclined to be the more transparent and fusible and therefore lends itself to richer colour and tone by allowing a further degree of glazing, the laying of one translucent layer of colour over another. It is this technique and the unprecedented variety with which he handled the oil paint that gives his fully mature painting the richness associated with the Venetian school.

It was his brother Gentile who was chosen by the government to continue the painting of great historical scenes in the Hall of the Great Council in Venice; but in 1479, when Gentile was sent on a mission to Constantinople, Giovanni took his place. From that time to 1480 much of Giovanni's time and energy was devoted to fulfilling his duties as conservator of the paintings in the hall, as well as painting six or seven new canvases himself. These were his greatest works, but they were destroyed when the huge hall was gutted by fire in 1577. We can now only gain an approximate idea of their design from "The Martyrdom of St. Mark" in the Scuola di S. Marco in Venice, finished and signed by one of Giovanni's assistants, and of their execution from Giovanni's completion of Gentile's "St. Mark Preaching in Alexandria" after his brother's death in Venice in 1507.

Yet a surprisingly large number of big altarpieces and comparatively portable works have survived and show the steady but adventurous evolution of his work. The principles and the technique of the Pesaro altarpiece find their full development in the still larger madonna altarpiece from S. Giobbe in the Venice Accademia, where the Virgin enthroned in a great apse and the saints beside her seem ready to melt into the reflected light. This seems to have been painted before the earliest of his dated pictures, the half-length "Madonna degli Alberetti," also in the Venice Accademia, of 1487. From this date on, there are pictures dated at intervals until 1515, and Giovanni seems to have been painting almost until the day of his death in the following year.

While for the first 20 years of Giovanni's career the subject matter was limited mainly to madonnas, pietàs, or crucifixions, toward the end of the century it began to be greatly enriched not so much by the wider choice of subjects, which were still mainly religious, as by the development of the *mise-en-scène*, the physical setting of the picture. He became one of the greatest of landscape painters. His study of outdoor light was such that one can deduce not only the season depicted but almost the hour of the day.

Late work

Bellini also excelled as a painter of ideal scenes; *i.e.*, scenes of primeval as opposed to individualized images. For the "St. Francis in Ecstasy" of the Frick Collection or the "St. Jerome at his Meditations," painted for the high altar of Sta. Maria dei Miracoli in Venice, the anatomy of the earth is studied as carefully as those of human figures; but the purpose of this naturalism is to convey idealism through the realistic portrayal of detail. In the landscape "Sacred Allegory," now in the Uffizi, he created the first of the dreamy enigmatic scenes for which Giorgione, his pupil, was to become famous. The same quality of idealism is to be found in his portraiture. His "Doge Leonardo Loredan" in the National Gallery, London, has all the wise and kindly firmness of the perfect head of state, and his "Pietro Bembo" (?) in the British royal collection portrays all the sensitivity of a poet.

Both artistically and personally the career of Giovanni Bellini seems to have been serene and prosperous. He lived to see his own school of painting achieve dominance and acclaim. He saw his influence propagated by a host of pupils, two of whom surpassed their master in world fame: Giorgione, whom he outlived by six years, and Titian.

The only personal description extant of Giovanni is from the hand of Albrecht Dürer, who wrote to the German humanist Willibald Pirkheimer from Venice in 1506 ". . . everyone tells me what an upright man he is, so that I am really fond of him. He is very old, and still he is the best painter of them all."

MAJOR WORKS

"Annunciation" (1450s; Thyssen-Bornemisza Collection, Castagnola, Switz.); "Transfiguration" (1450s; Museo Civico Correr, Venice); "Christ Blessing" (1450s; Louvre, Paris); "Virgin and Child" (1450s; Philadelphia Museum of Art, Philadelphia); "Pietà" (1460s; Museo Poldi Pezzoli, Milan); "The Agony in the Garden" (1460s; National Gallery, London); "The Blood of the Redeemer" (1460s; National Gallery, London); "Pietà with Virgin and St. John" (1460s; Brera, Milan); "St. Vincent Ferrer Polyptych" (mid-1470s; SS. Giovanni e Paolo, Venice); "Madonna with Sleeping Child" (1470s; Accademia, Venice); "Coronation of the Virgin" (1470s; Museo Civico, Pesaro); "St. Jerome" (1470s; Contini Bonacossi Collection, Florence); "Madonna with the Greek Inscription" (1470s; Museo di Castelvecchio, Verona); "Madonna and Child" (1470s; National Gallery of Art, Washington, D.C.); "Madonna with Standing Child" (1470s; Accademia, Venice); "Portrait of Georg Fugger" (1474; formerly Contini Bonacossi Collection, Florence); "St. Francis in Ecstasy" (*c.* 1480; Frick Collection, New York); "Enthroned Madonna from S. Giobbe" (1480s; Accademia, Venice); "Transfiguration" (1480s; Museo e Gallerie Nazionali di Capodimonte, Naples); "Resurrection" (1480s; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Madonna degli Alberetti" (1487; Accademia, Venice); "Madonna with Doge Agostino Barbarigo" (1488; Sta. Maria dei Frari, Venice); "Madonna of the Pomegranate" (*c.* 1490; National Gallery, London); "Madonna and Child with St. Paul and St. George" (*c.* 1490; Accademia, Venice); "The Allegory of the Souls in Purgatory" (1490s; Uffizi, Florence); "Portrait of a Young Man" (1490s; National Gallery of Art, Washington, D.C.); "Doge Leonardo Loredan" (*c.* 1501; National Gallery, London); "Enthroned Madonna with Four Saints" (1505; S. Zaccaria, Venice); "The Madonna of the Meadow" (*c.* 1505; National Gallery, London); "Madonna" (1510; Brera, Milan); "St. Jerome with St. Christopher and St. Augustine" (1513; S. Giovanni Grisostomo, Venice); "The Feast of the Gods" (landscape overpainted by Titian, 1500s; National Gallery of Art).

BIBLIOGRAPHY. The earliest biography of Bellini is by GIORGIO VASARI in *The Lives of the Painters, . . .* (1568; Eng. trans. by E.H. and E.W. BASHFIELD and A.A. HOPKINS, 4 vol., 1913). By the end of the 19th century, notable studies in English, more concerned with his painting than his life, had appeared in J.A. CROWE and G.B. CAVALCASELLE, *A History of Painting in North Italy, . . .*, vol. 1 (1871); BERNHARD BERENSON, *The Venetian Painters of the Renaissance* (1894); and ROGER FRY, *Giovanni Bellini* (1899). In the 20th century the study of Bellini's work has been extensive—the contributions of the Italian art historians ADOLFO VENTURI and, later, ROBERTO LONGHI and of the German G. GRONAU in particular marking advances in knowledge—and it received special impetus from the Bellini exhibition held in Venice in 1949, which brought together widely dispersed examples (nearly 130 paintings) for immediate comparison. RODOLFO PALLUCCHINI, who prepared the catalog for that exhibition, has written a standard work, *Giovanni Bellini* (1959; Eng. trans., 1962), large quarto, tracing the development of Bellini's painting, with notes relating to the 285 plates, 35 of them in colour, and with a bibliography. GIUSEPPE FIOCCO, *Giovanni Bellini* (Eng. trans. 1960), folio size, is a fairly brief monograph illustrated by 93 admirable plates, 48 of them in colour. FRITZ HEINEMANN, *Giovanni Bellini e i Belliniani*, 2 vol. (trans. into Italian, 1962), large quarto, is an annotated catalog of 1,902 works by Bellini and his followers, including copies of lost works and of works once wrongly attributed to Bellini, with a separate listing of works by MARCO BASAITI, and containing 910 illustrations including 24 coloured plates. STEFANO BOTTARI (ed.), *Tutta la pittura di Giovanni Bellini*, 2 vol. (1962), octavo, a catalog of Bellini's paintings, with notes and illustration in each case, making 343 page-size monochrome photographs in all. PHILIP HENDY and LUDWIG GOLDSCHEIDER, *Giovanni Bellini* (1945), folio size, is an introductory essay with five coloured plates and over 100 monochrome illustrations. A later important work is GILES ROBERT-

SON, *Giovanni Bellini* (1968), quarto, an excellent account of Bellini's career and *oeuvre* with the findings of art historical scholarship taken into consideration, accompanied by 120 plates in monochrome.

(P.He.)

Belorussian Soviet Socialist Republic

Lying in the western portion of the Soviet Union, the Belorussian (or Byelorussian) Soviet Socialist Republic (Belorussian S.S.R., or, more popularly, Belorussia, formerly often referred to in English as White Russia) combines a high density of settlement in its central portion with virtually unpeopled expanses of swamp and forest. One of the 15 constituent republics of the Soviet Union, the Belorussian S.S.R. was set up on January 1, 1919, in response to and counteraction against the Belorussian Democratic Republic, established and proclaimed independent on March 25, 1918.

On the northwest, the Belorussian S.S.R. adjoins the Latvian and Lithuanian S.S.R.'s; the Russian S.F.S.R. and the Ukrainian S.S.R. lie to the north and east and south, respectively, and the frontier with Poland is on the west. The total area of the republic is 80,200 square miles (207,600 square kilometres), or about half of the total area occupied primarily by Belorussians in the Soviet Union. By the mid-1970s the republic was the home of about 9,340,000 persons. Its capital, the ancient city of Minsk, is now the industrial powerhouse of the western Soviet Union.

For information on related subjects, see DNEPR RIVER; PRIPET MARSHES; SOVIET UNION; RUSSIA AND THE SOVIET UNION, HISTORY OF; SLAVIC LANGUAGES.

THE LAND

Low relief

Topography. Belorussia gives an impression of flat and sometimes monotonous landscapes, with extensive but low hilly tablelands alternating with plains and lowlands. The highest point, Dzerzhinsky (Dzyarzhynskaya; the spellings given in parentheses are Belorussian) Mountain, is only 1,135 feet (346 metres) above sea level, and the elevation of more than half of the republic is less than 330–660 feet. In the north, gently sloping ridges formed from glacial debris are separated by lake-dotted lowlands. The largest of these ridges, the Belorussian Ridge (Belaruskaya Hrada), runs from the northwest into the centre of the republic, where it widens into the Minsk Upland (Minskaye Ŭzvyshsha), a region covered with sandy loam and light soils. The large Central Berezinskaya Plain (Tsentralna-Byarezinskaya Raŭnina) is well drained and suitable for agriculture.

In the southwest, along the Pripyat (Prypyats) River, there are tracts of marshy land known as the Belorussian Lowland (Belaruskaye Palesse) merging with the swampy Dnepr Lowland (Prydnyaproŭskaya Nizina) in the east (all of which, with adjoining areas in the Ukraine, are commonly called the Pripet Marshes in the West). The structural trough of the Belorussian Lowland collected meltwater from the glaciers of the Pleistocene Epoch together with great quantities of glacial outwash, which were deposited as sands, lake clays, etc. These deposits and the overall lack of relief caused the formation of the extensive swamps of southern Belorussia and adjoining areas of the Ukraine.

Drainage. Rivers and lakes, major features in the Belorussian S.S.R., are used for shipping, floating timber, and generating power. The streams number about 20,800, with a total length of about 56,300 miles (90,600 kilometres), and the lakes total about 10,800, mostly in the north. The southward-flowing Dnepr River and its tributaries—the Pripyat, Berezina (Byarezina), and Sozh—dominate; a small portion of the Bug drains the extreme southwest; and in the north and west, respectively, the Western Dvina and the Neman (Nyoman) flow to the Baltic. The largest lakes are Narach, 31 square miles (79.6 square kilometres), Osveyskoye (Asveyskaye), and Drisvyaty (Drysvyaty).

Soils. Swamp and marsh soils cover about 10 percent of the republic's surface; because these soils are very fertile when drained (having high contents of both phosphorus and decomposed organic matter), land reclamation has high priority with Belorussian agricultural planners. The largest part of the republic, about 60 percent, is a region of podzols, parts of which are still covered by the postglacial coniferous and mixed forests beneath which they formed. The best soils in the podzol zone for agricultural purposes are those of the Central Berezinskaya Plain and the region around Orsha and Mogilyov. Between the podzol and swamp soils in the south is a zone of sandy soils that are utilized principally in the cultivation of potatoes.

Plant and animal life. The forests known as *pushchy*, covering almost a third of the Belorussian S.S.R., not only are of exceptional importance in its economy but also lend variety and character to the flat relief. In the north, conifers abound, pine and fir predominating, interspersed with birch and alder. The silver trunks of the birch trees stand out against the darker background formed by the conifers. Farther south, deciduous trees—oak, hornbeam, and ash—make an appearance, lending further variety to the landscape.

Belovezhskaya Pushcha

The Belovezhskaya (Belavezhskaya) Pushcha, the scenic forest, is a remnant of the forest that covered much of prehistoric Europe. The Belovezhskaya Pushcha Reserve is remarkable as an exceptionally successful example of international cooperation, comprising 185,000 acres (75,000 hectares) in Belorussia and 155,000 acres in Poland; the reserve is administered jointly by the two countries. The oldest preserve on the continent, it harbours rare animals and birds that have long since vanished elsewhere. The European bison, or wisent (sometimes mistakenly referred to as the aurochs), is the pride of the reservation. The animal life of the republic as a whole is also rich, including elk, deer, boar, hare, and squirrel, with occasional wolf, fox, badger, ermine, and marten, and, near water, such valuable fur animals as otter and mink, as well as beaver.

Game birds include black and hazel grouse, partridge, duck, woodcock, and snipe. Numerous fishes are found in the Dnepr Lowland, with carp prevailing in the lakes of the Baltic Basin. Some fish are raised for food in "farms," and lakes are stocked with food fish.

Climate. The proximity of the Baltic influences the climate of the republic, which varies from maritime to continental. The average July temperature is about 64° F (18° C) with humidity high. The winter is mild, with frequent thaws; the average January temperature is 21° F (−6° C). The average annual precipitation amounts to 22–28 inches (550–700 millimetres) and is occasionally excessive; the temperature range, however, is favourable for all temperate-zone crops. Except in the southern swamps, soils are generally fertile, especially in the river valleys.

THE PEOPLE

Composition and distribution. Ethnically, the 9,340,000 persons living in the republic by the mid-1970s were comparatively homogeneous, Belorussians making up 81 percent of the total. Other peoples included 938,000 Russians, 191,000 Ukrainians, 383,000 Poles, 148,000 Jews, and smaller numbers of Tatars, Lithuanians, Gypsies, and Latvians. In population, the Belorussian S.S.R. ranks fifth among the Soviet republics.

The overall density of population—116.5 persons per square mile (45 per square kilometre) in 1975—is relatively high, the central areas being most densely settled and the southern swamplands thinly peopled.

Patterns of settlement

Demographic trends. The annual birth rate in the mid-1970s averaged 15.8 per 1,000 and the death rate 7.9, giving a natural increase of 7.9. As a result of industrialization, city dwellers in 1974 accounted for 49 percent of the whole, compared with less than 25 percent in 1940. There are nearly 100 cities and towns and more than 100 settlements defined as urban-type. Ancient cities such as Minsk, Brest, and Grodno (Hrodna) have been joined by such entirely new centres as Soligorsk (Salihorsk), Novopolotsk (Navapolatsk), and Svetlogorsk (Svetlahorsk). Most cities cluster around centres of heavy industry and main communication routes in central and eastern Belo-

russia. Despite urbanization, rural residents—half of them living in villages numbering up to 30 households—remain in the majority.

THE ECONOMY

Resources. Traditionally, Belorussia was considered to be poor in mineral resources. Intensive prospecting, however, has revealed minerals, notably high-quality petroleum, now extracted in the south. The southeastern part of the Belorussian Lowland section of the Pripet Marshes has the largest oil reserves. The western part of the Belorussian Lowland and the centre of the republic have peat deposits covering, in total, some 6,170,000 acres (2,500,000 hectares). Coal, brown coal, and combustible shale underlie the Pripet Marshes. Salt deposits in the Belorussian Lowland are the second largest in Europe, estimated at more than 22,000,000,000 tons of rock salt and more than 8,300,000,000 tons of potassium salts; they are considered sufficient to meet the needs of neighbouring regions as well as of Belorussia itself. Other deposits include limestone, dolomite, marl, and sand, with the quartz sands of the Gomel (Homel) region a major component in the manufacture of high-quality glass. Refractory clays—used for pipes and tiles—are found near Brest, and deep mineral springs lie around the Minsk, Mogilyov (Mahilyoŭ), and Bobruysk (Babruysk) areas. There are also phosphate reserves and ferrous and nonferrous metal supplies.

Salt deposits

Overall patterns. As a result of the devastation of World War II, agriculture and industry were almost wiped out; an intensive postwar drive was required to restore the economy. The main industries are now engineering, chemicals, woodworking, light manufacturing, and food processing. The main agricultural activities include the raising of cattle (for meat and dairy products), pigs, and poultry and the production of potatoes and flax fibre.

The Belorussian S.S.R. receives from other Soviet republics coal, oil, natural gas, metals, cotton, synthetic rubber, and a variety of machinery. In its turn, it supplies vehicles, agricultural machinery, timber products, and agricultural goods, which are also exported by the Soviet Union to foreign countries.

Industry. Almost two-thirds of the republic's income is provided by industry. Although some industries are based on local resources—agricultural raw materials, timber, oil, peat, and potassium salts—most industrial plants work, at least partly, on imported raw materials and semimanufactured goods. The post-World War II years have witnessed major changes in the industrial pattern. New industries, including tool and instrument making, oil extraction and refining, and synthetic fibre production, have sprung up, and the geographic distribution of industry has been affected.

Engineering equipment. Production of heavy engineering equipment is an important industry in the republic, the largest manufacturing centres being Minsk, Gomel, and Mogilyov. The republic's heavy-duty trucks, tractors and other agricultural machinery, and metal-cutting tools are used throughout the Soviet Union. More specialized products include computers and such consumer goods as wristwatches, radios, television sets, pianos, bicycles, motorcycles, and sewing machines.

Chemical products. The republic's important chemical industry produces, among other products, some 40 percent of Soviet potassium fertilizers, with Soligorsk, Grodno, and Gomel as the main centres. A growing number of factories works on the by-products of oil refining; others turn out various rubber products, paints, and plastics. The Belorussian S.S.R. was the first Soviet republic to produce dimensionally stable glass pipes for transporting hot and cold liquids and gases.

Woodworking. The woodworking industry, drawing on the republic's forest reserves, notably pine, produces matches, plywood, pressboard, and furniture. It also supplies timber for the coal mines of the Donets Basin. The Belorussian S.S.R. also manufactures various types of paper and paperboard and sections of prefabricated houses.

Light industries. Belorussian textile mills produce linen, woollen, cotton, and silk fabrics. Linen mills in the flax-growing areas of the north and northwest are responsible for 10 percent of Soviet linen fabric. Synthetic fibres are also manufactured. Starch, syrup, alcohol, canned foods, and yeast are produced, and many river- and lake-based plants process fish.

Extractive, refining, and power industries. Petroleum deposits were discovered in the Belorussian Lowland in 1964, and by 1974 more than 8,700,000 tons were being removed annually. Oil from the Rechitsa (Rechytsa) and Ostashkovichi (Astashkavichy) fields is pumped, via the Druzhba pipeline, to a refinery in Polotsk. Another refinery at Mozyr (Mazyr) was constructed in 1975. The Belorussian S.S.R. is in second place in peat production among the Soviet republics (2,100,000 tons of briquettes in 1974).

Power generation (24,600,000,000 kilowatt-hours in 1974) is based on local peat and petroleum, as well as on coal and natural gas that are largely imported from the Ukraine.

Agriculture. The Belorussian environment is favourable to crop production, especially fodder crops. Some 60 percent of the total arable land—4,000,000 acres, or 1,600,000 hectares—is under crops, 25 percent is used for hay, and more than 10 percent is reserved for grazing. Crops

The main commodity crops are grain, flax, potatoes, and sugar beets. Grain (predominantly rye and oats) is sown on 44 percent and fodder crops on a third of the cultivated area. Potatoes and vegetables, on 17 percent of the cultivated land, account for about 15 percent of total Soviet and about 4 percent of total world production of those crops.

The temperate climate of northern Belorussia is particularly suited to flax, which predominates among industrial crops and accounts for one-fourth of Soviet production. Hemp, grown mainly in the south, is also important, and large areas are given over to sugar beets and tobacco as well.

Livestock production, based on good pasturage and a substantial acreage in fodder crops, accounts for more than half of the value of Belorussian agricultural output. Cattle raising accounts for about two-thirds of the total and hog production for most of the remaining third. Meat and poultry production predominates in the central regions, with dairying carried on in suburban areas. Potatoes being their chief feed, hogs are raised in central and southern potato-growing regions. Fur farming—Arctic fox, silver fox, and mink—is carried on, and beekeeping is traditional in Belorussia.

Agriculture in the Belorussian S.S.R. is fully collectivized.

Economic regions. *The Minsk region.* This advantageously situated region is the industrialized heart of Belorussia; with more than half of its population city dwellers, it harbours a wide variety of industries. The regional centre, Minsk, on the banks of the Svisloch (Svislach) River, the home of more than 1,100,000 persons in the mid-1970s, accounts for a third of the republic's total industrial output, mostly in heavy industry. Woodworking is also important in the region and includes the production of furniture in Minsk and Molodechno (Maladzechna) and matches in Borisov (Barysaŭ). Textile mills and dairying are also significant. The city of Minsk is the hub of rail, road, and air routes of national and international significance and is a major centre of Belorussian culture, having a dozen institutions of higher education.

The Vitebsk (Vitsebsk) region. Situated in the northeast, across the upper reaches of the Dnepr and Western Dvina, the Vitebsk region is notable for its toolmaking, textile production, and oil refining. It also leads the republic in flax production.

The Mogilyov region. About a third of the Mogilyov region, which lies in the east, is forested, crossed by the Dnepr, Berezina, and Sozh rivers. The main industrial centres are at Mogilyov and Bobruysk, and the region has engineering, chemical, woodworking, and building industries. Agricultural products include flax, grain, and potatoes, and cattle are raised.

The Gomel region. A third of the Gomel region (second to Minsk in size) is covered with peat bogs and swamplands, and one-fifth is occupied by forests. Its health and recreation spas are known throughout the Soviet Union. The region's industry includes production of machinery for land reclamation. Horticulture is carried on, and agricultural products include grain, hemp, flax, and milk. Gomel is the main urban centre.

The Brest region. The flat and largely swampy Brest region lies in the extreme southwest of the republic and is crossed by transport routes linking the Russian S.F.S.R. and the Ukraine with Poland, the German Democratic Republic, and other central European countries. There is specialization in woodworking, light manufacturing, and building. The region's agriculture, aided by a mild climate, emphasizes flax, sugar beets, potatoes, grain, cattle, and hogs.

The Grodno region. Lying in the Neman River Basin in the northwest, the Grodno region is the smallest in the Belorussian S.S.R. A quarter of the area is forested, and timber is floated down the Neman. The region's flatlands produce sugar beets, and cattle and hogs are raised. Fertilizers and leather are also produced.

Transportation. The flat Belorussian landscape has facilitated development of a transportation network. The main railways are the Moscow–Minsk–Brest and Gomel–Minsk–Vilnius lines, as well as a section of the Odessa–Leningrad route. The road network includes the Roslavl–Brest Highway (a section of the Moscow–Warsaw route) and the Vitebsk–Gomel link (part of the Leningrad–Kiev route). Waterways, too, are extensively used, with the Dnepr and its tributaries pre-eminent; Gomel, Bobruysk, Borisov, and Pinsk are river ports. Among several canals is the Dnepr–Bug link, important for the hauling of freight from the German Democratic Republic.

Pipelines

Ukrainian natural gas is pumped more than 1,200 miles from Dashava to the Belorussian S.S.R. through underground pipelines. The trans-European Druzhba oil pipeline also has a major section running through the Belorussian S.S.R.; its Unecha-Polotsk branch extends to the Latvian port of Ventspils.

Air transport plays an important role, with regional and international links. In addition to passenger, mail, and freight traffic, planes serve remote areas as ambulances.

ADMINISTRATIVE AND SOCIAL CONDITIONS

The constitutional framework. The Belorussian S.S.R. is formally an independent republic with a constitution that came into force on February 19, 1937. It is also a charter member of the United Nations, a status it shares only with the Ukraine among the Soviet republics. The highest body of state power is the Supreme Soviet, which is elected for a four-year period and selects from its ranks a Presidium consisting of a president, two deputy presidents, a secretary, and 11 members who carry on its functions between sessions of the Supreme Soviet. The Supreme Soviet also chooses the government, or Council of Ministers, the highest executive and administrative body. The Belorussian S.S.R. is represented by deputies in the federal Soviet of the Union and Soviet of Nationalities. The highest judicial organ in the republic is the Supreme Court, elected by the Supreme Soviet for a five-year period. The procurator general of the Soviet Union appoints the chief procurator of the Belorussian S.S.R. for a five-year term.

Communist Party

Political activity. The Communist Party of Belorussia, which by the mid-1970s had almost 490,000 members and candidate members, is the only political party in the republic, controlling its government and all political, economic, and cultural life. Part of the Communist Party of the Soviet Union (CPSU), it is guided by that body's program and statutes. The Communist Youth League (Komsomol) of Belorussia had 1,210,000 members in 1975. Trade unions (which do not function as bargaining agents between workers and management) at that time had a total membership exceeding 3,890,000.

Education. Literacy is universal and eight-year schooling obligatory. Ten-year secondary obligatory education has been introduced in general and professional schools. Total enrollment in all kinds of schools had reached about 2,000,000 by the mid-1970s. There were then 9,442 schools of general education, with about 1,800,000 pupils, and 130 or so specialized secondary schools with 152,700 students. The 30 institutions of higher learning include the Belorussian V.I. Lenin State University, the Gomel State University, the Belorussian Agricultural Academy in Gorky (Horki; in Mogilyov *oblast*), and medical, pedagogical, technological, and agricultural institutes. There are 164 students in higher education per 10,000 population, as compared with 39 in 1940. The Belorussian S.S.R. Academy of Sciences coordinates the work of 33 scientific research centres and laboratories with more than 12,000 research workers. The academy has an atomic reactor and a computer centre engaged in economic planning and management.

Housing and medical services. World War II caused the destruction of 74 percent of urban housing (and some 1,200,000 village dwellings) and almost all industrial buildings and schools. The situation was restored to normal after the war. Housing construction in cities and towns consists mainly of multi-story prefabricated units, while individual dwellings predominate in the villages.

Health services

Medical services have shown a steady improvement since the war, and there has been a planned growth in the number of places in hospitals, sanatoriums, convalescent centres, and dispensaries, with the total number of beds exceeding 104,000, or 112.1 per 10,000 population, by the mid-1970s. There were at that time 27,400 doctors in the republic, a ratio of 29.4 per 10,000 population. The improved medical services have resulted in an increased average life-span, from 67 years in 1955 to 72 in 1972.

CULTURAL LIFE

Belorussia is a land of ancient and rich culture. Architectural monuments of the early period include the 11th-century Cathedral of St. Sophia in Polotsk, while local architects of the 14th–16th centuries evolved (at Maloye Mozheykovo, or Maloye Mazheykava, and Synkovichi, or Synkavichy) an original design for a church-fortress. The 17th and 18th centuries are notable for the Baroque style of the Jesuit church in Grodno and for a remarkable variety of wood sculptures. Belorussia was also known for original items of applied art, including the 18th-century Slutsk belts, decorated with gold and silver threads.

Literature. Literary activity in Belorussia dates back to the 11th century. In the 12th century, Cyril of Turov preached and wrote his sermons and hymns and was deeply venerated among all the Orthodox Slavs as "the second Chrysostom." Frantsysk Skaryna of Polotsk in the 16th century translated the Bible into Belorussian, publishing lavish editions in Prague (1517–19) and Vilnius (1522–25), the first printed books not only in Belorussia but in all of eastern Europe. The Belorussian poet Simeon Polotsky (Polatsky; Symeon of Polotsk) was the first to bring Baroque versification and the Baroque sermon to Moscow (after 1655).

The classics of modern Belorussian literature include works of the poets Maksim Bahdanovich, Ales Harun, Yanka Kupala, and Yakub Kolas and the fiction writer Maksim Haretski. The last three lived long enough to contribute to Belorussian literature in the Belorussian S.S.R. This contribution includes, among others, such remarkable works as Kupala's collections of poems titled *Heritage* (1922), *The Nameless* (1925), and *1918–1928* (1930) and his play *Natives* (1924); Kolas' long narrative poems *The New Land* (1910–23) and *Symon the Musician* (1911–25) and his trilogy of novels, *On the Crossroads* (1921–54); and Haretski's short novel *The Quiet Current* (1918–30), his diary, *On the Imperialistic War* (1926), and his collection of short stories, *Before the Dawn* (1926).

The literary upheaval of the 1920s brought into Belorussian literature outstanding poets and writers, whose influence is felt in many respects to this day. Among these were the poets Uladzimir Duboŭka and Yazep Pushcha, the novelist Kuzma Chorny, and the fabulist, satirist, fiction writer, and playwright Kandrat Krapiva.

Their most notable works are Duboŭka's trilogy of long poems ("combines," in his terminology), *Circles* (1927), *And the Purple Sails Unfurled* (1929), and *Storm the Outposts of the Future* (1930); Pushcha's long poems *Song of the War* (1927), *Shadow of the Consul* (1928), and *Gardens of the Winds* (1929); Chorny's novels *Sister* (1927), *The Soil* (1928), and *Lyavon Bushmar* (1929); and Krapiva's *Fables* (1927), long poem, *Shkiruta* (1928), collection of short stories, *Folks, Neighbours* (1928), novel, *Myadzvedzichy* (1933), and play, *Who Laughs Last* (1939).

Contemporary writers

After the re-annexation of western Belorussia from Poland in 1939, the poet Maksim Tank entered Belorussian literature in the Belorussian S.S.R. and became one of its most outstanding contemporary poets; his best works include the long poems *Narach* (1937), *Kalinouski* (1938), and *Yanuk Syaliba* (1943). Among contemporary Belorussian fiction writers the most outstanding are Vasil Bykau, whose short novels *Alpine Ballad* and *The Ordeal* have been translated into English (1966 and 1972, respectively), as have some of the short stories of Yanka Bryl (*Short Stories,* 1957?).

The poets Pyatrus Brouka, Arkadz Kulashou, and Pimyen Panchanka, along with the novelists Ivan Shamyakin and Ivan Melezh and the playwright Andrey Makayonak, are the best representatives of Socialist Realism in Belorussian literature.

In 1973 the Union of Writers of Belorussia had 293 members. A special publishing house, Mastatskaya Litaratura (Belles-Lettres), was opened in 1972 to publish works in Belorussian in that category. Two monthly literary journals, *Polymya* ("Flame") and *Maladosts* ("Youth"), are published in Belorussian and a third, *Neman,* in Russian.

Music. The years since World War II have witnessed the intensive development of music in the Belorussian S.S.R. Notable composers include Dzmitry Lukas (the opera *Kastus Kalinouski,* 1947), Ryhor Pukst (the operas *Masheka,* 1947, and *Marynka,* 1955, and several symphonies), Yaŭhen Hlebaŭ (the opera *Your Spring,* 1963, and the ballet *Alpine Ballad,* 1967), Yaŭhen Tsikotski (the operas *Mikhas Padhorny*, 1939–57, and *Alesya,* 1944–67), and Yury Semyanyaka (the operas *Thorny Rose,* 1960, *When the Leaves Fall,* 1968, and *Star Venus,* 1970). The republic has a conservatory of music and a philharmonic society. Belorussian folk-music companies have made traditional music popular both in the republic and farther afield, the ensemble Pyesnyary (Songsters) being especially popular.

Theatre, film, and broadcasting. The Belorussian State Theatre of Opera and Ballet in Minsk and two Belorussian state dramatic theatres, in Minsk and Vitebsk, are functioning, as well as a dozen professional drama companies and a film studio. A television centre began operating in Minsk in 1956; Belorussian television is broadcast on three channels, totalling more than 30 hours of programming a day. Radio broadcasts consist of four programs, republic-wide and local, totalling 67 program hours a day. An hour of programming is broadcast daily for Belorussians abroad.

PROSPECTS

The manufacture of equipment remains the republic's key industry, but its pattern is changing appreciably, bearing added emphasis on such precision products as instruments, as well as on consumer goods. The chemical industry remains important, with a new development under way in the mid-1970s in the production of synthetic by-products of oil refining.

In agriculture, more effort is being put into the draining of the swampy regions, notably the Belorussian Lowland, to increase crop and fodder-crop area and, in turn, to increase meat production. The modernization of the Dnepr–Bug Canal will improve transport links. It is hoped that these measures will lead to a more effective use of labour resources, as well as of the natural potentialities of the republic, and result in a rise in living standards.

(M.I.R./A.Ad.)

Ben Bella, Ahmed

The best known and most outstanding of the *chefs historiques* of the Algerian War of Independence against France (1954–62), Ahmed Ben Bella played an important part in organizing Algeria's armed struggle for independence and, as the first prime minister (1962–63) and first elected president of the Algerian republic (1963–65), steered his country toward a socialist economy.

Pix



Ben Bella, 1962.

Ben Bella was born on December 25, 1918—the son of a farmer and small businessman—at Maghnia (Marnia) in the department of Oran. There, he successfully completed his early studies at the French school and continued his education in the neighbouring city of Tlemcen, where he first became aware of racial discrimination and also mingled with the fringes of the nationalist movement.

He was conscripted into the French Army in 1937, served in World War II, and was awarded the Croix de Guerre (1940) and the Médaille Militaire (1944). On his return to Maghnia, Ben Bella resumed his nationalist activities—refusing to be intimidated by the confiscation of his farm by the French authorities. He left Maghnia, joined Messali Hadj's underground movement, and soon became one of the "Young Turks" who, after the rigged election of governor Marcel-Edmond Naegelen (1948), considered as illusory any hope of achieving independence democratically. He founded with his friends within Messali Hadj's party the Organisation Spéciale (OS), whose aim was to take up arms as quickly as possible.

After robbing the post office at Oran to obtain funds (1950), Ben Bella was sentenced to prison for eight years but escaped after only two. He went underground again and moved to Egypt, where he was soon promised help by the revolutionary supporters of Gamal Abdel Nasser.

In November 1954 Ben Bella and the Algerian émigré leaders resident in Egypt, having met secretly in Switzerland those leaders still living in Algeria, came to two major decisions: to create the Front de Libération Nationale (FLN) and to order an armed insurrection against the French colonists.

Capture and imprisonment

Ben Bella played an important political role in the collective leadership of the FLN, simultaneously organizing the buying and shipment of foreign arms to Algeria. In 1956 he escaped two attempts on his life, one at Cairo and the other at Tripoli. In the same year he was arrested in Algiers by the French military authorities, acting on their own recognizance, at the very moment that he was negotiating peace terms with the French premier, Guy Mollet.

His captivity (1956–62) kept him apart from those errors of military conduct committed by the FLN and, when he was liberated after the Évian Agreements were signed in 1962, his reputation was intact.

The situation in independent Algeria was at that time chaotic. The leaders of the FLN had formed a provisional government (Gouvernement Provisoire or GPRA) of conservative trend, while the party's congress at Tripoli had elected a socialist-oriented government at the end of the war. It was this latter "Bureau Politique" that Ben Bella ran.

The intervention on his behalf of Colonel Houari Boumedienne, chief of the Armée de Libération Nationale (ALN), assured both the success of the Bureau Politique and of Ben Bella, who was elected unopposed and with an immense majority to the presidency of the Algerian republic in 1962.

Actions as president of Algeria

He re-established order in a country disorganized both by the massive departure of French colonists and by the clashes of armed groups. He created a state out of nothing and consecrated one-quarter of the budget to national education. Above all else he inaugurated—under the title autogestion—a series of important agrarian reforms, among which was the nationalization—but not direct state control—of the former French colonists' huge farms.

He allied himself with the anti-Zionist Arab states, developed cultural and economic relations with France, and quarrelled with the United States over his sympathy for Fidel Castro of Cuba. He also extricated himself, to some advantage, from an important border dispute with Morocco.

His humane, merciful, and youthful method of government pleased the Algerian people. But the effects of his policies were not always so beneficial as his generous intentions. Either through lack of time, political lucidity, or planning, Ben Bella governed from day to day in a series of improvised acts, some of which—like his appeal to Algerian women to donate their jewellry to the state—were more spectacular than useful.

He was unable to restore the FLN, nor was he able to win for it that popular support that would have made it capable of checking Boumedienne.

On June 19, 1965, he was surrounded by tanks at the Villa Joly and deprived of power. Thenceforth his internment became a mystery that may never be solved. Though in June 1971, the Algerian radio announced that he was living freely, but under supervision, at Blida, no impartial witness has ever been able to substantiate the fact.

BIBLIOGRAPHY. ROBERT MERLE, *Ahmed Ben Bella* (1965; Eng. trans., 1967), is the only authorized biography.

(Ro.M.)

Benedict of Nursia, Saint

St. Benedict of Nursia was the father of Western monasticism and the founder of the monastery at Monte Cassino. He formulated the Rule, or regulations governing monastic life and discipline, which from the mid-7th century onward was adopted by ever-widening circles of Frankish and German monasteries. In 1964, in view of the work of monks following the Benedictine Rule in the evangelization and civilization of so many European countries in the Middle Ages, Pope Paul VI proclaimed him the patron saint of all Europe.

Life. The only authority for the facts of Benedict's life is book 2 of the *Dialogues* of St. Gregory the Great (Pope Gregory I), who said that he had obtained his information from four of Benedict's disciples. Though Gregory's work includes many signs and wonders, his outline of Benedict's life may be accepted as historical. He gives no dates however. Benedict was born about 480 of good family at Nursia near Spoletium (Spoleto in Umbria) and was sent by his parents to Roman schools. His life spanned the decades in which the decayed imperial city became the Rome of the medieval papacy. In Benedict's youth Rome under Theodoric still retained vestigial remains of the old administrative and governmental system, with a Senate and consuls. In 546 Rome was sacked and emptied of inhabitants by the Gothic king Totila, and when the attempt of the emperor Justinian I to reconquer and hold Italy failed, the papacy filled the administrative vacuum and shortly thereafter became the sovereign power of a small Italian dominion virtually independent of the Eastern Empire.



St. Benedict of Nursia, detail of a polyptych, tempera on wood, by Segna di Buonaventura, early 14th century. In the Metropolitan Museum of Art, New York.

By courtesy of the Metropolitan Museum of Art, New York, Gift of Reinhardt and Co., 1924

Benedict thus served as a link between the monasticism of the East and the new age that was dawning. Shocked by the licentiousness of Rome, he retired as a young man to Enfide (modern Affile) in the Simbruinian hills, and later to a cave in the rocks beside the then existing lake near the ruins of Nero's palace above Subiaco, 40 miles east of Rome in the foothills of the Abruzzi. There he lived alone for three years, furnished with food and monastic garb by Romanus, a monk of one of the numerous monasteries nearby.

When the fame of his sanctity spread, he was persuaded to become abbot of one of these monasteries. His reforming zeal was resisted, however, and an attempt was made to poison him. He returned to his cave, but again disciples flocked to him and he founded 12 monasteries, each with 12 monks, with himself in general control of all. Patricians and senators of Rome offered their sons to become monks under his care, and from these novices came two of his best-known disciples, Maurus and Placid. Later, disturbed by the intrigues of a neighbouring priest, he left the area, while the 12 monasteries continued in existence.

Foundation of monastery at Monte Cassino

A few disciples followed him south, where he settled on the summit of a hill rising steeply above Cassino, halfway between Rome and Naples. The district was still largely pagan, but the people were converted by his preaching. His sister Scholastica, who came to live nearby as the head of a nunnery, died shortly before her brother. The only certain date in Benedict's life is given by a visit from the Gothic king Totila about 542. His feast day is kept by monks on March 21, the traditional day of his death, in about 547, and by the Roman Catholic Church in Europe on July 11.

His character, as Gregory points out, must be discovered from his Rule, and the impression given there is of a wise and mature sanctity, authoritative but fatherly, and firm but loving. It is that of a spiritual master, fitted and accustomed to rule and guide others, having himself found his peace in the acceptance of Christ.

Rule of St. Benedict. Gregory, in his only reference to the Rule, described it as clear in language and outstanding in its discretion. Benedict had begun his monastic life

as a hermit, but he had come to see the difficulties and spiritual dangers of a solitary life, even though he continued to regard it as the crown of the monastic life for a mature and experienced spirit. His Rule is concerned with a life spent wholly in community, and among his contributions to the practices of the monastic life none is more important than his establishment of a full year's probation, followed by a solemn vow of obedience to the Rule as mediated by the abbot of the monastery to which the monk vowed a lifelong residence.

The monastic constitution

On the constitutional level, Benedict's supreme achievement was to provide a succinct and complete directory for the government and the spiritual and material well-being of a monastery. The abbot, elected for life by his monks, maintains supreme power and in all normal circumstances is accountable to no one. He should seek counsel of the seniors or of the whole body but is not bound by their advice. He is bound only by the law of God and the Rule, but he is continually advised that he must answer for his monks, as well as for himself, at the judgment seat of God. He appoints his own officials—prior, cellarer (steward), novice master, guest master, and the rest—and controls all the activities of individuals and the organizations of the common life. Ownership, even of the smallest thing, is forbidden. The ordering of the offices for the canonical hours (daily services) is laid down with precision. Novices, guests, and the sick, readers, cooks, servers, and porters all receive attention, and punishments for faults are set out in detail.

Influence of the Rule of St. Benedict

Remarkable as is this careful and comprehensive arrangement, the spiritual and human counsel given generously throughout the Rule is uniquely noteworthy among all the monastic and religious rules of the Middle Ages. Benedict's advice to the abbot and to the cellarer, and his instructions on humility, silence, and obedience have become part of the spiritual treasury of the church, from which not only monastic bodies but also legislators of various institutions, have drawn inspiration.

St. Benedict also displayed a spirit of moderation. His monks are allowed clothes suited to the climate, sufficient food (with no specified fasting apart from the times observed by the Roman church), and sufficient sleep (7½–8 hours). The working day is divided into three roughly equal portions: five to six hours of liturgical and other prayer; five hours of manual work, whether domestic work, craft work, garden work, or field work; and four hours reading of the Scriptures and spiritual writings. This balance of prayer, work, and study is another of Benedict's legacies.

All work was directed to making the monastery self-sufficient and self-contained; intellectual, literary, and artistic pursuits were not envisaged, but the presence of boys to be educated and the current needs of the monastery for service books, Bibles, and the writings of the Church Fathers implied much time spent in teaching and in copying manuscripts.

Benedict's discretion is manifested in his repeated allowances for differences of treatment according to age, capabilities, dispositions, needs, and spiritual stature; beyond this is the striking humanity of his frank allowance for weaknesses and failure, of his compassion for the physically weak, and of his mingling of spiritual with purely practical counsel. In the course of time this discretion has occasionally been abused in the defense of comfort and self-indulgence, but readers of the Rule can hardly fail to note the call to a full and exact observance of the counsels of poverty, chastity, and obedience.

Until 1938 the Rule had been considered as a personal achievement of St. Benedict, though it had always been recognized that he freely used the writings of the Desert Fathers, of St. Augustine of Hippo, and above all of John Cassian. In that year, however, an opinion suggesting that an anonymous document, the "Rule of the Master" (*Regula magistri*)—previously assumed to have plagiarized part of the Rule—was in fact one of the sources used by St. Benedict, provoked a lively debate. Though absolute certainty has not yet been reached, a majority of competent scholars favours the earlier composition of the "Rule of the Master." If this is accepted, about one-third of Benedict's Rule (if the formal liturgical chapters are excluded) is derived from the Master. This portion contains the prologue and the chapters on humility, obedience, and the abbot, which are among the most familiar and admired sections of the Rule.

Yet, even if this be so, the Rule that imposed itself all over Europe by virtue of its excellence alone was not the long, rambling, and often idiosyncratic "Rule of the Master." It was the Rule of St. Benedict, derived from various and disparate sources, that provided for the monastic way of life a directory, at once practical and spiritual, that continued in force after 1,500 years.

BIBLIOGRAPHY

Life and spirit: O.J. ZIMMERMANN and B.R. AVERY, *Life and Miracles of St. Benedict* (1949), an Eng. trans. of bk. ii of the *Dialogues* of St. Gregory the Great; ILDEFONS HERWEGEN, *Der heilige Benedikt: ein Charakterbild* (1917; Eng. trans., *St. Benedict*, 1924); and J. MCCANN, *St. Benedict* (1937), good studies by competent scholars; E.C. BUTLER, *Benedictine Monachism*, 2nd. ed. (1924, reprinted 1962), the best account of the Rule and spirit of Benedict.

Rule: Latin-English editions by O. HUNTER-BLAIR, *The Rule of Our Most Holy Father St. Benedict*, 4th ed. (1934); and by J. MCCANN, *The Rule of St. Benedict, in Latin and English* (1952); detailed commentary by P. DELATTE, *Commentaire sur la règle de saint Benoit, par l'abbé de Solesmes* (1912; Eng. trans., *The Rule of Saint Benedict*, 1921). For controversy on authorship, see A. DE VOGUE, *La Règle du Maître*, 3 vol. (Latin text, French trans., excellent introduction and notes, 1964–65). For a short account, see D. KNOWLES in *Great Historical Enterprises* (1963). The standard critical (but also much criticized) Latin text of the Rule is R. HANSLIK, *Regula* (1960).

(M.D.K.)

Bengal, Bay of

The Bay of Bengal, lying roughly between latitude 5°–22° N and longitude 80°–95° E, forms a relatively shallow embayment of the northeastern Indian Ocean. It occupies an area of 839,000 square miles (2,172,000 square kilometres) and is bordered by India and Sri Lanka (Ceylon) on the west, Bangladesh to the north, and Burma and the northern part of the Malay Peninsula to the east. According to the definition of the International Hydrographic Bureau, the southern boundary extends from Dondra Head at the southern end of Sri Lanka to the northern tip of Sumatra. The bay is about 1,000 miles wide, with an average depth of more than 2,600 feet. Maximum depth is 14,764 feet (4,500 metres). A number of large rivers, namely, the Ganges and Brahmaputra on the north, the Irrawaddy on the east, and the Godāvari, Mahānadi, Krishna, and Cauvery on the west, flow into the Bay of Bengal. The Andaman and Nicobar groups are the only islands. Among the principal ports on the Bay are the Indian ports of Calcutta, Cuddalore, Kākināda, Machilīpatnam, Madras, Paradip, and Vishākhapatnam.

A number of expeditions have traversed the area since the latter part of the 19th century, and, in connection with the International Indian Ocean Expedition, the "Vityaz" from the Soviet Union, and the "Pioneer" and "Anton Bruun" from the United States, have more recently carried out detailed investigations. Extensive work has been done on the western coast of the bay by research workers of Andhra University.

Physiography. The Bay of Bengal is bordered to the north by a continental shelf 100 miles wide, but narrower to the south, and by slopes of varying gradient on the west, north, and northeast. Except for the submarine canyons around Sri Lanka, the troughlike valley off the Ganges Delta, and suspected north–south turbidity channels, the deep floor of the bay until recently was believed to be occupied by a vast plain sloping to the south and in places dissected by underwater valleys, perhaps caused by turbidity currents. As a result of the International Indian Ocean Expedition, however, the oceanography is now better understood, and many new physiographic features—mountain chains, submarine canyons, and deep channels—have been brought to light. One of the main submarine features is the north–south Indonesian Trench

Bottom topography

near the Nicobar-Sumatra mainland, which extends into the bay at a maximum depth of 14,800 feet. A large submarine canyon beginning at the head of the bay and called "Swatch of No Ground" extends across the continental shelf for 100 miles with a uniform breadth of eight miles. It appears to begin at about 60 to 180 feet and to cut into the shelf to depths of as much as 3,000 feet below the adjacent shelf level.

In 1963 the Andhra, Mahadevan, and Krishna canyons were discovered off the Andhra coast. A number of other canyons have been identified off the Kākināda–Madras (Coromandel) coast and named after rivers in the vicinity of their locations, namely, Swarnamukhi Canyon, Pennar Canyon, Madras Canyon, Nagarjuna Canyon, Godāvari Canyon, Gautami Canyon, and another south of Puri. Some of them are believed to have been formed during the Pleistocene (10,000 to 2,500,000 years ago) and to have been maintained since then by progressive slumps, density flows, and slow creep.

Hydrography. A unique feature of the bay is the extreme variability of its physical properties. Temperature in the offshore areas, however, is very uniform at all seasons, with decreasing temperatures toward the north. Surface densities are considerably greater in spring than in fall. The sea presents alternately slick and ruffled surfaces over shallow internal waves all along the east-coast shelf. Surface movements of the waters change direction with the season, the northeast monsoon giving them a clockwise circulation, the southeast monsoon a counterclockwise circulation. Severe storms occur at the change of monsoon, particularly to the south in October.

In addition to water-level changes resulting from waves and tide, the average sea level varies throughout the year.

Marine life and bottom deposits. Near its shores the Bay of Bengal is rich in phytoplankton and the related zooplankton. The coastal strips near the outlets of the Vamsadhāra, Nāgāvali, Vasishta Godāvari, and Vainateyam Godāvari rivers have deposits of heavy mineral sands, especially rich in manganese. The source of the manganese has been traced to the rocks in the drainage basins of the rivers emptying their waters and sediment load into the bay. The amount of organic matter present in the continental-shelf sediment of the northern part of the east coast is poor compared with the world's average for nearshore sediments.

Mineral assemblages

On the basis of the available data, the Bay of Bengal can be divided into two areas, one characterized by clay minerals such as illite and kaolinite (Andaman and Nicobar Islands) and the other containing montmorillonite—another clay mineral characterized by swelling in water (eastern and southern parts of the bay). These minerals appear to be derived mainly from the Indian peninsula and from the Himalayas, where they occur abundantly in sedimentary rocks and soils of the river basins. Modern geophysical methods of exploration employed in offshore prospecting are likely to open new vistas into the bay's structure and possible mineral resources under the ocean bed.

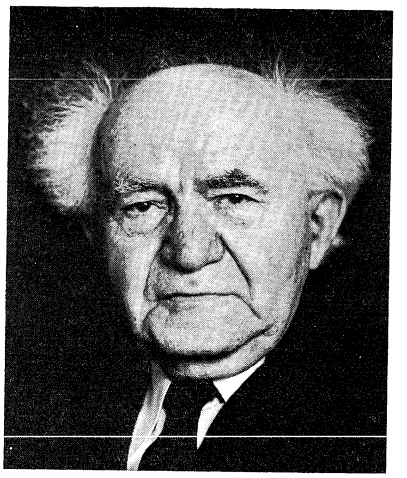
BIBLIOGRAPHY. E.C. LA FOND, "Bay of Bengal," in R.W. FAIRBRIDGE (ed.), *The Encyclopaedia of Oceanography* (1966), deals with oceanography, physical properties, chemistry, geology, submarine canyons, and sediments. The same author's "Andhra, Mahadevan and Krishna Submarine Canyons and Other Features of the Continental Slopes off the East Coast of India," *J. Indian Geol. Union*, vol. 1, no. 1 (1964), deals with the discovery of the above-mentioned canyons and details of profiles traversed. RAO C. BORRESHWARA and E.C. LA FOND, "Study of the Deposition of Heavy Mineral Sands at the Confluences of Some Rivers Along the East Coast of India," *Andhra Univ. Mem. Oceanog.*, vol. 2 (1958), describes the work done along the coastal strips near the confluences of some rivers to study the processes by which heavy mineral sands were deposited and concentrated.

(S.B.)

Ben-Gurion, David

David Ben-Gurion, statesman, political leader, was the first prime minister and chief architect of the State of Israel, whose declaration of independence he delivered on May 14, 1948. His charismatic personality and fighting spirit won him the adoration of the masses, and after his retirement from the government and, later, from the Knesset (the Israeli house of representatives), he was revered as the "Father of the Nation."

Horst Tappe—EB Inc.



Ben-Gurion.

David Ben-Gurion was born David Gruen in Plonsk, Poland, on October 16, 1886, the son of Victor Gruen, one of the local leaders of the "Lovers of Zion," a movement that was disseminating among the oppressed Jews of eastern Europe the idea of the return to their original homeland of Israel. Zionism fascinated the young David Gruen, and he became convinced that the first step for the Jews who wanted to revive Israel as a nation was to immigrate to Palestine and settle there as farmers. In 1906, the 20-year-old Gruen arrived in Palestine and for several years worked as a farmer in the Jewish agricultural settlements in the coastal plain and in Galilee, the northern region of Palestine. There he adopted the ancient Hebrew name Ben-Gurion. Suffering all the hardships of the early pioneers, including malaria and hunger, he never lost sight of his goal. It was owing to his efforts that the 1907 convention of his Zionist socialist party, ha-Po'el ha-Tza'ir ("The Young Worker"), included the following declaration in its platform: "The party aspires to the political independence of the Jewish people in this land."

Emigration to Palestine

With the outbreak of World War I, the Turkish governors of Palestine, their suspicions aroused by his Zionist activity, arrested Ben-Gurion and expelled him from the Ottoman Empire. During the height of the war, he travelled to New York, where he met and eventually married the Russian-born Pauline Munweis. In the last stages of World War I, the British supplanted Turkish rule in the Middle East; and with this change the Jewish settlers and their friends and supporters abroad began to realize that Zionism could rely for future assistance on Britain as well as on the wealthy and influential segments of U.S. Jewry. After the British government published the Balfour Declaration on November 2, 1917, which promised the Jews a "national home" in Palestine, Ben-Gurion enlisted in the British army's Jewish Legion and sailed back to the Middle East to join the war for the liberation of Palestine from Ottoman rule.

The British had already defeated the Turks when the Jewish Legion reached the battlefield, and when Britain received the mandate over Palestine, the work of realizing the "Jewish national home" had begun. For Ben-Gurion, the "national home" was a step toward political independence. To implement it, he called for accelerated Jewish immigration to Palestine in the effort to create a Jewish nucleus that would serve as the foundation for the establishment of a Jewish state. That nucleus was the Histadrut—the confederation of Jewish workers in Palestine founded in 1920 by Ben-Gurion—who was elected its first secretary general—and his colleagues. The His-

The Histadrut

tadrut rapidly became a central force in social, economic, and even security affairs, attaining the position of a "state within a state." Ten years later, in 1930, a number of labour factions united and founded Mapai, the Israeli Workers Party, with Ben-Gurion at its head. In 1935 he was elected chairman of the Zionist Executive, the highest directing body of world Zionism, and head of the Jewish Agency, the movement's executive branch.

As the Jewish settlement strengthened and deepened its roots in Palestine, anxiety mounted among the Palestinian Arabs, resulting in violent clashes between the two communities. In 1939 Britain radically changed its Middle East policy, abandoning its sympathetic stand towards the Jews and adopting a pro-Arab line that led to severe restrictions on Jewish immigration and settlement in Palestine. Ben-Gurion reacted by calling upon the Jewish community to rise against England, thus heralding the decade of "fighting Zionism." On May 12, 1942, he assembled an emergency conference of U.S. Zionists in New York; the convention decided upon the establishment of a Jewish commonwealth in Palestine after the war. At the end of World War II, Ben-Gurion again led the Jewish community in its successful struggle against the British mandate; and in May 1948, in accordance with a decision of the United Nations General Assembly, with the support of the United States and the Soviet Union, the State of Israel was established.

Prime minister

David Ben-Gurion was elected prime minister and minister of defense. Through internal political struggles that incensed both the right and the left, he succeeded in breaking up the underground armies that had fought the British and in fusing them into a national army, which became a model and symbol of the maturing Israeli nation and an effective force against the invading Arab armies from Syria, Jordan, Iraq, and Egypt. Although the fighting ended with an Israeli victory, the Arab leaders refused to enter into formal peace negotiations with the Jewish state.

Ben-Gurion viewed the newborn state as the direct continuation of Jewish history that, in his opinion, had been interrupted 2,000 years earlier when the Roman legions had crushed the Hebrew freedom fighters and banished the Jews from Palestine. He saw the Jews' period of exile as a prolonged interlude in the history of Israel and declared that they had now regained their rightful home. In order to strengthen and develop the young nation, Ben-Gurion presented the people of Israel with a series of challenges: the absorption of mass immigration from all over the world; the assimilation of newcomers of diverse communities and backgrounds; the creation of a unified public education system; the settlement of the desert lands. In his foreign policy, he adopted an independent and pragmatic course. He used to say: "What matters is not what the Gentiles will say, but what the Jews will do." His defense policy was firm, and he answered violations of the cease-fire agreements by neighbouring Arab states with military reprisals. His stronghanded policy inspired little sympathy for him from the governments of the United States and Britain, for both feared that their support of Israel would injure their standing with the oil-exporting Arab governments. They preferred more moderate leaders such as Chaim Weizmann, first president of Israel, and Moshe Sharett, who was elected prime minister for a brief term (1953–55) when Ben-Gurion temporarily retired from office. Striving to gain a foothold in the Middle East, the U.S.S.R. alienated Israel by providing the Arabs with vast quantities of arms. At that time, Ben-Gurion found an ally in France. During the war in Algeria, France encountered the opposition of the united Arab front, led by Egyptian President Nasser, and consequently drew closer to Israel, supplying her with considerable amounts of military equipment; when Nasser nationalized the Suez Canal in July 1956, French initiative brought Israel to join the Franco-British military campaign against Egypt. On October 29, 1956, following a secret visit to France and a meeting with French and British leaders, Ben-Gurion ordered the army to take over the Sinai Peninsula, while France and Britain were making an abortive attempt to seize the canal. Israel subsequently withdrew from Sinai after having been assured freedom of navigation in the Strait of Tiran and de facto peace along the Egyptian-Israeli border, which was to be supervised by a special United Nations force.

Following the Sinai Campaign, Israel entered a period of diplomatic and economic prosperity. Ben-Gurion continued as head of government until 1963. During his last years of office, he initiated several plans (which proved fruitless) for secret talks with Arab leaders with a view to establishing peace in the Middle East. In June 1963 Ben-Gurion unexpectedly resigned from the government for unnamed "personal reasons." His move apparently resulted in part from the bitter internal controversy between his supporters and his rivals in the party, who rose against him for the first time because of the political implications of the 1954 "Lavon Affair," involving Israeli-inspired sabotage of U.S. and British property in Egypt. The affair led Ben-Gurion in 1965 to leave Mapai with a number of his supporters and to found a small opposition party, Rafi, at the head of which he fought, with little success, against his successor, Levi Eshkol.

Resignation and retirement

In 1970 Ben-Gurion retired from the Knesset and from all political activity, devoting himself to the writing of his memoirs in Sede-Boqer, a kibbutz in the Negev desert. He published a number of books, mostly collections of speeches and essays. Through most of his life he had also engaged in researches into the history of the Jewish community in Palestine and in the study of the Bible, for it was from the Bible that he drew his ideals. He died in Tel Aviv on December 1, 1973.

BIBLIOGRAPHY. MICHEL BAR-ZOHAR, *Ben Gourion, le prophète armé* (1966; Eng. trans., *Ben-Gurion: The Armed Prophet*, 1967), is based on the private documents, files, and diaries of Ben-Gurion, as well as on extensive interviews with him, his supporters, and rivals. Other works on Ben-Gurion include: BARNET LITVINOFF, *Ben-Gurion of Israel* (1954), a good, serious biography covering his life until 1953; and ROBERT SAINT-JOHN, *Ben-Gurion* (1959), a rather anecdotal biography.

(M.Ba.)

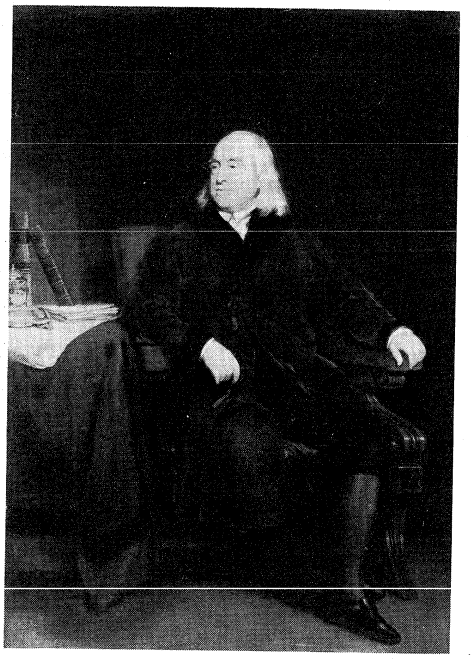
Bentham, Jeremy

English Utilitarian philosopher, economist, and theoretical jurist, Jeremy Bentham exerted a predominant influence on reforming thought of the 19th century. His work, though sometimes crude and superficial, was ahead of its time in that it proposed a scientific discipline for achieving solutions to social problems.

Early life and works. Bentham was born on Feb. 15, 1748, in Red Lion Street, Houndsditch, London, the son of an attorney. At the age of four he is said to have read eagerly and to have begun the study of Latin. Much of his childhood was spent happily at his two grandmothers' country houses. At Westminster School he won a reputation for Greek and Latin verse writing. In 1760 he went to Queen's College, Oxford, and took his degree in 1763. In November, he entered Lincoln's Inn to study law and took his seat as a student in the King's Bench division of the High Court, where he listened with rapture to the judgments of Chief Justice Lord Mansfield. In December 1763 he managed to hear Sir William Blackstone lecture at Oxford, but said that he immediately detected fallacies that underlay the grandiloquent language of the future judge. He spent his time performing chemical experiments and speculating upon the more theoretical aspects of legal abuses rather than in reading law books. On being called to the bar, he "found a cause or two at nurse for him, which he did his best to put to death," to the bitter disappointment of his father, who had confidently looked forward to seeing him become lord chancellor.

Bentham's early publications

Bentham's first book, *A Fragment on Government*, appeared in 1776. The subtitle, "being an examination of what is delivered, on the subject of government in general, in the introduction to Sir William Blackstone's *Commentaries*," indicates the nature of the work. Bentham found the "grand and fundamental" fault of the *Commentaries* to be Blackstone's "antipathy to reform." Bentham's book, written in a clear and concise style different



Bentham, oil painting by H.W. Pickersgill, 1829. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

from that of his later works, may be said to mark the beginning of philosophic radicalism. It is also a very good essay on sovereignty. Lord Shelburne (afterward 1st marquess of Lansdowne), the statesman, read the book and called upon its author in 1781. Bentham became a frequent guest at Shelburne's home. At this period Bentham's mind was much occupied with the publication of his work in French by his admirer Étienne Dumont entitled *Théorie des peines et des récompenses*, 2 vol., in 1811. This work eventually appeared in English as *The Rationale of Reward* (1825) and *The Rationale of Punishment* (1830). In 1785 Bentham started, by way of Italy and Constantinople, on a visit to his brother, Samuel Bentham, an engineer in the Russian navy; and it was in Russia that he wrote his *Defence of Usury* (published 1787). This, his first essay in economics, presented in the form of a series of letters from Russia, shows him as a disciple of the economist Adam Smith, but one who insisted on the extreme logical application of Smith's principles. He argued that every man was the best judge of his own advantage, that it was desirable from the public point of view that he should seek it without hindrance, and that there was no reason for limiting the application of this doctrine in the matter of lending money at interest. His later works on political economy followed the laissez-faire principle, though with modifications. In the "Manual of Political Economy" he gives a list of what the state should and what it should not do, the second list being much longer than the first.

Mature works. Disappointed, after his return to England in 1788, in the hope of making a political career, he settled down to discovering the principles of legislation. The great work on which he had been engaged for many years, *An Introduction to the Principles of Morals and Legislation*, was published in 1789. In this book he defined the principle of utility as "that property in any object whereby it tends to produce pleasure, good or happiness, or to prevent the happening of mischief, pain, evil or unhappiness to the party whose interest is considered." Mankind, he said, was governed by two sovereign motives, pain and pleasure; and the principle of utility recognized this state of affairs. The object of all legislation must be the "greatest happiness of the greatest number." He deduced from the principle of utility that, since all punishment involves pain and is therefore evil, it ought only to be used "so far as it promises to exclude some greater evil."

The fame of the *Principles* spread widely and rapidly. Bentham was made a French citizen in 1792, and his advice was respectfully received in several of the states of Europe and America. With many of the leading men of these countries Bentham maintained an active correspondence. In 1817 he became a bencher (governor) of Lincoln's Inn. The codification of law was one of his chief preoccupations, and it was his ambition to be allowed to prepare a code of laws for his own or some foreign country, but he seems to have underestimated both the intrinsic difficulties of the task and the need for diversity of institutions adapted to the tradition and civilization of different countries. Bentham, however, must be reckoned among the pioneers of prison reform. It is true that the particular scheme that he worked out was bizarre and spoiled by the elaborate detail that he loved. "Morals reformed, health preserved, industry invigorated, instruction diffused" and other similar desiderata would, he thought, be the result if his scheme for a model prison, the "Panopticon," were to be adopted; and for many years he tried to induce the government to adopt it. His endeavours, however, came to nothing; and though he received £23,000 in compensation in 1813, he lost all faith in the reforming zeal of politicians and officials.

Work on prison reforms

In 1823 he helped to found the *Westminster Review* to spread the principles of philosophic radicalism. Bentham had been brought up a Tory, but his experience over the "Panopticon" scheme served to make a democrat of him. As far back as 1809 he had written a tract, *A Catechism of Parliamentary Reform*, that advocated annual elections, equal electoral districts, a wide suffrage, and the secret ballot, which was, however, not published until 1817. He drafted a series of resolutions on reform, based on this tract, which were introduced in the House of Commons in 1818. A volume of his *Constitutional Code*, which he did not live to complete, was published in 1830.

Bentham died in London on June 6, 1832, in his 85th year. In accordance with his directions, his body was dissected in the presence of his friends. The skeleton was then reconstructed, supplied with a wax head to replace the original (which had been mummified), dressed in Bentham's own clothes and set upright in a glass-fronted case. Both this effigy and the head are preserved in University College, London.

Bentham's life was a happy one. He gathered around him a group of congenial friends and pupils, such as the philosophers James and John Stuart Mill, with whom he could discuss the problems upon which he was engaged. These friends, too, practically rewrote several of his books from the mass of rough though orderly memoranda that Bentham himself prepared. Thus the *Rationale of Judicial Evidence*, 5 volumes (1827), was put in its finished state by J.S. Mill and the *Book of Fallacies* (1824) by Peregrine Bingham. The services of Étienne Dumont in recasting as well as translating the works of Bentham were still more important. It is often difficult to distinguish what part of the work is Bentham's and what is due to his assistants.

Assessment. Bentham was less a philosopher than a critic of law and of judicial and political institutions. Unfortunately, he was not aware of his limitations. He tried to define what he thought were the basic concepts of ethics. Most of his definitions are oversimple or ambiguous or both, and his "felicific calculus," a method for calculating amounts of happiness, as even his warmest admirers have admitted, cannot be used. It is not simply impractical but logically absurd, for not even an omniscient God could make the calculations imagined by Bentham. As a moralist and psychologist, Bentham is inadequate; his arguments, though sometimes elaborate, rest too often on insufficient and ambiguous premises. His analyses of the concepts men use to describe and explain human behaviour are too simple. He seems to have believed both that man is completely selfish and that everyone ought to promote the greatest happiness, no matter whose. Not even the formula of which he made so much, "the greatest happiness of the greatest number," has a definite meaning. The opening chapters of *An Introduction*

to the Principles of Morals and Legislation, in which Bentham seeks to explain the ideas fundamental to his philosophy, are remarkably confused. Fortunately, they are more confused than confusing. They spring from simple fallacies easily detected. It must be said in Bentham's favour that he tried to be clear and precise; he did not hide the poverty of his arguments in a fog of rhetoric. He was maladroit but no mystifier.

Bentham as a critic of institutions

As a critic of institutions Bentham was admirable. In his *Rationale of Judicial Evidence* he describes the methods that a court should use to get at the truth as quickly as possible; and in the *Essay on Political Tactics* he describes what he considers the most effective forms of debate for a legislative assembly—an account largely based on the procedure of the House of Commons. In these works and in others Bentham is concerned to discover what makes for efficiency. Though he defines efficiency in terms of happiness, his reader need not do so; or, if he does, he need not think of happiness as Bentham did. Bentham's assumptions about what makes for happiness are often quite ordinary and sensible; the reader can accept them and still insist that happiness is not to be defined in terms of pleasure and is not to be measured. Whatever is excellent, ingenious, and original in Bentham—and there is a great deal of it—need not depend on the "felicific calculus" and "the greatest happiness of the greatest number."

BIBLIOGRAPHY. JOHN BOWRING (ed.), *The Works of Jeremy Bentham*, 11 vol. (1838–43, reprinted 1962), hitherto to the nearest to a complete edition but badly printed; J.H. BURNS (gen. ed.), *The Collected Works of Jeremy Bentham* (1968–), more complete, more accurate, and much easier to read than Bowring's edition; LESLIE STEPHEN, *The English Utilitarians*, vol. 1 (1900, reprinted 1968), still one of the best general accounts of Bentham's life and thought; ELIE HALEVY, *La Formation du radicalisme philosophique*, 3 vol. (1901–04; Eng. trans., *The Growth of Philosophical Radicalism*, 1928), a study of Bentham's philosophy seen in the context of the history of English thought in his time; MARY MACK, *Jeremy Bentham: An Odyssey of Ideas* (1962), the most recent extensive biography; CHARLES W. EVERETT, *The Education of Jeremy Bentham* (1913), the best account of how his mind was formed; S.R. LETWIN, *The Pursuit of Certainty*, pt. 2 (1965), a good recent assessment of the importance of Bentham's ideas in his own day. Other good assessments of Bentham's philosophy are: J.S. MILL's essay on "Bentham" in F.R. LEAVIS (ed.), *Mill on Bentham and Coleridge* (1950); HENRY SIDGWICK, "Bentham and Benthamism in Politics and Ethics," *Fortnightly Review, New Series*, 21:627–652 (1877); JACOB VINER, "Bentham and J.S. Mill: The Utilitarian Background," *American Economic Review*, 39:360–383 (1949); and H.L.A. HART, "Bentham: Lecture on a Master Mind," *Proceedings of the British Academy*, 48: 297–320 (1962).

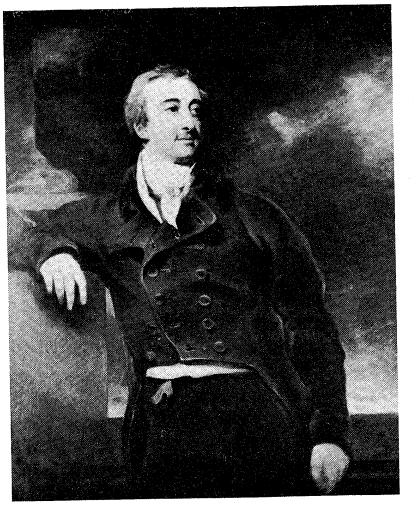
(J.P.Pl.)

Bentinck, Lord William

The governor general of India from 1828 to 1835, Lord William Henry Cavendish Bentinck, an aristocrat in sympathy with many of the liberal ideas of his day, introduced a great number of humanitarian and administrative reforms in both Indian government and society. The innovations effected in his years of office were decisive milestones on the road that was ultimately to lead to independence more than a century later.

The second son of the 3rd duke of Portland, Bentinck was born on September 14, 1774. At the age of 17 he received a commission as ensign in the Coldstream Guards and by 1794 was a lieutenant colonel. Born to wealth and rank, he was a promising, if not outstanding, young officer. Nevertheless, his appointment as governor of Madras in 1803, at the early age of 29, caused surprise.

Although he performed his duties satisfactorily enough, his administration was clouded by his disagreements with his council and was abruptly terminated by the mutiny at Vellore. An unwise order by the commander in chief of the Madras Army had forbidden the native troops to wear their traditional beards and turbans; Bentinck, even more unwisely, would not allow the order to be rescinded. The consequence was a serious mutiny in July 1806, accompanied by attacks on officers and British troops. The outbreak was suppressed with heavy loss of life, and the ill-considered order was finally withdrawn. Bentinck, nevertheless, was held responsible and recalled. Believing he had been treated unjustly, he pressed for the next 20 years for a chance to vindicate his name by service in India.



Bentinck, oil painting by Sir Thomas Lawrence (1769–1830). In the collection of the Duke of Portland.
By courtesy of His Grace the Duke of Portland; photograph, Gordon Hull

The Napoleonic Wars were on, and he was next assigned to Spain, where he commanded a brigade at Corunna, after which he was appointed commander of the British troops in Sicily. Italy was then in the hands of Napoleon, but in Sicily the Bourbon monarchs of Naples still reigned under the protection of the British fleet. Bentinck's orders were to raise a Sicilian Army of 10,000 men to supplement his 5,000 British soldiers and land on the east coast of Spain with his combined forces to assist in the campaign against Napoleon. Had Bentinck been no more than a soldier, his course would have been clear. But he was a man of imagination, a Whig (a liberal) by family tradition, and a radical in the eyes of his contemporaries. Therefore, besides merely raising a Sicilian Army, he engineered the deposition of the Bourbon king —in favour of the heir apparent—as well as the adoption of a liberal Sicilian constitution with a legislative body modelled on the English Parliament. Further, he planned to invade Italy and rally the people not only to expel Napoleon but to set up a constitutional monarch. The British government would never have supported such a plan; in fact, it intended eventually to restore Austrian rule in Italy. The Italian landing did not take place at that time, however, and Bentinck delayed his landing in Spain beyond the date when he was most needed. When he finally did land in Italy, at Genoa in 1814, his liberal proclamations again embarrassed his government, and he was recalled to England in 1815. On his return, he was elected to the House of Commons.

Tour as commander in Sicily

He refused reappointment to the governorship of Madras in 1819, waiting to attain his real ambition—the appointment as governor general of India and governor of Bengal, which came in 1827. Bentinck's immediate instructions were to rescue India from its financial difficulties; at this time the government in India operated on an annual deficit of about £1,500,000. Bentinck soon succeeded in turning the deficit into a surplus of about the same amount. He next turned to judicial and administrative reforms, which included making more administrative and judicial positions available to Indians and improving the salaries and status of Indian judges. Bentinck also made English, instead of Persian, the language of the higher courts and of higher education and arranged for financial aid to colleges, which were to be adapted to the Western models.

Governor general of India

Bentinck showed great courage and humanity by his decision to abolish *satī* (suttee), the custom of burning

widows alive with the corpse of their husbands. Previous governors general had shrunk from prohibiting the custom as an interference in religion and one particularly likely to upset the Indian Army; but Bentinck cut through these hesitations without facing much open opposition. He also was responsible for the measures taken to suppress the murder of unwanted children, human sacrifice, and the *thags* (thugs)—bands of robbers, bound together by oaths and ritual, who murdered unsuspecting travellers in the name of the goddess Kālī. Flogging in the Indian Army was also abolished, long before it ended in the British Army.

Bentinck left India in March 1835 and died in Paris on June 17, 1839. It has been argued that the reforms he initiated and those that followed in the next 20 years—which accelerated the westernization of India—were partly responsible for the Indian Army's Mutiny of 1857. That argument may have little force; his reforms were liberal, not radical, and would eventually have become inevitable. Bentinck was not an original thinker; his philosophical masters were the Utilitarians Jeremy Bentham and James Mill; his practical instructor, especially in the field of education, was the historian Thomas Babington Macaulay, among others. He borrowed useful elements from the creed of his liberal Whig ancestors and of Bentham and combined them in policies that were sensible, practical, and humanitarian. Toward the end of his career he had lost the impetuosity that had characterized his earlier years in Sicily and the tactlessness that had appeared when he first held office in Madras. Though certainly not the most brilliant of the governors general, in solid achievements and in utilizing the resources available to him, he must rank among the most successful.

Assessment

BIBLIOGRAPHY. There is no good biography. D.C. BOULGER, *Lord William Bentinck*, vol. 17 in *Rulers of India* (1892), is below the usual standard of that series. See also JOHN ROSSELLI, *Lord William Bentinck and the British Occupation of Sicily, 1811–1814* (1956). Two large volumes of Bentinck's correspondence, ed. and with an introduction by C.H. PHILIPS (in prep.), to some extent will fill the gap.

(P.Ma.)

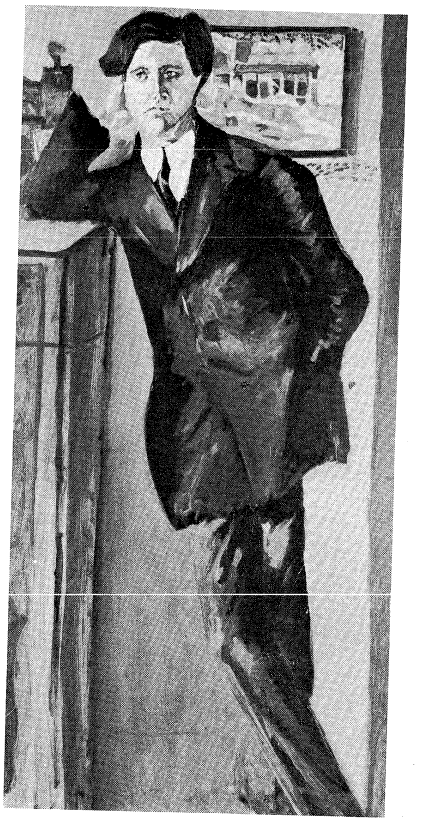
Berg, Alban

The Austrian composer Alban Berg was one of the principal figures responsible for bringing to maturity the atonal style of 20th-century music. His powerful and complex works draw from a broad range of musical resources but are chiefly shaped by a few central techniques: the use of a complex chromatic expressionism, which nearly obscures, yet actually remains within, the framework of traditional tonality; the recasting of classical musical forms with atonal content—*i.e.*, abandoning traditional tonal structure dependent upon a centrally important tone; and a deft handling of the 12-tone approach developed by his teacher, the composer Arnold Schoenberg, as a method of structuring atonal music. Berg dealt with the new medium so skillfully that the classical heritage of his compositions is not obliterated, thus justifying the term frequently applied to him—the "classicist of modern music."

Alban Berg was born in Vienna on February 9, 1885. Apart from a few short musical trips abroad and annual summer sojourns in the Austrian Alps, his life was spent in the city of his birth. At first, the romantically inclined youth leaned toward a literary career. But, as in most Viennese middle class homes, music was regularly played in his parents' house, in keeping with the general musical atmosphere of the city. Encouraged by his father and older brother, Alban Berg began to compose music without benefit of formal instruction. His output consisted of over 100 songs and piano duets, most of which remain unpublished.

Influence of Schoenberg

In September 1904 he met Arnold Schoenberg, an event that decisively influenced his life. The death of Berg's father had left little money for composition lessons, but Schoenberg was quick to recognize Berg's talent and accepted the young man as a free pupil. The musical precepts and the human example provided by Schoenberg shaped Berg's artistic personality for the next six years.



Berg, oil painting by Arnold Schoenberg, *c.* 1900. In the Historisches Museum der Stadt Wien.
By courtesy of the Historisches Museum der Stadt Wien

In the circle of Schoenberg's students, Berg presented his first public performance in the fall of 1907: *Piano Sonata*, Opus 1 (1908). This was followed by *Four Songs*, Opus 2 (1909), and *String Quartet*, Opus 3 (1910), each strongly influenced by the young composer's musical gods, Gustav Mahler and Richard Wagner.

Having come into a small inheritance, Berg married Helene Nahowski, daughter of a high-ranking Austrian officer, in 1911. The Bergs took an apartment in Vienna, where he settled down to devote the remainder of his life to music, although they participated freely in the intellectual life of the city. Among their closest friends were Adolph Loos (1870–1933), one of the pioneers of modern architecture, and the painter Oskar Kokoschka.

A characteristic of Berg's creative activity was the slow, often hesitant, manner in which he gave final form to the musical ideas that, for the most part, were the result of sudden inspiration. This fastidious, perfectionist manner of composing explains his relatively small number of works. In 1912 Berg finished his first work since his student days with Schoenberg, *Five Orchestral Songs*, Opus 4. The inspiration for this composition came from postcard messages addressed to both his friends and foes by the eccentric Viennese poet Peter Altenberg. These sometimes erotic postcard texts were sufficiently nonconformist to prompt Berg to use them as background for even less traditional music than he had composed in the past. But when two of these songs were presented at a concert of the Academic Society for Literature and Music in March 1913, they provoked a near riot, in which performers and audience freely participated.

The genesis of Berg's first work for the stage was a memorable theatrical experience: the performance of Georg Büchner's (1813–37) *Woyzeck*, a drama built around a poor working man who murders his faithless sweetheart and then commits suicide while their child, unable to comprehend the tragedy, plays nearby.

The theme fascinated Berg, but World War I, during which Berg, always in frail health, worked in the War Ministry, prevented him from beginning work on his

The writing of *Wozzeck*

opera, which he, varying the spelling, called *Wozzeck*. He was confronted by the gigantic task of compressing 25 scenes into three acts. Although he managed to write the libretto in 1917, he did not begin composing the score until the war was over. He completed the opera in 1921, dedicating it to Alma Mahler, the widow of Gustav Mahler, the composer and conductor who had dominated Vienna's musical life during Berg's youth.

Wozzeck—perhaps the most frequently performed theatrical work in the atonal idiom—represents Berg's first attempt to deal with social problems within the framework of opera. From numerous statements he made, it is evident that he intended the opera to portray far more than the tragic fate of the protagonist. He wanted, in fact, to make it symbolical of human existence. Musically, its unity stems from large overall symmetries within which are set traditional forms (such as the passacaglia and sonata), excerpts in popular music style, dense chromaticism (use of notes not belonging to the composition's key), extreme atonality, and passing approaches to traditional tonality, all of which function to create a work of notable psychological and dramatic impact. Although it antedates Schoenberg's early 12-tone compositions, the opera also includes a theme using the 12 notes of the chromatic scale.

After 137 rehearsals, *Wozzeck* was presented in its entirety for the first time on December 14, 1925, at the Berlin State Opera, with Erich Kleiber conducting. Critical response was unrestrained. Typical of the prevailing attitude was the reaction of a reviewer in the *Deutsche Zeitung:*

> As I was leaving the State Opera I had the sensation of having been not in a public theatre but in an insane asylum. . . . I regard Alban Berg as a musical swindler and a musician dangerous to the community.

But another critic described the music as "drawn from Wozzeck's poor, worried, inarticulate, chaotic soul. It is a vision in sound."

Upon completion of *Wozzeck*, Berg, who had also become an outstanding teacher of composition, turned his attention to chamber music. His *Chamber Concerto*, Opus 8, for violin, piano, and 13 wind instruments was written in 1925, in honour of Schoenberg's 50th birthday. At the same time, Berg searched for a new opera text. He found it in two plays by the German dramatist Frank Wedekind (1864–1918). From *Erdgeist* ("Earth Spirit") and *Büchse der Pandora* ("Pandora's Box"), he extracted the central figure for his opera *Lulu*. This work

Composition of *Lulu*

engaged him, with minor interruptions, for the next seven years, and its third act remained incomplete at his death. Symbolically and musically complex, and highly expressionistic in idiom, *Lulu* was composed entirely in the 12-tone system.

With the seizure of power by the Nazis in Germany in 1933, Berg lost most of his income. Although, unlike their teacher Schoenberg, Berg and his friend and colleague Anton von Webern were of non-Jewish descent, they, with Schoenberg, were regarded as representatives of "degenerate art" and were increasingly excluded from performances in Germany. The meagre response that Berg's works evoked in Austria caused him particular anguish; abroad, he was considered more and more as the representative Austrian composer, and his works were performed at leading musical festivals.

Berg's last complete work, the *Violin Concerto*, originated under unusual circumstances. In 1935 the American violinist Louis Krasner commissioned Berg to compose a violin concerto for him. As usual, Berg procrastinated at first. But after the death of Manon, the beautiful 18-year-old daughter of Alma Mahler (by then the wife of the architect Walter Gropius), Berg was moved to compose the work as a kind of requiem and to dedicate it to the "memory of an angel"—Manon. Having found his inspiration, Berg worked at fever pitch, in the seclusion of his villa in the Austrian province of Carinthia, and completed the concerto in six weeks. By the time the work was finally presented by Krasner in Barcelona in April 1936, it had become a requiem not only for Manon Gropius but for Alban Berg as well. One of the major violin concerti of the 20th century, it is a work of highly personal, emotional content achieved through the use of 12-tone and other resources—symbolic as well as musical.

In mid-November 1935 he returned, a sick man, to Vienna. Although his mind was completely absorbed in his desire to finish the opera *Lulu*, he had to be hospitalized in December with septicemia; and after a deceptive initial improvement, he died suddenly on December 24.

A man of strikingly attractive appearance and reserved, aristocratic bearing, Berg had also a generous personality that found expression in his correspondence and among his friends. He was an outstanding teacher of composition who encouraged his pupils to undertake significant work of their own. Few honours were accorded Berg in his lifetime, but within a few years after his death, he became widely recognized as a composer who broke with tradition, mastered a radical technique, and blended the two to create, with Schoenberg and Webern, a 20th-century Viennese school of music.

MAJOR WORKS

Orchestral music

Three Pieces for Orchestra, op. 6 (composed 1914); *Chamber Concerto* for piano, violin and 13 wind instruments, op. 8 (1924); *Violin Concerto* (1935).

Chamber music

Piano Sonata, op. 1 (1908); *String Quartet*, op. 3 (1910); *Four Pieces for Clarinet and Piano*, op. 5 (1913); *Lyric Suite*, op. 10 (1925–26).

Vocal music

OPERAS: *Wozzeck*, libretto by Berg, based on a play by Georg Büchner, op. 7 (first performed 1925); *Lulu*, libretto by Berg, based on two tragedies by Frank Wedekind (1937), orchestration of third act incomplete.

VOICE AND ORCHESTRA: *Five Orchestral Songs*, on picture-postcard texts, by Peter Altenberg, op. 4 (composed 1912); *Der Wein*, concert aria with orchestra with text translated from Charles Baudelaire, op. 11 (1929).

VOICE AND PIANO: *Seven Early Songs* (1905–08); *Four Songs*, op. 2 (1909).

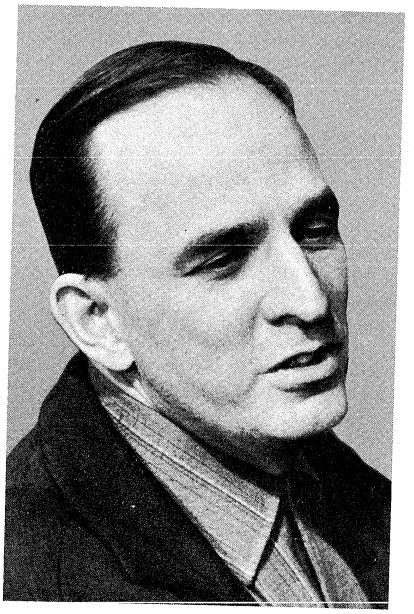
BIBLIOGRAPHY. H.F. REDLICH, *Alban Berg: The Man and His Music* (1957), especially noteworthy for its prodigious use of examples in presenting a basic analysis of the music; WILLI REICH, *The Life and Work of Alban Berg* (1965), an authoritative text by a pupil and intimate friend of Berg; ALBAN BERG, *Briefe an seine Frau* (1965), Berg's letters to his wife—provides important source material.

(W.R.)

Bergman, Ingmar

One of the most eminent and influential contemporary film makers, Ingmar Bergman has established a worldwide reputation for writing and directing films that, in an unmistakably individual style, examine the issues of morality by exploring man's relationship to himself, to others, and to God. His work and the worldwide vogue it enjoyed in the late 1950s and early 1960s introduced many people for the first time to the idea of the total film maker, the writer-director who throughout a sizable body of work uses the medium of film to express his own ideas and perceptions, with as much ease and conviction as artists in earlier generations used the novel or the symphony or the fresco. In addition, the immense international popularity of his films has tended to ensure that Bergman's picture of Sweden and the Swedish temperament is the first and often the only impression received by the outside world; and when other Swedish films seem to present much the same image, it is usually because the influence of Bergman on his Swedish colleagues is so pervasive rather than because his highly personal vision should be taken as an objectively true portrait of his country.

Ernst Ingmar Bergman was born in Uppsala, Sweden, on July 14, 1918. The son of a Lutheran pastor, he has frequently remarked on the importance of his childhood background in the development of his ideas and moral preoccupations. Even when the context of his film characters' sufferings is not overtly religious, they are always implicitly engaged in a search for moral standards of judgment, a rigorous examination of action and motive,



Bergman, 1969.
Svenskt Pressfoto—Keystone

in terms of good and bad, right and wrong, which seems particularly appropriate to someone brought up in a strictly religious home. Another important influence in his childhood was the religious art Bergman encountered, particularly the primitive yet graphic representations of Bible stories and parables found in rustic Swedish churches, which fascinated him and gave him a vital interest in the visual presentation of ideas, especially the idea of evil as embodied in the devil.

Bergman attended Stockholm University, where he studied art, history, and literature. There for the first time he became passionately involved in the theatre and began writing and acting in plays and directing student productions. From these he went on to become a trainee director at the Mäster Olofsgärden Theatre and the Sagas Theatre, where he produced a spectacularly unconventional and disastrous production of the Swedish playwright August Strindberg's *Ghost Sonata*, which was withdrawn after four performances. In 1944 he was given his first full-time job as a director, at Hälsingborg municipal theatre. Also, and more importantly, he met Carl-Anders Dymling, the head of the Svensk Filmindustri. Dymling was sufficiently impressed by him to commission an original screenplay, *Hets* (called *Torment* in the U.S., *Frenzy* in Britain). This was directed by Alf Sjöberg, then Sweden's leading film director, and had an enormous success, both at home and abroad. Largely as a result of this success, Bergman was, in 1945, given a chance to write and direct a film of his own, *Crisis*, and from this point on, his career was under way.

Success in Sweden

The films that Bergman wrote or directed, or both, in the next five years were, if not directly autobiographical, at least very much concerned with the sort of problems that he himself was encountering at that time: the role of the young in a changing society, ill-fated young love, and military service. He directed his first film based on an original screenplay of his own, *Fängelse* (*Prison*), in 1948. It recapitulated all the themes of his previous films in a complex, perhaps overambitious story, built around the romantic and professional problems of a young film director who considers making a film based on the idea that the devil rules the world. While this is not to be taken without qualification as Bergman's message in his early work, it may at least be said that his imaginative world is divided very sharply between the worlds of good and evil, the latter always overshadowing the former, the devil lying in wait at the end of each idyll.

In 1951 Bergman's career in films, like nearly the whole of Swedish filmmaking, came to an abrupt halt as the result of a major economic crisis in Sweden. But in 1952 he returned with two films, *Secrets of Women* and *Monika* (British title, *Summer with Monika*), that marked the beginning of his mature work. He was also appointed to his most important position to date in the theatre, director of the Malmö municipal theatre, where he remained until 1959. This new phase introduced two markedly new characteristics in his work. In subject matter, Bergman, now himself married, returned again and again to the question of marriage. Viewing it from many angles, he examined the ways by which two people adjust to living together, their motives for being faithful or unfaithful to each other, and their reactions to bringing children into the world. At this time Bergman began to gather around him, in his film and stage productions, a faithful "stock company" of actors with whom he worked regularly to give his work and their interpretation of it a manifest consistency and style.

International recognition

In 1955 Bergman had his first great international success with *Smiles of a Summer Night*, a bittersweet romantic comedy-drama in a period setting. In the next few years, a kind of Bergman fever swept over the international film scene: concurrently with the succession of his new films, which included two additional masterpieces, *The Seventh Seal* (1956), a medieval morality, and *Wild Strawberries* (1957), a meditation on old age, all of his early work was shown, and Bergman was universally recognized as one of the most important figures in the contemporary cinema. Indeed, a far wider section of the cultured public became aware of his work than of that of any previous film maker; for the first time a film maker was as widely and as highly regarded as artists in any of the more traditional media.

Inevitably, a reaction set in, though Bergman continued to make films and direct plays with undiminished activity; and his trilogy of films, *Through a Glass Darkly* (1961), *Winter Light* (1962), and *The Silence* (1963), dealing with related themes—the border line between sanity and madness and that between human contact and total withdrawal—was by many regarded as his crowning achievement. About this time, he acquired a country home on the bleak island of Fårö; and the island has provided a characteristic stage for the dramas of most of his more recent films, notably *Persona* (1966), *Hour of the Wolf* (1967), *Shame* (1968), and *The Passion of Anna* (1969), all intimate dramas of inner conflicts involving a small, closely knit group of characters. With *The Touch* (1971), his first English-language film, Bergman returned to an urban setting and more romantic subject matter, though fundamentally the characters in the film's marital triangle are no less mixed up than any in the Fårö cycle of films. Bergman's anguished appraisal of the human situation has lost nothing of its intensity through the years; rather, he has progressively stripped away the distracting decorations in his films to create an abstract drama of man's relation with man and perhaps with God (if he exists). He deals with man's attempt to define his own personality by removing his masks to see if there is a face underneath. The images of the creator as actor and the creator as magician recur throughout Bergman's work. He himself embodies elements of both the thinker and the actor, the preacher and the charlatan; in Bergman they all fuse to create an artist of great force and individuality whose work is always unmistakably his own.

MAJOR WORKS

Crisis (1945); *A Ship to India* (1947); *Port of Call* (1948); *The Devil's Wanton* (1948; also released as *Prison*); *Three Strange Loves* (1949; British title, *Thirst*); *To Joy* (1949); *Illicit Interlude* (1951; British title, *Summer Interlude*); *Secrets of Women* (1952; British title, *Waiting Women*); *The Naked Night* (1953); *A Lesson in Love* (1954); *Dreams* (1955; British title, *Journey Into Autumn*); *Smiles of a Summer Night* (1955); *Wild Strawberries* (1957); *The Seventh Seal* (1956); *Brink of Life* (1958); *The Magician* (1958; British title, *The Face*); *The Virgin Spring* (1960); *The Devil's Eye* (1960); *Through a Glass Darkly* (1961); *Winter Light* (1962); *The Silence* (1963); *All These Women* (1964); *Persona* (1966); *Hour of the Wolf* (1967); *Shame* (1968); *The Passion of Anna* (1969); *The Touch* (1971); *Cries and Whispers* (1972).

BIBLIOGRAPHY. JORN DONNER, *Djävulens ansikte: Ingmar Bergmans filmer* (1962; Eng. trans., *The Personal Vision of Ingmar Bergman*, 1964), a perceptive study of Bergman's themes and images by a fellow film maker; JOHN RUSSELL TAYLOR, *Cinema Eye, Cinema Ear* (1964), contains a critical study of Bergman, filmography, and bibliography; HENRIK SJOGREN, *Ingmar Bergman på teatern* (1968), a complete account of Bergman's theatrical work, with pictures and credits; *Four Screenplays of Ingmar Bergman: Smiles of a Summer Night, The Seventh Seal, Wild Strawberries, The Magician (The Face)* (1960), and *A Trilogy: Through a Glass Darkly, Winter Light, The Silence* (1965), two collections of English translations of Bergman's principal screenplays.

(J.R.T.)

Bergson, Henri

Henri Louis Bergson, the leading French philosopher of the early 20th century, was the first to elaborate what has come to be called a process philosophy, one that stresses the open flow of time. In attacking the pretensions of Positivistic, or narrowly scientific, interpretations and in defending humanistic and spiritual values, Bergson was an exponent of the "revolt against reason." The master of a literary style noted for lucidity and grace, he was one of the most widely read philosophers of the century and exerted an influence that reached far beyond the confines of academic and professional circles.

Parentage and education

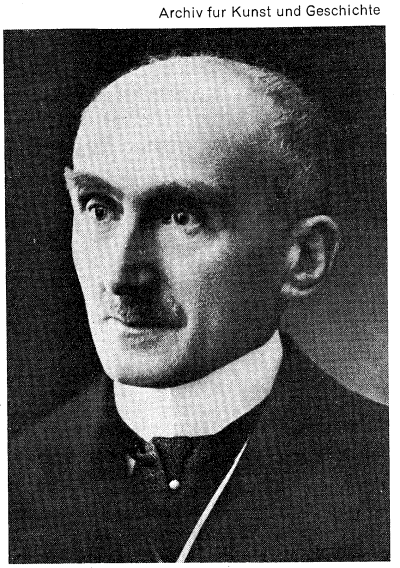
Early years. Bergson was born in Paris on October 18, 1859. His father, a talented musician and at one time director of the Geneva Conservatory, was descended from a rich Polish Jewish family—the sons of Berek, or Berek-son, from which the name Bergson is derived. His mother came from an English Jewish family. Bergson's upbringing, training, and interests were typically French, and his professional career, as indeed all of his life, was spent in France.

He received his early education at the Lycée Condorcet in Paris, where he showed equally great gifts in the sciences and the humanities. From 1878 to 1881 he studied at the École Normale Supérieure in Paris, the institution responsible for training university teachers. The general culture that he received there made him equally at home in reading the Greek and Latin classics, in obtaining what he wanted and needed from the science of his day, and in acquiring a beginning in the career of philosophy, to which he turned upon graduation.

His teaching career began in various lycées outside of Paris, first at Angers (1881–83) and then for the next five years at Clermont-Ferrand. While at the latter place, he had the intuition that provided both the basis and inspiration for his first philosophical books. As he later wrote to the eminent U.S. Pragmatist William James:

> I had remained up to that time wholly imbued with mechanistic theories, to which I had been led at an early date by the reading of Herbert Spencer. . . . It was the analysis of the notion of time, as that enters into mechanics and physics, which overturned all my ideas. I saw, to my great astonishment, that scientific time does not *endure* . . . that positive science consists essentially in the elimination of duration. This was the point of departure of a series of reflections which brought me, by gradual steps, to reject almost all of what I had hitherto accepted and to change my point of view completely.

Archiv fur Kunst und Geschichte



Bergson, 1928.

First work

The first result of this change was his *Essai sur les données immédiates de la conscience* (1889; Eng. trans., *Time and Free Will: An Essay on the Immediate Data of Consciousness*, 1910), for which he received the doctorate the same year. This work was primarily an attempt to establish the notion of duration, or lived time, as opposed to what he viewed as the spatialized conception of time, measured by a clock, that is employed by science. He proceeded by analyzing the awareness that man has of his inner self to show that psychological facts are qualitatively different from any other, charging psychologists in particular with falsifying the facts by trying to quantify and number them. Fechner's Law, claiming to establish a calculable relation between the intensity of the stimulus and that of the corresponding sensation, was especially criticized. Once the confusions were cleared away that confounded duration with extension, succession with simultaneity, and quality with quantity, he maintained that the objections to human liberty made in the name of scientific determinism could be seen to be baseless.

Philosophical triumphs. The publication of the *Essai* found Bergson returned to Paris, teaching at the Lycée Henri IV. In 1891 he married Louise Neuburger, a cousin of the French novelist Marcel Proust.

Philosophy of mind and body. Meanwhile, he had undertaken the study of the relation between mind and body. The prevailing doctrine was that of the so-called psychophysiological parallelism, which held that for every psychological fact there is a corresponding physiological fact that strictly determines it. Though he was convinced that he had refuted the argument for determinism, his own work, in the doctoral dissertation, had not attempted to explain how mind and body are related. The findings of his research into this problem were published in 1896 under the title *Matière et mémoire: Essai sur la relation du corps à l'esprit* (Eng. trans., *Matter and Memory*, 1911).

Publication of *Matière et mémoire*

This is the most difficult and also, in the opinion of some critics, the most perfect of his books. The approach that he took in it is typical of his method of doing philosophy. He did not proceed by general speculation and was not concerned with elaborating a great speculative system. He began in this, as in each of his books, with a particular problem, which he analyzed by first determining the empirical (observed) facts that are known about it according to the best and most up-to-date scientific opinion. Thus, for *Matière et mémoire* he devoted five years to studying all of the literature available on memory and especially the psychological phenomenon of aphasia, or loss of the ability to use language. According to the theory of psychophysiological parallelism, a lesion in the brain should also affect the very basis of a psychological power. The occurrence of aphasia, Bergson argued, showed that this is not the case. The person so affected understands what others have to say, knows what he himself wants to say, suffers no paralysis of the speech organs, and yet is unable to speak. This fact shows, he argued, that it is not memory that is lost but, rather, the bodily mechanism that is needed to express it. From this observation Bergson concluded that memory, and so mind, or soul, is independent of body and makes use of it to carry out its own purposes.

The *Essai* had been widely reviewed in the professional journals, but *Matière et mémoire* attracted the attention of a wider audience and marked the first step along the way that led to Bergson's becoming one of the most popular and influential lecturers and writers of the day. In 1897 he returned as professor of philosophy to the École Normale Supérieure, which he had first entered as a student at the age of 19. Then, in 1900, he was called to

the Collège de France, the academic institution of highest prestige in all of France, where he enjoyed immense success as a lecturer.

From then until the outbreak of World War I, there was a veritable vogue of Bergsonism. William James was an enthusiastic reader of his works, and the two men became warm friends. Expositions and commentaries on the Bergsonian philosophy were to be found everywhere. It was held by many that a new day in philosophy had dawned that brought with it light to many other activities as well: to literature in the work of Proust in *À la recherche du temps perdu* (1913–27; Eng. trans., *Remembrance of Things Past*, 1922–31) and George Bernard Shaw (especially in *Back to Methuselah*, 1921); to politics in the work of Georges Sorel, a French syndicalist philosopher; to painting in Claude Monet, a French Impressionist; to music in Claude Debussy; and to religion for many thinkers who found support for spiritual values in his work. In fact, his influence caused something of a crisis among French Catholics when the enthusiasm of his followers threatened to substitute Bergsonism for the traditional philosophical preparation for theology. Bergson himself wrote, in a letter to a French Jesuit theologian:

> The considerations exposed in my *Essai* throw light upon the very fact of liberty; those of *Matière et mémoire* permit tangibly to ascertain, as I hope, the reality of the spirit; those of *L'Évolution créatrice* present creation as a fact. From all this clearly follows the notion of a God both creating and free.

Philosophy of biology and metaphysics. *L'Évolution créatrice* (1907; Eng. trans., *Creative Evolution*, 1911), the greatest work of these years and his most famous book, reveals him most clearly as a philosopher of process at the same time that it shows the influence of biology upon his thought. In examining the idea of life, Bergson accepted evolution as a scientifically established fact. He criticized, however, the philosophical interpretations that had been given of it for failing to see the importance of duration and hence missing the very uniqueness of life. He proposed that the whole evolutionary process should be seen as the endurance of an *élan vital* ("vital impulse") that is continually developing and generating new forms. Evolution, in short, is creative, not mechanistic. In this developing process, he traced two main lines: one through instinct, leading to the life of insects; the other through the evolution of intelligence, resulting in man; both of which, however, are seen as the work of one vital impulse that is at work everywhere in the world. The final chapter of the book, entitled "The Cinematographical Mechanism of Thought and the Mechanistic Illusion," presents a review of the whole history of philosophical thought with the aim of showing that it everywhere failed to appreciate the nature and importance of becoming, falsifying thereby the nature of reality by the imposition of static and discrete concepts.

Exposition of *élan vital*

The minor works that Bergson wrote in this period should not be overlooked, because they contribute importantly to an understanding of his thought. In 1900, he published *Le Rire: Essai sur la significance du comique* (Eng. trans., *Laughter: An Essay on the Meaning of the Comic*, 1911) and, in 1903, *Introduction à la metaphysique* (Eng. trans., *An Introduction to Metaphysics*, 1913). The latter provides perhaps the best introduction to his philosophy by offering the clearest account of his method. There are two profoundly different ways of knowing, he claimed. The one, which reaches its furthest development in science, is analytic, spatializing, and conceptualizing, tending to see things as solid and discontinuous. The other is an intuition that is global, immediate, reaching into the heart of a thing by sympathy. The first is useful for getting things done, for acting on the world, but it fails to reach the essential reality of things precisely because it leaves out duration and its perpetual flux, which is inexpressible and to be grasped only by intuition. Bergson's entire work may be considered as an extended exploration of the meaning and implications of his intuition of duration as constituting the innermost reality of everything.

Later years. In 1914 Bergson retired from all active duties at the Collège de France, although he did not formally retire from the chair until 1921. During the war years he undertook several diplomatic missions, including one to the United States. Then, with the formation of the League of Nations, he became the first president of its Commission for Intellectual Cooperation, which he held until poor health compelled his resignation in 1925. Having received the highest honours that France could offer him, including membership, since 1915, among the "forty immortals" of the Académie Française, he was awarded the Nobel Prize for Literature in 1928.

After *L'Évolution créatrice*, 25 years elapsed before he published another major work. During that time he was known to be thinking about the moral and spiritual life of man, and in 1932 he published his results in *Les Deux Sources de la morale et de la religion* (Eng. trans., *The Two Sources of Morality and Religion*, 1935). As in the earlier works, he claimed that the polar opposition of the static and the dynamic provides the basic insight. Thus, in the moral, social, and religious life of men he saw, on the one side, the work of the closed society, expressed in conformity to codified laws and customs, and, on the other side, the open society best represented by the dynamic aspirations of heroes and mystical saints reaching out beyond and even breaking the strictures of the groups in which they live. There are, thus, two moralities, or, rather, two sources: the one having its roots in intelligence, which leads also to science and its static, mechanistic ideal; the other based on intuition, and finding its expression not only in the free creativity of art and philosophy but also in the mystical experience of the saints.

Bergson's notion of the two moralities

Bergson in *Les Deux Sources* had come much closer to the orthodox religious notion of God than he had in the vital impulse of *L'Évolution créatrice*. He acknowledged in his will of 1937, "My reflections have led me closer and closer to Catholicism, in which I see the complete fulfillment of Judaism." Yet, although declaring his "moral adherence to Catholicism," he never went beyond that. In explanation, he wrote: "I would have become a convert, had I not forseen for years a formidable wave of anti-Semitism about to break upon the world. I wanted to remain among those who tomorrow were to be persecuted." To confirm this conviction, only a few weeks before his death, he arose from his sick bed and stood in line in order to register as a Jew, in accord with the law just imposed by the Vichy government and from which he refused the exemption that had been offered him.

Since resigning from the Collège de France in 1921, Bergson had been in very poor health, so subject to pain and headaches that he often could work only a few hours a day. He died in Paris at the age of 81 on January 4, 1941. In those days, especially dark for France, the poet and writer Paul Valéry, in the eulogy that he delivered at the funeral, declared that,

> While misery, anguish, and constraints of every variety depress and discourage enterprises of the mind, Bergson seems already to belong to an age that is past, and his name seems to be the last great name of the European intellect.

Influence. Although it did not give rise to a Bergsonian school of philosophy as such, his influence has been considerable. It reached beyond philosophy, as has already been noted. Among philosophers, his influence has been greatest in France, but it has also been felt in the United States and Great Britain, especially in the work of William James; of George Santayana, a Spanish–U.S. naturalist; and of Alfred North Whitehead, the other great process metaphysician of the 20th century.

BIBLIOGRAPHY. A good introduction to the study of Bergson is provided by HAROLD A. LARRABEE (ed.), *Selections from Bergson* (1949), with an introduction by the editor. The best edition of works is the *Oeuvres*, with notes by ANDRE ROBINET and an introduction by HENRI GOUHIER (1959). Perhaps the best sympathetic account (in French) of the man and his work is that of VLADIMIR JANKELEVITCH, *Henri Bergson* (1959), a complete revision of his earlier work (*Bergson*, 1931), containing a highly congratulatory letter from Bergson himself. Critical, yet also sympathetic, are JACQUES MARITAIN, *La Philosophie bergsonienne*, 2nd ed. (1930); and

JACQUES CHEVALIER, *Bergson* (1926; Eng. trans., 1928). A highly critical account is BERTRAND RUSSELL, *The Philosophy of Bergson* (1914). A good account of Bergson the man may be found in JEAN GUITTON, *La Vocation de Bergson* (1960). On Bergson's influence on literature and the arts, see ROMEO ARBOUR, *Henri Bergson et les lettres françaises* (1955).

(A.Tt./O.A.B.)

Bering Sea and Strait

The Bering Sea (Beringovo More in Russian and in the transliteration system of the Akademiya Nauk) is the northernmost part of the Pacific Ocean, separating the continents of Asia and North America at their closest point. To the north it connects with the Arctic Ocean through the Bering Strait (Proliv Beringa in Russian), at the narrowest point of which the two continents are about 53 miles apart. The boundary between the United States and the Union of Soviet Socialist Republics passes through the sea and the strait.

The Bering Sea roughly resembles a triangle with its apex to the north and its base formed by the Aleutian Islands. To the west lies the coast of Asia and to the east the Alaskan peninsula. Its area is 890,000 square miles (2,304,000 square kilometres), including its islands. The maximum width from east to west is about 1,488 miles and from north to south about 992 miles.

The Bering Strait is a relatively shallow passage averaging 98 to 164 feet in depth. During the Ice Age the sea level fell by several hundred feet, making the strait into a bridge between the continents of Asia and North America, over which a considerable migration of plants and animals occurred.

The strait as a bridge

There are numerous islands in both the sea and strait. The largest are the Aleutians (14,610 sq mi), Nunivak (1,940 sq mi), St. Lawrence (about 1,000 sq mi), Karaginsky (983 sq mi), Nelson (885 sq mi), the Commander or Komandorskiye group (712 sq mi), Hagemeister (125 sq mi), Arakamchechen (100 sq mi), the Pribilof or Fur Seal Islands (87 sq mi), St. Matthew (75 sq mi), Yttygran (25 sq mi), and the two Diomede Islands (about 6 sq mi).

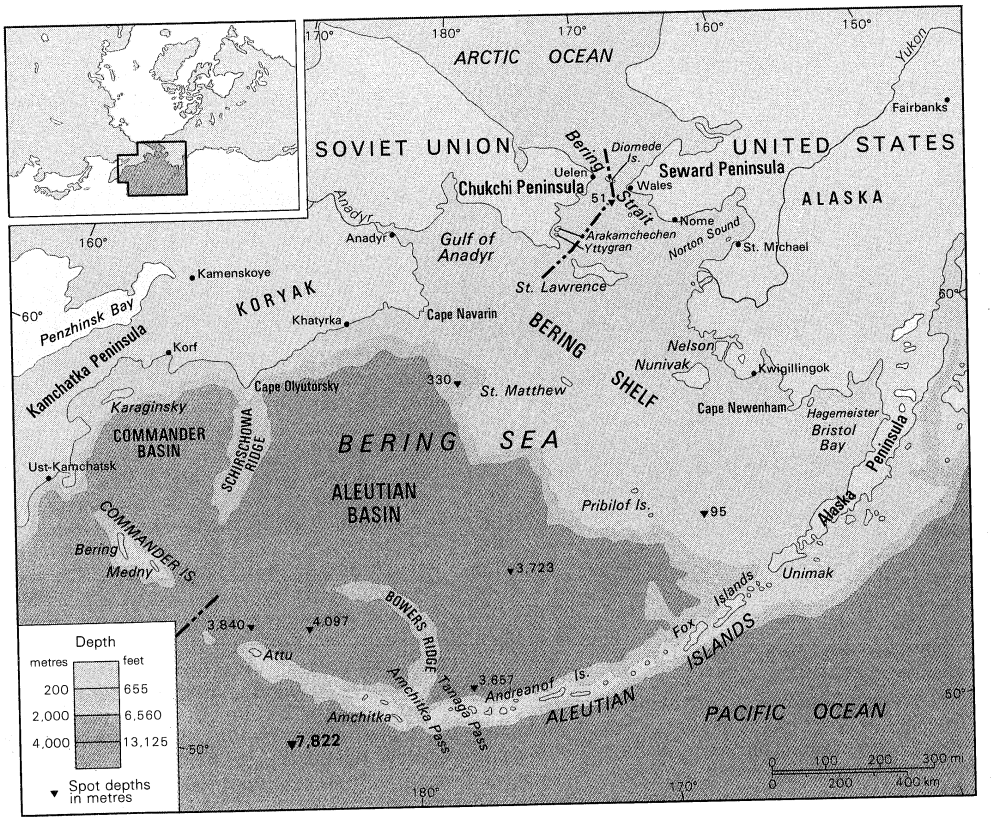
Relief. The Bering Sea may be divided into two nearly equal parts: a shallow area along the continental and insular shelves in the north and east and a deep area in the southwest. In the shelf area, which is an enormous underwater plain, the depths are usually less than 500 feet. The deep part in the southwestern portion is also a plain, lying at depths of 12,000 or 13,000 feet and divided by a ridge into two basins: the Commander Basin and the Aleutian Basin.

The continental crust is more than 12 miles thick along the shallow shelves and in the Aleutian Islands. The thickness decreases in the slope areas, and in the deep part of the sea the crust is six to nine miles thick.

Climate. Although the Bering Sea is situated in the same latitude as Great Britain, its climate is much more severe. The southern and western parts are characterized by cool, rainy summers with frequent fogs, and comparatively warm, snowy winters. Winters are extreme in the northern and eastern portions, with temperatures of −31° to −49° F (−35° to −45° C) and high winds. The summers in the north and east are cool with comparatively low precipitation. Snow persists on the Koryak coast for as long as eight months, and on the Chukchi Peninsula (Chukotsky Poluostrov) for nearly ten months, with a snow cover one to two feet thick. The annual precipitation in the southern part of the sea is more than 40 inches (1,000 millimetres), mainly in the form of rainfall, while in the northern part the precipitation is less than half as much and is mainly snow.

Mean annual air temperatures range from −14° F (−10° C) in the northern areas to about 39° F (4° C) in the southern parts. Water temperatures on the surface average from 34° F (1° C) in the north to 41° F (5° C) in the south. The period without frosts lasts for about 80 days in the northern part of the sea, where snow is common even in the summer and maximum temperatures are only 68° F (20° C). In the southern area there are nearly 150 days without frost, and the temperature seldom falls much below freezing. January and February are the coldest months, July and August the warmest. Typhoons occasionally penetrate the southern part of the sea.

Hydrology. Practically all of the Bering Sea water comes from the Pacific. The salinity of the surface water is 31 to 33 parts per thousand; in the deeper parts of the sea the salinity increases to 34.8 parts per thousand near the bottom. In winter the northern portion of the sea is



The Bering Strait and the Bering Sea.

covered with ice, and even in summer the water below the surface retains a subfreezing temperature. The structure of the Bering Sea waters in general is subarctic, characterized by the presence of a cold intermediate layer with warmer waters above and below. The surface water is heated during the summer, but a considerable layer of water that was cooled during the winter remains cold and is known as the cold intermediate layer. The maximum thickness of this layer is about 475 feet in the northern part of the sea. Underneath this layer is one that is slightly warmer, below which lie the colder bottom waters. In the northern and eastern shallow regions of the sea, only two upper layers develop: surface water and a cooler intermediate layer.

Effect of water layers on plant life

The existence of the cold intermediate layer separating the deep waters, which are rich in nutrient salts, from the upper photic layer (*i.e.*, the layer exposed to sunlight) results in two growths of floating plant life during the year. The first growth occurs in the spring after the mixing of waters in winter; and the second during the autumnal mixing, when the cold surface waters descend and the deeper waters, rich in nutrient salts, come to the surface while there is still sufficient sunlight for plant growth.

Currents and water exchange. Warm oceanic waters from the south enter the Bering Sea in three regions: through the strait between the Medny and Commander islands, through the numerous straits of the Fox Islands, and through the Amchitka and Tanaga passes. The Attu, Tanaga, and Transverse currents carry the warm water to the northwest. The Transverse Current, proceeding along the Asian continental slope in the direction of Cape (Mys) Navarin, branches in two: one branch forms the Lawrence Current moving northward and the other joins the Anadyr Current, which in turn gives birth to a powerful Kamchatka Current that governs the southward movement of the Bering Sea waters along the Asian coasts. Near the Alaska coast the general direction of the water is to the north, a factor responsible for the less severe ice conditions in that part of the sea as compared to the western part. Some of the Bering Sea water passes through the Bering Strait into the Arctic Ocean, but the bulk of it returns to the Pacific. The deep Bering Sea waters rise gradually to the surface and return to the Pacific as surface waters. Thus, the Bering Sea is an important factor in the general circulation of the northern part of the Pacific Ocean waters. The rise to the surface of oceanic waters rich in nutrient salts gives the sea a high biological productivity.

Navigability

The Bering Sea is considered by navigators to be one of the most difficult of seas. Winter storms are frequent and severe, often coating ships' superstructures with ice. Waves may reach over 40 feet in height. Added to these hazards are fog, rain, and floating ice in the northern part of the sea. In winter the northern area is covered by ice fields about four or five feet thick, with hummocks in some places more than 100 feet high. At its maximum extent in April, the ice reaches as far south as Bristol Bay and the Kamchatka coasts. Melting begins in May, and by July there is no ice in the sea except for drift ice in the Bering Strait.

Bottom sediments. From 325,000,000 to 425,000,000 tons of sedimentary material enter the sea annually from the land as a result of erosion of the shore. Plant and animal life at the surface produce 4,742,000,000 tons of sedimentary material, but very little reaches the bottom, and consequently most of the sediment on the floor of the sea is from the land. Along with a great deal of silica, the bottom ooze holds a large quantity of boulders, pebbles, and gravel torn from the shores by ice and carried out to sea. In the southern part, the sediments are rich in material of volcanic origin.

Flora and fauna. The floating plant life of the Bering Sea consists of 163 species, of which the most common are diatom algae. The largest concentration of diatoms have been found in the shallow part of the sea. Diatoms are the principal producers of organic matter, and they are consumed by small copepods (microscopic crustaceans), which in turn become the food of fish and mammals. On the continental shelf there are vast quantities of mollusks, sea urchins (*Echinorachnius parma*), and barnacles. Also abundant on the shelves are sponges, echinoderms (marine animals, such as starfish and sea urchins), marine worms, and crustaceans. In the southern regions, down to depths of 100 or 130 feet, populations of giant brown algae grow like forests on the rocky bottom. There are about 200 species of them, some reaching lengths of 200 or 300 feet.

The Bering Sea has about 315 species of fish, including 50 deep-sea species, of which 25 are caught commercially. The most important among them are salmon, herring, cod, flounder, halibut, and pollack. The islands are breeding grounds for the fur seal and the sea otter. The northern areas are inhabited by the walrus, seal, and sea lion.

There are oil and gas deposits on the continental shelf in the north and deposits of gold and tin on the sea bottom. These have not yet been exploited.

History of exploration. The Bering Strait and the Bering Sea were first explored by Russian ships under Semyon Dezhnev, in 1648. They are named after Vitus Bering, a Danish captain who was taken into Russian service by Peter the Great, in 1704. He sailed into the strait in 1728 but did not see the Alaskan coast, although he discovered the islands of St. Lawrence and Diomede. In 1730 the strait was charted for the first time by Mikhail Gvozdev and Ivan Fyodorov. Bering sailed again in 1733, leading a large expedition from St. Petersburg along the northern coast of Siberia, and he reached the Gulf of Alaska in the summer of 1741. He reconnoitred the southwestern coast of Alaska, the Alaskan peninsula, and the Aleutians, but misfortune befell him, and he perished along with many of his men. In 1780 Russian merchants founded a private company to trade in fur-bearing animals in northwest America. A geographic study of the sea was made at the end of the 18th century and later supplemented by hydrographic studies.

Vitus Bering

Deep-sea studies were begun in 1827 by British explorers. Extensive work was also done by an American group aboard the U.S. Research Vessel "Albatross," in 1893–1906. Since then the sea has been systematically studied by Soviet, U.S., and Japanese investigators.

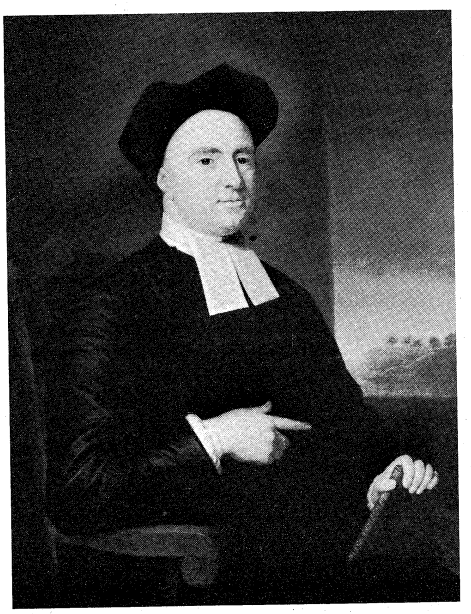
BIBLIOGRAPHY. For more information on the Bering Sea and Strait, see D.M. HOPKINS (ed.), *The Bering Land Bridge* (1967); A.P. LISITSYN, *Recent Sedimentation in the Bering Sea* (1969; orig. pub. in Russian, 1966); A.J. DODIMEAD, F. FAVORITE, and T. HIRANO, "Salmon of the North Pacific Ocean, pt. 2, Review of Oceanography of the Subarctic Pacific Region," *Bull. Int. N. Pacif. Fish. Commn.*, 13:1–192 (1963); the UNITED STATES WEATHER BUREAU and HYDROGRAPHIC OFFICE, *Climatological and Oceanographic Atlas for Mariners*, vol. 2, *North Pacific Ocean* (1961); FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, *Yearbook of Fishery Statistics* (annual); and L.A. ZENKEVICH, *Biology of the Seas of the U.S.S.R.* (1963; orig. pub. in Russian, 1963).

(A.P.L.)

Berkeley, George

George Berkeley, Irish philosopher, economist, mathematician, physicist, and bishop, advanced a new theory of sense perception, discarding the traditional concept of material substance—*i.e.*, of an imperceptible matter underlying the world of sense. His theory was misunderstood in the 18th and 19th centuries, and it has been reassessed today in the light of his revision notebooks, the *Philosophical Commentaries*, formerly known as the *Commonplace Book*.

Berkeley was born at Kilkenny, March 12, 1685, the eldest son of William Berkeley, described as a "gentleman" in George's matriculation entry, and as a commissioned officer, a cornet of dragoons, in the entry of a younger brother. Brought up at Dysart Castle, near Thomastown, Berkeley entered Kilkenny College in 1696, and Trinity College, Dublin, in 1700, where he graduated with a B.A. degree in 1704. While awaiting a fellowship vacancy, he made a critical study of time, vision, and the hypothesis that there is no material substance. The principal influences upon his thinking were Empiricism, represented by the English philosopher,



Berkeley, oil painting by John Smibert, c. 1732. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

John Locke, and Continental Skepticism, represented by Nicolas Malebranche and Pierre Bayle. His first publication, *Arithmetica* and *Miscellanea Mathematica* (published together in 1707), was probably a fellowship thesis.

Period of the revision. Elected fellow of Trinity College in 1707, Berkeley began to "examine and revise" his "first arguings" in his revision notebooks. The revision was drastic and its results revolutionary. His old principle was largely superseded by his new principle; *i.e.*, his original line of argument for immaterialism, based on the subjectivity of colour, taste, and the other sensible qualities, was replaced by a simple, profound analysis of the meaning of "to be" or "to exist." "To be," said of the object, means to be perceived; "to be," said of the subject, means to perceive. Berkeley called attention to the whole situation that exists when a person perceives something, or imagines it. He argued that, when a person imagines trees or books "and no body by to perceive them," he is failing to appreciate the whole situation: he is "omitting" the perceiver, for imagined trees or books are necessarily imagined as perceivable. The situation for him is a two-term relation of perceiver and perceived; there is no third term; there is no "idea of" the object, coming between perceiver and perceived.

"To be is to be perceived"

The revision was a gradual development. At the start Berkeley held that nothing exists but "conscious things." "On second thoughts," he was certain of the existence of bodies and knew intuitively "the existence of other things besides ourselves." His expressions, "in the mind" and "without the mind," must be understood accordingly. As he wrote in his notebook, heat and colour (which philosophers had classed as secondary qualities because of their supposed subjectivity) are "as much without the mind" as figure and motion (classed as primary qualities) or as time; for both primary and secondary qualities are *so* in the mind as to be in the thing, and are *so* in the thing as to be in the mind. The mind does not become red, blue, or extended when those qualities are in it; they are not modes nor attributes of mind. Colour and extension are not mental qualities for Berkeley: colour can be seen, and extension can be touched; they are "sensible ideas," or sense-data, the direct objects of percipient mind.

Berkeley accepted possible perception as well as actual perception; *i.e.*, he accepted the existence of what a person is not actually perceiving but might perceive if he took the appropriate steps. The opposite view was held by some philosophers, including Materialists, who (in Berkeley's words) "are by their own principles forced" to accept it. They are forced to accept that objects actually seen and touched have only an intermittent existence, that they come into existence when perceived and pass into nothingness when no longer perceived. Berkeley treated those views with respect; he denied that they are absurd; but he did not hold them, and he explicitly denied that they follow from his principles. In effect he said to his readers, "You may hold, if you will, that objects of sense have only an 'in-and-out' existence, that they are created and annihilated with every turn of man's attention; but do not father those views on me. I do not hold them." In his notebook he wrote, "Existence is *percipi* or *percipere*. the horse is in the stable, the Books are in the study as before." Horse and books, when not being actually perceived by man, are still there, still perceivable "still with relation to perception." To a nonphilosophical friend Berkeley wrote, "I question not the existence of anything that we perceive by our senses."

Berkeley's immaterialism is open to "gross misinterpretation," as he said in his preface; rightly understood, it is common sense. Like most people, he accepted and built on "two heads," "two kinds entirely distinct and heterogeneous": (1) active mind or spirit, perceiving, thinking, and willing; and (2) passive objects of mind, viz., sensible ideas (sense-data) or imaginable ideas.

Period of his major works. Berkeley's golden period of authorship followed the revision. In *An Essay Towards a New Theory of Vision* (1709), he examined visual distance, magnitude, position, and problems of sight and touch, and concluded that "the proper (or real) objects of sight" are not without the mind, though "the contrary be supposed true of tangible objects." In his *Treatise Concerning the Principles of Human Knowledge*, Part I (1710), he brought all objects of sense, including tangibles, within the mind; he rejected material substance, material causes, and abstract general ideas; he affirmed spiritual substance; and he answered many objections to his theory and drew the consequences, theological and epistemological. His *Three Dialogues between Hylas and Philonous* (1713), by its attractive literary form and its avoidance of technicalities, reinforced the main argument of the *Principles;* the two books speak with one voice about immaterialism.

Empirical epistemology and immaterialism

Berkeley was made a deacon in 1709 and ordained a priest in 1710. He held his fellowship for 17 years, acting as librarian (1709), junior dean (1710–11), and tutor and lecturer in divinity, Greek, and Hebrew. In politics he was a Hanoverian Tory, and he defended the ethics of that position in three sermons, published as *Passive Obedience* (1712). Thus, with four major books in five years, the foundations of his fame were laid; and, when he first left Ireland in 1713 on a leave of absence, he was already a man of mark in the learned world; his books were reviewed on the Continent, and Gottfried Wilhelm Leibniz, the wide-ranging author of the *Monadology*, knew of his immaterialism and commented upon it.

Among the London wits he was an immediate success. Jonathan Swift, dean of St. Patrick's Cathedral, Dublin, presented him at court. For Sir Richard Steele, an essayist, he wrote essays in *The Guardian* against the freethinkers. He was in the theatre with Joseph Addison, essayist and poet, on the first night of *Cato*, and has left a spirited description of the experience. Alexander Pope credited him with "ev'ry virtue under heav'n." In 1713–14 he went on an embassy to Sicily as chaplain with Charles Mordaunt, 3rd earl of Peterborough, whom Berkeley called an "ambassador extraordinary." In 1715 during the Jacobite rebellion (on behalf of the exiled Stuarts) he proved his loyalty by publishing his *Advice to the Tories Who Have Taken the Oaths*. He was abroad again from 1716 to 1720 in Italy, acting as tutor to George Ashe, son of the Bishop of Clogher (later, of Derry); his four travel diaries give vivid pictures of sightseeing in Rome and of tours in southern Italy. On his return he published his *De motu* (1721), which rejected Sir Isaac Newton's absolute space, time, and motion, gave a veiled hint of his immaterialism, and has recently earned him the title "precursor of Mach and Einstein."

Resuming his work in Dublin, he took a full part in teaching and administration for more than three years.

In 1724 he was appointed dean of Derry, and his 24 years' connection with Trinity College ended.

His American venture and ensuing years. The deanery and legacy from Hester van Homrigh (Swift's Vanessa) were seen by Berkeley as providences, furthering his "scheme of Bermuda," in the New World. The frenzied speculation that preceded the bursting of the South Sea Bubble had shaken his faith in the Old World, and he looked in hope to the New. His *Essay Towards preventing the Ruin of Great-Britain* (1721) was soon succeeded by his prophetic verses on "Westward the course of empire takes its way." Already by 1722 he had resolved to build a college in Bermuda for the education of young Americans (Indians), publishing the plan in *A Proposal For the better Supplying of Churches . . .* (1724). The scheme caught the public imagination; the King granted a charter; the Archbishop of Canterbury acted as trustee; subscriptions poured in; and Parliament passed a contingent grant of £20,000. But there was opposition; an alternative charity for Georgia was mooted; and the prime minister, Sir Robert Walpole, hesitated.

In 1728 Berkeley married Anne, daughter of Chief Justice Forster, a talented and well-educated lady, who defended her husband's philosophy after his death. Soon after the wedding, they sailed for America, settling at Newport, Rhode Island, where Berkeley bought land, built a house (Whitehall), and waited. Berkeley preached often in Newport and its neighbourhood, and a philosophical study group met at Whitehall. Eventually, word came that the grant would not be paid, and Berkeley returned to London in October 1731. Several American universities, Yale in particular, benefitted by Berkeley's visit; and his correspondence with Samuel Johnson, later president of King's College (Columbia University), is of philosophical importance.

Works on religion and mathematics

Alciphron: or, The Minute Philosopher (1732) was written at Newport, and the setting of the dialogues reflects local scenes and scenery. It is a massive defense of theism and Christianity with attacks on deists and freethinkers and discussions of visual language and analogical knowledge and of the functions of words in religious argument.

Upon his return to London in 1731, Berkeley's pen, never idle for long, became active. A writer in the *Daily Post-boy* commended *Alciphron* but attacked the appended *Essay* on vision. Berkeley replied with *The Theory of Vision, or Visual Language . . . Vindicated and Explained* (1733). This fine work brought the metaphysics (theory of Being) of the *Essay* into line with the *Principles* and added his doctrine of cause, admitting defects in the premises of the original *Essay*. *Alciphron* provoked replies from the satirist Bernard de Mandeville; John Hervey, Baron Hervey of Ickworth; the statesman Henry St. John, 1st Viscount Bolingbroke; and Peter Browne, his former teacher and provost. To Browne, Berkeley sent a long, private letter on analogy—discovered only recently and first published in *Mind* (July 1969)—which comprises an important supplement to his 4th dialogue.

In 1734 Berkeley published *The Analyst; or, a Discourse Addressed to an Infidel Mathematician*, which Florian Cajori, a historian of mathematics, has called "the most spectacular event of the century in the history of British mathematics." Besides being a contribution to mathematics, it was an argument *ad hominem* for religion. "He who can digest a second or third fluxion . . .," wrote Berkeley, "need not, methinks, be squeamish about any point in divinity." A long and fruitful controversy followed. James Jurin, a Cambridge physician and scientist, John Walton of Dublin, and Colin Maclaurin, a Scottish mathematician, took part. Berkeley answered Jurin in his lively satire *A Defence of Free-Thinking in Mathematics* (1735) and answered Walton in an appendix to that work and again in his *Reasons For not Replying . . .* (1735).

Years as bishop of Cloyne. Berkeley was consecrated as bishop of Cloyne in Dublin in 1734. He found Trinity College flourishing: its new library was completed and John Stearne's Doric printing house was being built. To the latter, Berkeley contributed a fount of Greek type and also founded the Berkeley gold medal for Greek. His episcopate, as such, was uneventful. He took a seat in the Irish House of Lords in 1737 and, while in Dublin, published *A Discourse addressed to Magistrates and Men in Authority* (1738), condemning the Blasters whose Hell-Fire Club, now in ruins, still can be seen near Dublin.

The see-house at Cloyne was a cultured home and a social centre and, during epidemics, a dispensary. On arrival the family consisted of his wife and two sons; and two more sons and two daughters were born at Cloyne. George, the only child to prolong the line, became canon of Canterbury.

In 1745 Berkeley addressed open letters to his clergy and to the Roman Catholics of his diocese about the Stuart uprising. In letters to the press over his own name or through a friend, he expressed himself on several public questions, political, social, and scientific. Two major works stand out, *The Querist* and *Siris*, both occasioned by the poverty, distress, and sickness around him, but rising far above their source.

Works on economics and the chain of being

The Querist, published in three parts from 1735 to 1737, deals with basic economics—credit, demand, industry, and "the true idea of money"—and with special problems, such as banking, currency, luxury, and the wool trade. The final query puts the central question, "Whose fault is it if poor Ireland still continues poor?"

Siris (1744) passed through some six editions in six months. It is at once a treatise on the medicinal virtues of tar-water, its making and dosage, and a philosopher's vision of a chain of being, "a gradual evolution or ascent" from the world of sense to "the mind, her acts and faculties" and, thence, to the supernatural and God, the three in one.

In August 1752, Berkeley commissioned his brother, Dr. Robert Berkeley, as vicar general and arranged with the bishop of Cork as to his episcopal duties and, with his wife and his children George and Julia, went to Oxford and took a house in Holywell Street. George entered at Christ Church.

Berkeley died suddenly on January 14, 1753, and was buried in Christ Church Chapel.

BIBLIOGRAPHY

Biography: A.A. LUCE, *The Life of George Berkeley, Bishop of Cloyne* (1949, reprinted 1968), a definitive work; B. RAND, *Berkeley's American Sojourn* (1932); A. BRAYTON, *George Berkeley in Apulia* (1946) and *George Berkeley in Newport* (1954).

Works: A.A. LUCE and T.E. JESSOP (eds.), *The Works of George Berkeley, Bishop of Cloyne*, 9 vol. (1948–57), definitive edition with valuable introductions and notes; *Philosophical Commentaries, Generally Called the Commonplace Book*, ed. by A.A. LUCE (1944); *A Treatise Concerning the Principles of Human Knowledge*, ed. by C.M. TURBAYNE (1957); *The Principles of Human Knowledge, and Three Dialogues Between Hylas and Philonous*, ed. by G.J. WARNOCK (1962); *Bishop Berkeley's Querist in Historical Perspective*, text with critical comment, and discussion of the basis of credit and of 18th-century monetary, banking, and currency problems, by J. JOHNSTON (1970).

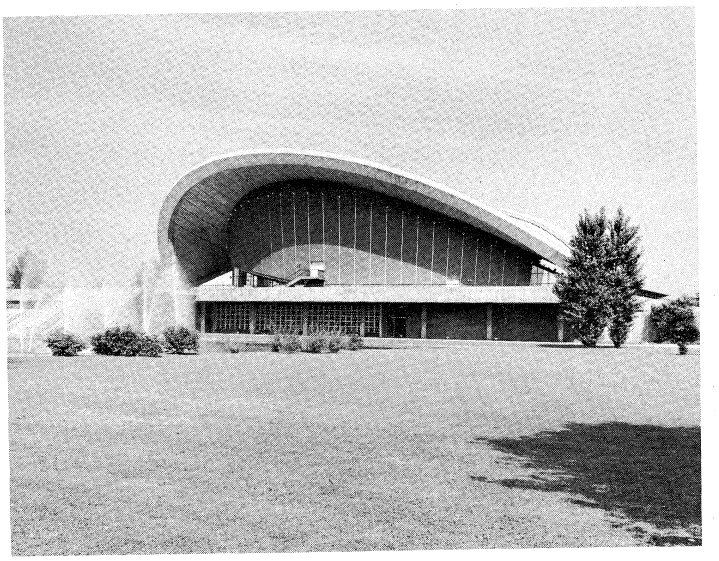
Works on Berkeley: G.A. JOHNSTON, *The Development of Berkeley's Philosophy* (1923, reprinted 1965); J.M. HONE and M.M. ROSSI, *Bishop Berkeley*, introduction by W.B. YEATS (1931); G.D. HICKS, *George Berkeley* (1932); A.A. LUCE, *Berkeley and Malebranche* (1934, reprinted 1967), *Berkeley's Immaterialism* (1945, reprinted 1968), and *The Dialectic of Immaterialism* (1963); J. WILD, *George Berkeley* (1936, reprinted 1962); G.J. WARNOCK, *Berkeley* (1953), by a Logical Analyst; M. GUEROULT, *Berkeley* (1956); A.L. LEROY, *George Berkeley*, (1959); H.M. BRACKEN, *The Early Reception of Berkeley's Immaterialism, 1710–1733*, rev. ed. (1965); D.M. ARMSTRONG, *Berkeley's Theory of Vision* (1960), and (ed.), *Berkeley's Philosophical Writings* (1965); G.W.R. ARDLEY, *Berkeley's Renovation of Philosophy* (1968); W.E. STEINKRAUS (ed.), *New Studies in Berkeley's Philosophy*, 13 contributions, with selected bibliography by T.E. JESSOP (1966); A.D. RITCHIE, *George Berkeley: A Reappraisal*, ed. by G.E. DAVIE (1967); G.W. ENGLE and G. TAYLOR (eds.), *Berkeley's Principles of Human Knowledge: Critical Studies* (1968). See V.I. LENIN, *Materialism and Empirio-Criticism* (English edition 1950), for an *ex parte* study of Berkeley in connection with revisionism.

Bibliography: For earlier works on Berkeley, see *A Bibliography of George Berkeley* by T.E. JESSOP (1934); the list is continued through 1962 by C.M. TURBAYNE and R. WARE in the *Journal of Philosophy*, pp. 93–112 (1963).

Berlin

A city of the 20th century divided by a wall more appropriate to the Middle Ages, Greater Berlin lies sprawled in the heart of Europe. It is situated athwart a natural east–west commercial and geographical axis, dominating the North European Plain, that helped make it the capital of the kingdom of Prussia and then of the German *Reich*. Berlin today stands wholly within the territory of the German Democratic Republic but echoes Germany's division since World War II into two separate states by being itself conformably divided into independent halves—East Berlin being the capital of the Socialist German Democratic Republic and West Berlin being the 11th *Land*, or state (though not constitutionally incorporated as such), of the German Federal Republic in the west, from which it is physically isolated.

With 1945, Berlin's former glory ended. But a durable city mastered two of three tasks posed for it in the aftermath of World War II: Berlin survived; Berlin was rebuilt, with amazing economic growth in the western half; the third task, mastery of its future in the shadow of a wall, remained. That barrier, since its erection in 1961 by Communist authorities in East Berlin, quickly dominated and determined the contemporary city's character. With the wall and a steel fence, the East German Democratic Republic surrounding West Berlin strengthened the position of East Berlin as its capital, ended a massive drain of manpower to the west, and completed the isolation of West Berlin from its natural hinterland. (For related information see GERMAN DEMOCRATIC REPUBLIC; GERMANY, FEDERAL REPUBLIC OF; GERMANY, HISTORY OF.)



Congress Hall, an international meeting hall in the Tiergarten, West Berlin; completed in 1957.
Wayne Andrews

HISTORY

Origins. Berlin appeared for the first time in recorded history in 1244, seven years after its sister city, Kölln, with which it eventually merged. Both were founded about the beginning of the 13th century for geographical and mercantile reasons, as they commanded a natural east–west trade route over the Spree River.

Founding

The way for their founding was opened by a Germanic resurgence in the area, which had been abandoned to the Slavs by the original Germanic tribes as they migrated westward. The Slavs were subdued by Albert I the Bear, a Saxon, who crossed the Elbe River from the west. His successors took the title margrave of the mark (border territory) of Brandenburg. To this day Greater Berlin retains as its symbol a black bear standing on its hind legs in a defiant posture.

Man had first appeared in the area in 50,000 BC. The ice age that covered the sandy north German plain left behind natural waterways that would further Berlin's commercial importance throughout time. Large forests once covered much of the Berlin area, and remnants have survived as the Grunewald, Jungfernheide, and the Spandauer Stadtforst wooded areas of West Berlin. Great stones called *Findlinge*, which came with the ice that moved south from Scandinavia, still dot Berlin fields and forests.

The settlements of Spandau and Köpenick, which survive in the form of metropolitan districts, preceded the establishment of Berlin-Kölln. The Ascanians, followers of Albert I the Bear, established their fortress in 1160 at Spandau in the north, where the Spree flows into the Havel River; by 1232 the fortress had earned the privileges of a town. Berlin-Kölln emerged between Spandau to the north and Köpenick to the south. Kölln stood on an island in the Spree, while Berlin developed on the river's east bank. This opening to the east soon allowed it to pass its confined sister city in importance. By 1250 Berlin-Kölln dominated the German east to the Oder River where a fort had been built in 1214. A common civil government, courts, and justice were established in 1307. With the demise of the Ascanians in the 14th century, Berlin-Kölln joined the Hanseatic League of North German mercantile towns. Herring, grain, wood, and wool were early products.

The Hohenzollerns. In 1411 the mark of Brandenburg came under the governorship of the Nürnberg feudal baron Frederick VI (1371–1440), beginning Berlin's association with the Hohenzollerns, who from the end of the 15th century as prince electors of Brandenburg established Berlin-Kölln as their capital and permanent residence.

The Thirty Years' War of 1618–48 laid a heavy financial burden on the city. When Frederick William the Great Elector (1620–88) assumed power in 1640, he found the population reduced by war levies from 12,000 to 7,500. He embarked on a building program that included fortification of his residence, enabling him to expell Swedish invaders. His time also marked the beginning of the development of canals, which by 1669 provided a direct link between Breslau in the east and Hamburg and the open sea in the west.

His son, who became Frederick I (1657–1713), the first king in Prussia, in 1701, laid the foundation for a gradually emerging new German national state. Already in 1698, Andreas Schlüter (1664–1714), the first of Berlin's great builders, began fashioning the palace residence into a Baroque masterpiece. Schlüter's fame as a sculptor was ensured by his creation of the equestrian statue of the Great Elector. In 1950 the Communist authorities in East Berlin razed the war-damaged royal palace to make way for their Marx-Engels-Platz. Schlüter's statue, however, survived and was placed in West Berlin before the Charlottenburg Palace, designed by the Baroque architect Johann Arnold Nering (1659–95). This is Greater Berlin's most elegant structure remaining from royal times, painstakingly restored after wartime bombing.

The founding of the Academy of the Arts in 1696 and the Academy of Sciences in 1700 gave impetus to the city's academic and cultural life. In 1709 the framework of Greater Berlin was laid when Berlin-Kölln and the newer towns of Friedrichswerder, Dorotheenstadt, and Friedrichstadt were put under a single magistrate. Population grew from 12,000 in 1670 to 61,000 in 1712, including a garrison of 8,000 men and 6,000 French Huguenot refugees.

The first half of the 18th century saw Berlin expand in all directions. Frederick II the Great (1712–86) adorned the city with a number of new buildings, among which the opera house (1741–43) by the Rococo architect Georg Wenzeslaus von Knobelsdorff (1699–1753) is notable. Situated in Unter den Linden (Under the Linden Trees), one of the finest and most spacious avenues in Europe, it was destroyed in World War II but later restored. At the western end of Unter den Linden stood the Brandenburg Gate (Brandenburger Tor; 1789–93) by the Neoclassical architect Carl Gotthard Langhans (1732–1808), a ceremonial Doric gateway that was to become the symbol of Greater Berlin.

The role of Frederick the Great

In 1809 the scholar Wilhelm von Humboldt (1767–1835) founded Berlin's Friedrich-Wilhelm-Universität, renamed the Humboldt-Universität zu Berlin after World War II, and by 1840 it was the largest in Germany, with 1,772 students. It early attracted such outstanding thinkers as the philosopher Georg Wilhelm Friedrich Hegel (1770–1831) and the prophet of communism, Karl Marx (1818–83). Berlin had its first popular uprising in 1830 when tailor apprentices took to the streets over working conditions. A quarrel over the price of potatoes set the stage for the Revolution of 1848, which produced 200 dead in a clash between soldiers and citizenry. The city's population rose to 415,000, from about 100,000 a century before. The railroad age dawned with the opening of the Berlin–Potsdam line in 1838, and Berlin became the centre of a rail net that incorporated 12 lines, enabling easy movement of troops.

The period of railway growth was also that of the Iron Chancellor, Otto von Bismarck (1815–98), whose successful military ventures paved the way for the creation of a united Germany. The First Reich came into being when the king of Prussia was crowned Kaiser (Emperor) William I in 1871, at which time the population of Berlin, his capital, was 826,000. Berlin was to retain its role as capital of the German *Reich* until 1945. The Second Reich began with the reign of William II and lasted until his abdication at the end of World War I.

The republic and Hitler. On November 9, 1918, Berlin became the capital of the first German republic, proclaimed from the Reichstag building by Philip Scheidemann (1865–1939), the Social Democratic leader. The period 1918–33, in Berlin as elsewhere in Germany, was one of great disorder, political murder, runaway inflation, failure of the republic, and the rise to power of Adolf Hitler. Political and economic chaos spurred the latter's ascendency. Meanwhile Berlin played an important part—particularly in theatre, cinema, and other entertainment—in the art boom of the feverish 1920s, producing a brilliance that contrasted with the darker forces that were to gain the day: by 1932 the unemployed in Berlin alone totalled 636,000, and on January 31, 1933, Hitler became chancellor, his storm troopers marching through the Brandenburg Gate with massed flags and torches. He took absolute power in February of that year by using a mysterious fire that swept the Reichstag as an excuse for the enactment of emergency laws.

In 1936 Berlin was the scene of the most spectacular of modern Olympic Games in a specially built 100,000-seat stadium complex (whose Olympic village became the postwar British occupation headquarters). A landmark of another kind, "Crystal Night" (November 9–10, 1938), so called because of the glass shattered in shop windows of Jewish-owned stores, initiated a pogrom that in Berlin reduced a Jewish population of 170,000 to 5,000 by 1945, a figure that had risen only to 6,000 by 1970.

Impact of World War II

Allied aerial bombing during World War II cost Berlin an estimated 52,000 dead. Another 100,000 civilians died in the battle for Berlin launched by the Soviet Red Army on April 16, 1945. Berlin's residential districts, factories, military facilities, streets, and cultural buildings were pounded into the sandy plain (one-sixth of all the wartime rubble in Germany lay in Berlin). In the aftermath of bombardment, the Trümmerfrauen, the rubble women—the war left Berlin with a population 70 percent female—cleared the refuse. On April 30, 1945, Hitler committed suicide in his bunker below the Chancellery. On May 8 the Russians staged an elaborate German surrender ceremony in Berlin, rivalling that held by the Western Allies the day before in Reims, France. This divided ritual began the drama of the contemporary city, focal point of an East–West political struggle.

Berlin divided. Greater Berlin had begun to assume its contemporary face in 1912 when seven districts and two agrarian counties formed a "union of expediency." The city now had a population of over 2,000,000 for the first time. On October 1, 1920, the creation of a metropolitan Berlin took place. It fused seven districts, 59 country communities, and 27 landed estates into a single association. Twenty resultant districts were integral parts of metropolitan Berlin, but still largely autonomous. At the end of World War II in 1945 the Soviet Union took eight districts as its sector of occupation, Berlin Mitte, Prenzlauer Berg, Friedrichshain, Treptow, Köpenick, Lichtenberg, Weissensee, and Pankow. Berlin Mitte included the sites of original Berlin-Kölln, with most historic establishments, Unter den Linden, and the Brandenburg Gate. What was called the New West End, developed after old Berlin had outgrown its space, became West Berlin. The U.S. sector was formed by the southern districts of Zehlendorf, Steglitz, Tempelhof, Neukölln, Kreuzberg, and Schöneberg, the last named the seat of the West Berlin city government after western representatives had been locked out of the old city hall and its assembly in 1948. The British sector embraced the central and western districts of the Tiergarten, including the park laid out by Frederick the Great, and the broad Kurfürstendamm, which became the symbol of the new West Berlin, as well as Wilmersdorf, Charlottenburg, and Spandau. The French were allotted the northern Wedding and Reinickendorf districts.

These zones were based on an agreement reached in London in 1944 among the United States, Britain, and the Soviet Union, acting on a British plan that divided Germany into occupation zones and Greater Berlin into sectors within, but not part of, the Soviet zone of occupation. A significant feature of agreements concerning Berlin was the inability of the Western side to get a written Soviet guarantee of access.

In March 1948 the Western powers decided to unite their zones of Germany into a single economic unit. In protest, the Soviet representative withdrew from the Allied Control Council. In June 1948, the West introduced a currency reform in what became West Germany, including West Berlin. The Soviet Union responded by launching a land blockade of the western sectors of the city on June 24, 1948.

Blockade and airlift

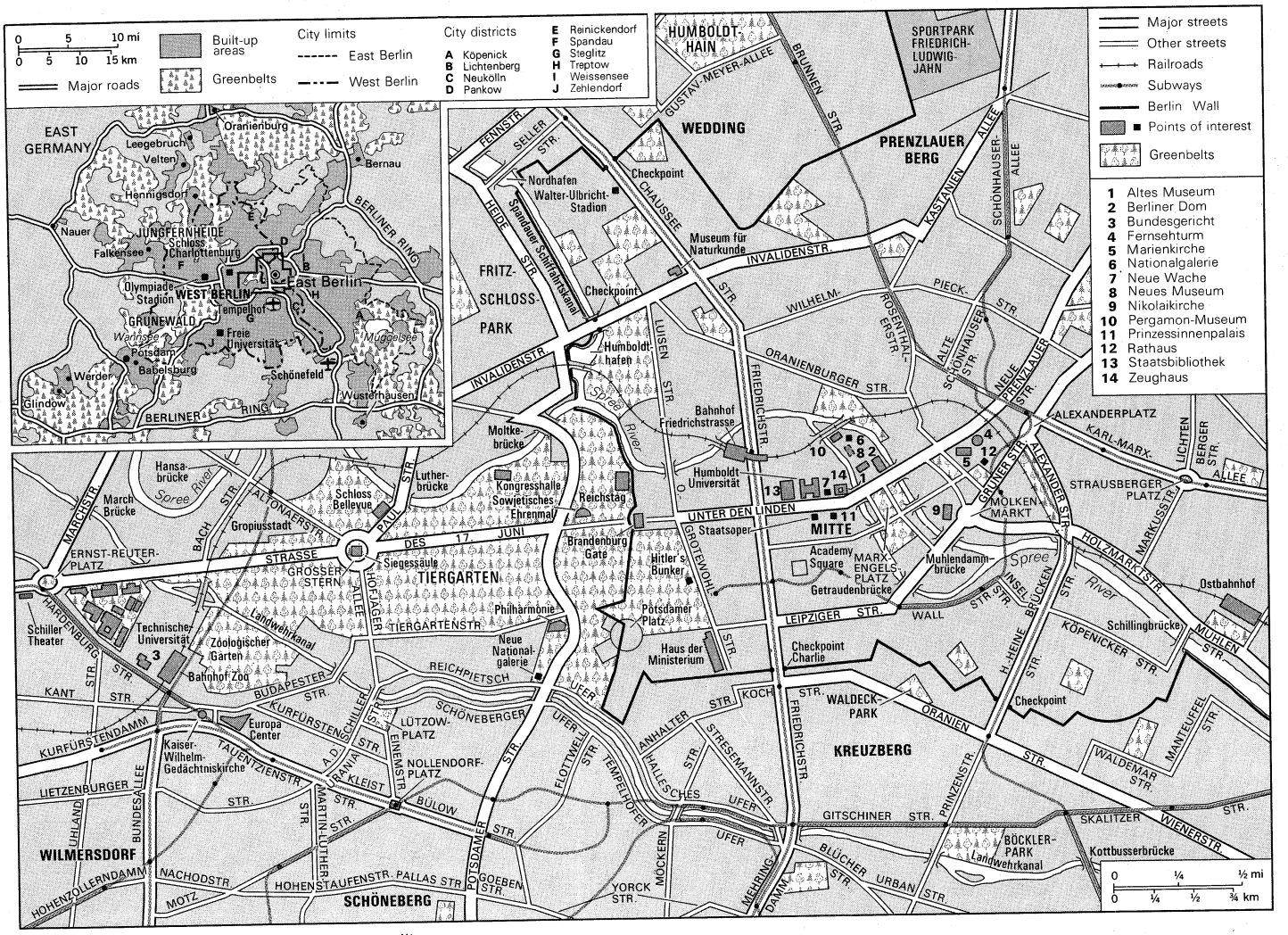
A great airlift broke this attempt to cut off the city from vital supplies, Western Allied planes hauling 1,831,200 tons of food, coal, and other necessities. The Russians abandoned the blockade on May 12, 1949, but the Western Allies kept flying until September, building up a year's supply of essential goods which were stockpiled in Berlin from then on. Seventy allied airmen and eight German workers died keeping the airlift going in and out of Tempelhof, Gatow, and Tegel airfields and off the water of the Havel River. From time to time afterward access by road to West Berlin was denied for short periods as part of a pattern of harassment. In the meantime, on November 30, 1948, a separate municipal government with its own chief burgomaster, or mayor, was set up in East Berlin, and this completed the splitting of Berlin between East and West.

On June 17, 1953, some 50,000 workers rebelled in East Berlin. The uprising, which spread throughout the German Democratic Republic, was crushed by Russian intervention. The Russians also backed the East German regime when it decided to build a wall in Berlin to stop a massive outflow of refugees to the West. On August 13, 1961, West Berlin was physically separated from East Berlin by the erection of a nine-foot wall topped with barbed wire, and its isolation from the East German countryside elsewhere secured by an encircling electrical contact fence surmounted by watchtowers with searchlights overlooking a strip of cleared ground. In an attempt to negate this manifest East-West confrontation, the United States, Britain, and France joined the Soviet Union in signing a Berlin agreement in September 1970 that was designed to ease tensions. The working out of practical details for easier access to West Berlin from the West and for visits beyond the wall for West Berliners was left to officials of the two German states. The Soviet Union obtained the right to open a Consulate General in West Berlin.

The wall

THE CONTEMPORARY CITY

The city and its people. Berlin lies 110 miles from the borders of West Germany proper, about 112 miles south of the Baltic Sea, and 118 miles north of Czechoslovakia.



Central Berlin and (inset) its metropolitan area.

In 1945, with German lands east of the Oder and Neisse rivers placed under Russian and Polish control as the conquering Red Army advanced to Berlin, the defeated *Reich's* capital found itself no longer in the centre of Germany. It again was on an eastern fringe, much as when it was founded in the 13th century, 52 miles from Frankfurt-on-the-Oder and present-day Poland.

Sprawling in area, Berlin, from the 18th century on, had broad streets, such as the north-south Friedrichstrasse, and a central east–west axis, running from the site of the former royal palace (Marx-Engels Platz) along the Unter den Linden to the Brandenburg Gate (just east of the walled frontier with West Berlin) and then across the Tiergarten through Charlottenburg to the Havel River in the west. The wall dividing West and East Berlin runs in an irregular line roughly from north northwest to south southeast.

Waterways

With the Spree River running through its centre, touching East and West, and with a chain of lakes formed by the Havel River on its western outskirts, Berlin continued as a city on the water, having also the further attraction of large birch and Scotch pine forests.

Berlin's waterways were formed by the glaciers of the Pleistocene Epoch, the last of three ice carpets slowly draining off 20,000 years ago. The city was established and prospered where the meteorological influence of the Atlantic Ocean fades and where the great expanse of continental plain extending to Asia begins to hold sway. This climatic crossroads gave Berlin what the inhabitants call *Berliner Luft*, a type of air colder but not as damp as that of western Germany and fresher because of quick replenishment from across lake-studded flatlands, despite increasing industrial smog. Berlin's average mean temperature is 49° F (8.8° C). The seasonal breakdown is: winter 30° F; spring 48° F; summer 63° F; and autumn 49° F. Average precipitation is 23 inches (580 millimetres) with a summer high. About one-fourth to one-fifth of the total falls as snow, which covers the ground, on average, for about 50 days a year.

Berlin has a mean elevation of 115 feet (35 metres) above sea level. In the business quarter the highest natural point is the Kreuzberg, a hill 216 feet (66 metres) high, in the southern district of Schöneberg, where in 1821 the state architect of Prussia, Karl Friedrich Schinkel (1781–1841), created a monument to the German nation commemorating the wars of liberation against Napoleon. The elevation of the Schäfer Berg near the biggest of the Havel lakes in West Berlin, the Wannsee, is 338 feet (103 metres). It is crowned by a 695-foot (212-metre) radio tower. In natural height it is lower than the 377-foot- (115-metre-) high hills rising around East Berlin's Müggelsee, the largest lake in Greater Berlin. This height was matched by the largest of many man-made hills formed by an estimated 104,000,000 cubic yards of rubble from World War II that dot the city, East and West, 16 alone in West Berlin. One of the latter, the 377-foot "Devil's Mountain," was turned into a winter sports area for skiing and sledding.

Transportation. In inner city transportation, the bus is the mainstay, although East Berlin maintains streetcar service as well. New rapid-transit systems have brought some postwar unity to the sphere of transport: East Berlin runs the S-Bahn (Stadt-Bahn, "city railway") elevated railway system started in 1871 as a connecting system to a rail net in and out of the city. By 1963, the S-Bahn had a citywide net of 202 miles with 77 stations in the West and 51 in the East. The East also controls railroad service under the old name, Reichsbahn. Through trains are

Rail systems

few, with only one major station in West Berlin. Subway construction began in 1897, and by World War II the city possessed one of the finest systems in Europe, consisting of 92 stations forming a net stretching 46.6 miles. Administration was divided between East and West in 1948. Western additions increased the total to 55.9 miles, with East Berlin adding 14.9 miles.

Air traffic has played an important role from 1945, particularly in the West, a role increased in importance at the time of the airlift. Tempelhof provides a midcity field for West Berlin with Tegel, the site of early rocket launches, furnishing an extensive auxiliary field for large jets. East Berlin's Schönefeld also accommodates the largest aircraft.

Only planes of the United States, Great Britain, and France are able to use the air corridors to the West. The East German line, Interflug, uses Schönefeld mostly for eastwardly traffic. By the early 1970s, however, each half of the city was seeking a broader air traffic program that would take advantage of Berlin's natural east–west and north–south axis for traffic across Europe in every direction.

The Reichsautobahn (National Expressway) in Berlin is the Berlin terminus of a superhighway net developed before World War II that covers Germany. West Berlin has built a crosstown extension and northern and southern interchanges. The whole and East Berlin's main traffic arteries are linked with the Berliner Ring, a circle of Autobahn connecting road around the city, putting Berlin in the centre of access spokes.

Demography. Two Berlin half-cities virtually equal in area emerged from the 1945 division into four occupation sectors. Of the 20 metropolitan districts the Soviet Union took the eight eastern ones, while the Americans took six, the British four, and the French two of the 12 in the West. From the time Greater Berlin unified itself under metropolitan administration in 1920, an overall expanse of 339 square miles (878 square kilometres) increased slightly so that now West Berlin comprises 185 square miles (480 square kilometres) and East Berlin 156 square miles (403 square kilometres), a total of 341 square miles (883 square kilometres).

The population of West Berlin in 1970 was over 2,000,000 and of East Berlin a little over 1,000,000. Some 800,000 people living in the West had close relatives in the East. A modest quarter-century increase from 2,800,000 in the whole city at the end of World War II still represented a considerable drop from more than 4,300,000 in 1939. The erection of modern buildings and the restoration of historic ones has made Greater Berlin, with broad, bustling streets in each half, retain its position as the largest German city anywhere. It also retains its role as Germany's greatest population centre, once sixth and now 16th among the 20 largest cities in the world, and reclaims its functions as the centre of Germanic culture, technology, medicine, and industrial enterprise, despite its political fragmentation.

Berlin's role in German life

The unitary character of Greater Berlin is illustrated in the age structure of its population. As the city moved into the final third of the 20th century, one-fifth of its inhabitants were over 65 years old. Although the forecast of West Berlin deaths remained fixed at about 40,000 annually, birth rate projections indicated a drop of from 22,419 in 1970 to 17,578 by 1978. If so, West Berlin's population would fall below 2,000,000 by 1980. In essence, the city never recovered from a low rate of 8.6 births per thousand in 1944 and the high of 55.5 deaths per thousand in the collapse of 1945. At the beginning of the war the birth rate was 17.2 compared with a death rate of 13.2. For 1969, East Berlin registered 3.1 less births than deaths per thousand, while West Berlin had 9.8 less.

West Berlin managed to even out its death losses with an inflow of about 30,000 West German workers per year. Likewise East Germany had a backlog of workers wanting to move to East Berlin. Each side began employing foreign workers, with the figure in West Berlin rising above 100,000 by 1970. In East and West, more than two children per family was highly unusual, since young couples in their family planning paid close attention to schooling, housing, and job opportunities. The high proportion of elderly persons contributed to such social problems, aggravated by loneliness, as suicide and drunkenness.

Housing. Emphasis on new construction in both parts of Berlin has been in housing and office buildings. Each side has built clusters of new high-rise apartment districts, with the Gropiusstadt of the West the most ambitious. Named after its designer, Berlin-born architect Walter Gropius (1883–1969), it houses 50,000 people in 17,000 apartments, the tallest building rising 31 floors. A subway line and Autobahn have been so planned to permit extension beyond the nearby wall boundary to East Berlin's Schönefeld airfield lying in view beyond. City planners in East and West continue to consider eventual reunion so that new streets will fit and new buildings will mesh within a single concept.

Architectural features. An effort to blend the new with the traditional is evident. In West Berlin, the 1957 Congress Hall (Kongress Halle; called "the pregnant oyster" by Berliners, because of its shape) and the restored Reichstag building, rebuilt at a cost of 100,000,000 Deutsche Marks, are examples of this trend; also significant are the 1963 Philharmonic Concert Hall (Philharmonie) and a new National Gallery of modern art (Nationalgalerie), the last creation of the architect Ludwig Mies van der Rohe (1886–1969), who first worked in Berlin before World War I. This western complex also includes the Victory Column (Siegessäule) from the wars of 1864–70, Schloss Bellevue (castle), new hotels, and a 20-story glass and steel Europa Centre near the new buildings of the Kaiser Wilhelm-Gedächtniskirche (Memorial Church), whose original blackened main tower has been left as a war memorial.

Buildings in the East

East Berlin also has its own new symbol and church war memorial. Berlin's oldest building, the Nikolai-Kirche, dating from around 1200, was gutted by bombing, and its red brick walls were left standing as a reminder. The central East Berlin area, however, is dominated by the Communist regime's first great postwar prestige project, a 1,170-foot (357-metre) television tower. It commands the Berlin landscape and has a revolving restaurant at the 800-foot level. Because it had to be erected on Berlin's sandy soil, the tower represents an engineering feat and is seen by East Germans as a symbol of their capital.

The tower, completed in 1969, stands adjacent to a

Berlin Funk—Ullstein



The new memorial church built next to the ruined Kaiser Wilhelm Memorial Church, West Berlin.



Alexander Platz, meeting point of Karl-Marx-Allee and Lenin-Allee, two main avenues in East Berlin.
Siegfried Sammer—Bavaria Verlag

refashioned Alexander Platz. This square, once a crossroads of Greater Berlin, leads to the 1952 Stalin-Allee, a postwar housing project where the revolt of 1953 began; it was renamed Karl-Marx-Allee after Stalin's death. The 40-floor City of Berlin Hotel on Alexander Platz is Berlin's tallest business or residential building.

Unter den Linden also combines old and new, featuring modern hotels and shops along with the restored Zeughaus (Armoury), Neue Wache (New Watch), Crown Prince Palace, Prinzessinnenpalais (Princess Palace), State Opera, Old Library, Kaiser Wilhelm Palace and University, renamed after its founder, Humboldt. The Brandenburg Gate regained its sculptured chariot with four horses abreast in 1959, re-created after a model in West Berlin. At the head of the boulevard lies the great blackened mass of the Berlin Cathedral (Berliner Dom), which has not been rebuilt. Nearby stands Berlin's oldest remaining church, the Marienkirche, and also the so-called museum island with the Old (Altes) and New (Neues) museums, National Gallery (National-Galerie) and Pergamon Museum, containing, among other treasures from Pergamon, the altar of Zeus. In the area where the royal palace once stood are grouped the traditional red brick city hall, a foreign ministry, State Council Building, and the rebuilt St. Hedwig's Cathedral, Berlin's first Roman Catholic church to be put up after the Reformation. South of Unter den Linden is the old Gendarme Market, renamed Academy Square, once one of the finest architectural centres in Berlin, where restoration has begun of the twin German and French cathedrals and the State Theatre. The Wilhelmstrasse, seat of Prussian and *Reich* governments, has mostly gone. On one side lies the wall, Hitler's bunker, now covered by a grassy mound, and the empty Potsdamer Platz; on the other, the Nazi propaganda ministry of Joseph Goebbels still stands, taken over for Communist use. The marble from Hitler's nearby *Reich* Chancellery was used by the Russians for their war memorial just inside West Berlin.

Economy, administration, and social services. *Economic life.* To a large extent, traditional economic activities have been revived throughout Greater Berlin. These include textiles, ironworks and steelworks, rail cars, sewing machines, chemicals, china, breweries, and machine works. Electronics production is a principal industry in both East and West, and each is a fashion centre, as before the war. West Berlin has developed cigarette production as a vital industry. Bicycles from East Berlin have formed a substantial element of war aid to North Vietnam. Berlin continues to be a central and chief market for wheat, rye, and cereals from nearby and from farther east. This trade is channelled largely through East Berlin, since West Berlin is isolated from its hinterland.

East Berlin authorities try to maintain morale through cheaper basic elements in the cost of living, particularly in accommodation rents and foodstuffs such as bread and potatoes. But shortages keep prices high for such items as automobiles, refrigerators, washing machines, and colour television sets. Salaries are lower than in the West, but comparative buying power is difficult to assess.

Administration. In the East and West separately, district government after World War II has continued much as before: with chief burgomaster, or mayor, city assembly, or parliament, and district mayors and councils, although a trend toward centralization runs strongly in such matters as citywide integration of education. On the higher level, subordination is the rule in East and West. In the East this results from its status as capital of East Germany. The chief burgomaster runs East Berlin through a city parliament but is overshadowed by the presence of the apparatus of the central Communist regime, including the Volkskammer (People's Chamber). East Berlin delegates to this body have something of a reduced position because the system of four-power responsibility is still kept alive, notably concerning free access to East Berlin for members of the western military garrisons. West Berlin being accounted the 11th *Land* of the German Federal Republic, its city parliament elects 22 delegates to the West German Bundestag (federal parliament) in Bonn but without full voting powers since the *Land* Berlin is not yet constitutionally part of the German Federal Republic, retaining occupied status. The U.S., British, and French commandants retain ultimate authority in their respective sectors. Soviet four-power participation continues at an air-safety centre for western flights to and from the city. The western commandants also have final say over police matters within their sectors. A 15,000-man West Berlin force includes large elements of paramilitary police who form an adjunct to the western garrisons and total about 12,000 men: this arrangement reflects the highly ambiguous condition within what was to have become a demilitarized Berlin under military occupation after the war.

Subordinate role of city parliaments

Direct remilitarization by Germans is apparent only in East Berlin, however. The West German Bundeswehr (Federal Defense Force) is barred from the city, and its residents can volunteer for military service but are not liable to draft. East Germany openly drafts East Berliners, keeps garrisons for units of the People's Army within the city, and maintains a ring of 14,000 men in three brigades along the wall. Recruits are sworn in at public ceremonies, a weekly changing of the guard is staged with Prussian pomp and ceremony on Unter den Linden, and each May Day the People's Army parades in force. Allied protests are ignored. The Russians also maintain a Berlin garrison, with some 50,000 more men ringed about Greater Berlin.

Politics. East Berlin political parties are identical with those of East Germany, as is true of West Berlin and West Germany. The Socialist Unity (Communist) Party takes a commanding position in a National Front, comprising remnants of Social Democrats, Christian Democrats, Liberals, and nationalist groupings. The Social Democrats have kept control of the West Berlin city government from the days of Lord Mayor Ernst Reuter, with the Christian Democrats the second largest party, followed by the Free Democrats. There also is a small West Berlin Communist Party.

Justice. In matters of justice, East Berlin is fully integrated within an overall East German court system. In West Berlin, the presence of the West German Constitutional Court is barred because of Allied reservations. In practical law, however, West German justice and legislation apply under the federal constitution as they do under the East German constitution in East Berlin.

Health and education. Far-reaching health insurance is available throughout the city, which once again forms Germany's largest centre of medical activity. In East Berlin, the Charité has resumed its pivotal role, first taken up as a royal hospital on its foundation in 1710. In West Berlin, a modern Klinikum, or teaching hospital, has introduced new methods to medical practice.

Berlin has traditionally played a leading role in German education, which in the postwar West has been pursued



Altes Museum, East Berlin, an example of German Neoclassical architecture, by Karl Friedrich Schinkel, 1822–30.
By courtesy of the Staatliche Museen zu Berlin

with unconstrained and reformist zeal. Communist ideology forms the basis of education in East Berlin, where reform mostly involves better organized and more efficient teaching methods. East Berlin has had an earlier start through its physical control of the old Berlin University. Because of Communist hegemony, the non-Communists left in 1948 and founded the Free University in West Berlin. From its inception, the Free University has drawn political activists from all over Germany. At first, many students participated in daring operations to bring refugees out of the East. By 1965, a new left had emerged whose militancy was carried onto the streets, leading to clashes with the police. This activism also appeared at the Technical University and Teachers' Training College. In 1970, West Berlin students totalled more than 30,000. Limited dissidence surfaced in East Berlin, particularly after the 1968 invasion of Czechoslovakia, with the sons and daughters of prominent Communists taking part in public demonstrations.

Students and politics

Culture and recreation. By 1750, the Prussian State Opera on Unter den Linden already was the finest opera house in Europe. Berlin between World Wars I and II blossomed in all areas of recreation and cultural life. After 1945, each half of the city successfully exerted great effort to re-establish this tradition. Along with shops and restaurants, one of the first business activities to be resumed was night life, the city having long been famous for its clubs, cabarets, and other amusement enterprises. A feature of the early postwar period was the political cabaret, in which the Berliners' characteristic sharpness of tongue and quick, cockney-like humour found expression.

In East Berlin, both the State Opera and the Comic Opera offer first-rate productions, with the latter having a reputation for colourful presentation as well as musical excellence. The rebuilt German Opera in West Berlin attracts the major dramatic singers of the West. In 1928, Berlin had premiered *The Three-penny Opera* by the dramatist and producer Berthold Brecht (1898–1956) and composer Kurt Weill (1900–50). Brecht founded the postwar Berliner Ensemble in East Berlin, and after his death its work, which had won international acclaim, continued along with that of the German Theatre. The Schiller Theatre and old Hebbel Theatre have formed the nucleus of a theatrical revival in West Berlin. Festivals of East and West, in music and other arts, including films and (in West Berlin) jazz, help to fill a year-round program.

But it is in classical music that Greater Berlin again shows consistent high quality. Foremost among a number of accomplished musical aggregations of East and West is the West Berlin Philharmonic Orchestra, founded in 1882 and led after 1954 by a great conductor, Herbert von Karajan. East Berlin continues as a centre of film production through its Defa studios. West Berlin attempted to bring back the days when Greater Berlin rivalled Hollywood, but it was a largely unsuccessful effort because of the city's isolation from the West. The film industry's weakness also reflects a loss of management leadership that has parallels in general business as well. In broadcasting, East and West are roughly equal. Each half of the city is served by five radio programs and two television stations, with colour viewing in both East and West. There also are U.S., British, and French military radio stations and a U.S. television station with limited range. A Soviet Army station broadcasts from outside the city. The symbol of broadcasting in West Berlin is the 400-foot Funkturm built in 1926.

The role of the media

Greater Berlin has managed to retain a greater number of newspapers than is usual even for large cities, with nine dailies in West Berlin and five in the East. This is still far from the 147 newspapers it had in 1928. West Berlin also is the headquarters for Europe's largest publishing house, that of Axel Springer, located directly at the Communist wall. Greater Berlin once rivalled Leipzig as a German publishing centre, but postwar development failed to regain such a ranking, partly because the former publishers' row, Kochstrasse, almost totally disappeared under wartime bombing.

East and West also have a wide-ranging sports program. Spacious parks, woods, and numerous streams and lakes help to make Berlin an outdoor city in its leisure pursuits. The city's traditionally respected zoo (founded 1844), southwest of the Tiergarten, is a major attraction in West Berlin, while East Berlin founded its own zoo.

BIBLIOGRAPHY. O.F. GANDERT *et al.*, *Heimatchronik Berlin* (1962), is a scholarly, definitive, and clearly written treatment of Berlin and its history down to the present day; H. KOTSCHENREUTHER, *Kleine Geschichte Berlins* (1967), an easy to follow chronological treatment of Berlin events. W. KRUMHOLZ, *Berlin ABC*, 2nd ed. (1969), is a city government sponsored work specializing in events since 1945. W. HEGEMANN, *Das Steinerne Berlin* (1930), concentrates on the buildings in Berlin with critical comment concerning its rows of stony rental barracks; P.O. RAVE, *Berlin* (1962; Eng. trans., 1966), is mainly a pictorial work showing the various historical buildings still standing in Berlin. J.W. KELLER, *Germany: The Wall and Berlin* (1964), is useful in charting the background of postwar events; A. BERGER, *Berlin 1945–1969* (1970), gives quick reference to events applying to the presently divided city; C. RYAN, *The Last Battle* (1966), gives perspective to the fall of the city and to rival East–West views already forming before the war's end. A. SPEER, *Inside the Third Reich* (1970), is a revealing look at the Berlin of the Third Reich by a man uncommonly close to Hitler. W. VOGEL, *Führer Durch die Geschichte Berlins* (1966), is another historical treatment prepared as a guide, incorporating some new material.

(H.J.Er.)

Berlioz, Hector

The French composer, critic, and conductor Hector Berlioz was a many-sided genius whose contributions to 19th-century music only began to be fully understood some 50 years after his death. They have now, 100 years later, earned him a position among the greatest composers in the history of Western music.

Study and early career. Louis-Hector Berlioz was born on December 11, 1803, in the village of La Côte-Saint-André 35 miles northwest of Grenoble in the French Alps. France was at war; the schools were disrupted; and Berlioz received his education from his father, an enlightened and cultured physician, who gave him his first lessons in music as well as in Latin. But, like many composers, Berlioz received in his early years little formal training in music. He worked out for himself the elements of harmony and by his 12th year was composing for local chamber-music groups. With help from performers, he learned to play the flute and the guitar, becoming a virtuoso on the latter instrument.

In 1821 his father sent him to Paris to study medicine, and for a year he followed his courses faithfully enough to obtain his first degree in science. But he took every opportunity to go to the Paris Opéra, where he studied, score in hand, the whole repertory, in which the works of Gluck had for him the most appeal and authority. His musical vocation had become so clear in his mind that he contrived to be accepted as a pupil of Jean-François Lesueur, professor of composition at the Paris Conservatoire. This led to disagreements between Berlioz and his parents that embittered nearly eight years of his life.

He persevered, took the obligatory courses at the Conservatoire, notably counterpoint under the influential Czech composer and teacher Anton Reicha, and in 1830 won the Prix de Rome, having received second prize in an earlier competition. These successes pacified his family but were, in a sense, incidental to his career, for in the same year he had finished and obtained a performance of his first great score, which is also a seminal work in 19th-century music, the *Symphonie fantastique*.

Performance of first great score

It was in some respects unfortunate that, instead of being able to follow up this success, Berlioz was required, under the terms of his prize, to spend three years abroad, two of them in Italy. During his long Paris apprenticeship, he had experienced the "revelation" of two modern musicians, Beethoven and Weber, and of two great poets, Shakespeare and Goethe. He had meanwhile fallen in love, at a distance, with Harriet Smithson, a Shakespearean actress who had taken Paris by storm; and, on the rebound from this rather one-sided attachment, he had become engaged to a brilliant and beautiful pianist, Camille Moke (later Mme Pleyel). In leaving Paris, Berlioz was not only leaving a flirtatious fiancée and the artistic environment that had stimulated his powers; he was also leaving the opportunity to demonstrate what his genius saw that modern French music should be. The public was content with the "Paris school," dating back to the 1780s and 1790s, and there is evidence that all Europe (including the Vienna of Beethoven and Schubert) accepted the productions of André Grétry, Étienne Méhul, Luigi Cherubini, and their followers as leading the musical world.

The Bettmann Archive



Berlioz.

Berlioz wanted to bring forward the work of Weber and Beethoven (including the last quartets) and add contributions of his own. He also preached, for the sake of dramatic expression in music, a return to the master of the stage, Gluck, whose works he knew by heart. These three musicians were all in some sense dramatists, and to Berlioz music must first and foremost be dramatically expressive. This doctrine he had begun to expound in his first musical reviews, as early as 1823, and, with the sharpness and strength of an early vision, it remained the artistic creed of his mature years. When one understands its intellectual and intuitive basis, one understands also the reasons for his dynamic career. What may look like self-seeking—the unceasing effort to have his music played—was, in fact, the dedication of his tremendous energies to a cause, often at the expense of his own creative work. The result of his many journeys to Germany, Belgium, England, Russia, and Austria-Hungary was that he taught the leading orchestras of Europe a new style and, through them, taught a new idiom to the young composers and critics who flocked wherever he went.

Italian interlude

But, before these "campaigns" began, Berlioz had his time of reflection in Italy. He has told in his *Mémoires* (published 1870) how bored and unproductive he was after the rich output of the Paris years, which had brought forth an oratorio, numerous cantatas, two dozen songs, a mass, part of an opera, two overtures, a fantasia on Shakespeare's *Tempest*, and eight scenes from Goethe's *Faust*, as well as the *Symphonie fantastique*. Even in Italy, however, Berlioz filled notebooks, met the Russian composer Mikhail Glinka, made a lifelong friend of Mendelssohn, and tramped the hills with his guitar over his shoulder, playing for the peasants and *banditti* whose meals he shared. The impressions gathered in Italy remained a source of both musical and dramatic inspiration down to the last of his works, *Les Troyens* and *Béatrice et Bénédict* (first performed 1862).

Meanwhile, his love affair not prospering and his impatience with life at the Villa Medici in Rome becoming acute, he returned to France after 18 months, thus forfeiting part of his prize.

Mature career. Back in Paris, he set about conquering it anew. He put together a collection of earlier pieces in a form then fashionable, the monodrama, or recitation by one actor interspersed with musical scenes. Since the *Symphonie fantastique* had ended with the death and demonic torments of the protagonist, Berlioz called his new work *Le Retour à la vie* (later *Lélio*, after the hero's name). First performed in 1832, this concoction, which contains three or four delightful pieces, enjoyed great success, and Berlioz had reason to think himself launched again.

A series of accidents brought him in touch with the actress Harriet Smithson, still beautiful but in financial straits, and his mental and artistic attachment to her reviving, he married her on October 3, 1833. The marriage did not last, though for some years the couple led a peaceful existence at Montmartre in the house that Maurice Utrillo later never tired of painting. Among the visitors there were the young poets and musicians of the Romantic movement, including Alfred de Vigny and Chopin. It was there that Berlioz's only child, Louis, was born and also where he composed his great *Requiem*, the *Grande Messe des morts* (1837), the symphonies *Harold en Italie* (1834, inspired by Byron's *Childe Harold*) and *Roméo et Juliette* (1839), and the opera *Benvenuto Cellini* (Paris, 1838).

It was after the premiere of *Harold en Italie* that Berlioz had the astonishing experience of seeing the world-famous violin virtuoso Paganini fall at his feet and declare that he was a genius destined to carry on the new musical tradition initiated by Beethoven. The next day Berlioz received 20,000 francs with a letter from Paganini repeating this judgment. Using the money to free himself from journalistic drudgery, Berlioz composed the choral symphony *Roméo et Juliette*, dedicated to Paganini.

But in Paris it was always expected that a composer, regardless of his bent, should be tested at the Opéra. Berlioz's friends intrigued to procure the assignment of a libretto. An adaptation of Benvenuto Cellini's autobiography was secured, and Berlioz finished his score in a short time. But the intrigue now passed to the other side, which saw to it that the production of *Benvenuto Cellini* at the Opéra failed. From this blow the work itself and the composer's reputation in France never recovered during his lifetime. The score, still little known in the early 1970s, is a masterpiece, and the attribution of the failure to the libretto shows ignorance of the qualities of both the libretto and the music.

The *Requiem* of 1837 had been a government commission for a military and ceremonial occasion designed to encourage the Rome laureate. The request to compose another work for a public ceremony—the *Symphonie funèbre et triomphale* (*Funeral Symphony*) for military band, chorus, and strings, commissioned for the tenth anniversary of the July Revolution (1840)—was intended as a partial solace for the defeat of *Benvenuto Cellini*. A few years before, Berlioz's literary gifts had won him

the post of music critic for the leading Paris newspaper, the *Journal des Débats*, and his employers wielded political influence. Once again, there were intrigues, but the score of the *Funeral Symphony* was ready for the inauguration of the Bastille column. Unfortunately, the music was drowned out by the assembled drum corps, a disaster that Berlioz repaired by giving the work the following month at a concert hall. This was the score that Wagner, then seeking fame and fortune in Paris, admired so wholeheartedly.

Conflicting views of Berlioz and Wagner

Berlioz was able to put Wagner in the way of some musical journalism and thus began a fitful connection of 30 years between the two men whose influence on modern music still resembles a battle of ideals: Berlioz aiming at the creation of drama in and through music alone; Wagner at a marriage of symphony with opera. Although Berlioz and Wagner met again in London in 1855 and found each other congenial, their philosophical differences generally kept them apart, as did also Wagner's habitual duplicity.

After 1840 Berlioz's life consisted of a series of tours across Europe. The last of these was an exhausting series of concerts in St. Petersburg and Moscow in 1867, when he was desperately ill. But it had the effect of introducing the Russian Five, notably Mussorgsky, to his style through his manuscript scores and his conducting. For Berlioz was the first of the virtuoso conductors, having made himself such in order to supply the deficiencies of men who were unable to direct the new music according to the new canon: play what is written. Moreover, the rhythmical difficulties of his scores and the unfamiliar curve of his melodies disconcerted many. The orchestras themselves had to be taught a new precision, vigour, and ensemble, and this was Berlioz's handiwork. Wagner's memoirs bear testimony to this "revelation of a new world," which he experienced at Berlioz's hands in 1839.

On orchestration itself (and, even more important, on instrumentation) Berlioz produced the leading treatise, *Traité d'instrumentation et d'orchestration modernes* (1844). Much more than a technical handbook, it served later generations as an introduction to the aesthetics of expressiveness in music. As Albert Schweitzer has shown, its principle is as applicable to Bach as to Berlioz, and it is in no way governed by considerations of so-called program music. To this last-named genre of dubious repute, Berlioz did not contribute more than the printed "story" of his first symphony, which was and is intelligible as music, without any program.

Dramatic works

Among Berlioz's dramatic works, two became internationally known: *La Damnation de Faust* (1846) and *L'Enfance du Christ* (1854). Two others began to emerge from neglect after World War I: the massive two-part drama *Les Troyens* (composed 1855–58; part II produced in Paris, 1863; entire, in Cologne, 1898), based on Virgil's story of Dido and Aeneas, and the short, witty comedy *Béatrice et Bénédict*, written between 1860 and 1862 and based on Shakespeare's *Much Ado About Nothing*. For all these Berlioz wrote his own librettos. He also wrote a *Te Deum* (1849; performed 1855), which is a fitting counterpart to the *Requiem*, and between 1843 and 1856 he orchestrated his songs, including the song cycle *Les Nuits d'été* (*Summer Nights*). Among his best-known overtures are *Le Roi Lear* (1831), *Le Carnaval romain* (1844), based on material from *Benvenuto Cellini*, and *Le Corsaire* (1831–52).

Berlioz's last years were marked by fame abroad and vulgar hostility at home. In his private life he was incapacitated by illness and saddened by many deaths. His first wife, from whom he was separated but to whom he still felt a deep attachment, died in 1854; his second wife, Maria Recio, who had been his companion for many years and whom he had married when he became a widower, died suddenly in 1862. Finally, his son, who was a sea captain and on whom he concentrated the affection of his declining years, died of yellow fever in Havana at the age of 33, two years before Berlioz's own death in Paris on March 8, 1869.

Assessment. The outstanding characteristics of Berlioz's music—its dramatic expressiveness and variety—account for the feeling of attraction or repulsion that it produces in the listener. Its variety also means that devotees of one work may dislike others, as one finds lovers of Shakespeare who detest *Othello*. But Berlioz also presents a particular difficulty of musicianship in being closer to the true sources of music than to its German, Italian, or French conventions; his melody is abundant and extended and is often disconcerting to the lover of four-bar phrases; his harmony may be obvious or subtle, but it is always functional and frequently depends on elements of timbre; his modulations can be harsh and may even seem harsher than they would in another composer, because he uses his effects sparingly and achieves much by small means and adroit contrasts. This is also true of his orchestration, generally light and transparent, never pasty. As Shaw said: "Call no conductor sensitive in the highest degree to musical impressions until you have heard him in Berlioz and Mozart."

Stylistic traits

The Belgian composer César Franck once said that Berlioz's whole output is made up of masterpieces. He meant by this that each of the composer's dozen great works was the realization of a conception distinct from all the others, rather than successive efforts to attain perfection in the last or best of a series. Franck's judgment is borne out by the fact that, unlike many composers, Berlioz almost never repeats himself. Rather, he created a fresh style for each of his subjects, with the result that familiarity with one is no guaranty of ready access to another. Nothing could be less alike than the *Symphonie funèbre et triomphale* and *Roméo et Juliette* or than the *Requiem* and *L'Enfance du Christ*. To be sure, Berlioz's harmonic system seems the same throughout, partly because it deviates so noticeably from common expectation and partly because its nuances are only now being appreciated for what they are, instead of being looked upon as clumsy attempts to do something else. Again, his melody and free counterpoint everywhere carry his mark—the sinewy originality and dynamic equilibrium of the former, the ingeniously careless independence of the latter. Yet, out of these characteristic elements Berlioz makes a radically different atmosphere for each of his dramas and within them for each of his dramatis personae. Only a repeated hearing of any given work discloses all the power and art (including what would now be called psychology) that it contains. This does not mean that these works are without flaw; it does mean that they embody unique conceptions, to be taken for what they have to give and which no other composer provides.

In the creation of drama and atmosphere, Berlioz excels in scenes of melancholy, introspection, love—gentle or passionate—the contemplation of nature, and the tumult of crowds. His intention throughout is to combine truth with musical sensations, be they powerful or (to quote Shaw again) "wonderful in their tenuity and delicacy, unearthly, unexpected, unaccountable."

Much might be added or quoted that would show the extent to which Berlioz's music still needs careful and dispassionate study. In 1935 the respected British musicologist Sir Donald Tovey, who had not before heard *Les Troyens*, declared that it is "one of the most gigantic and convincing masterpieces of music drama." And, he went on, "You never know where you are with Berlioz." What is certain is that books that date from the 19th century or echo its views, with or without a bias toward Wagner or Debussy, will mislead the student and possibly close the ears of the listener. It is easy to represent Berlioz as merely a craftsman in tone colour who helped develop the resources of the orchestra. But with the repeated performance of the major works all over the Western world, the more comprehensive judgment has come to prevail that Berlioz is a dramatic musician of the first rank. Before 1945 the Berlioz repertoire was limited to the *Symphonie fantastique* and a few brief extracts. The great works, done once and usually with insufficient preparation, produced little effect and confirmed the wisdom of letting them lie. The advent of long-playing records radically altered the situation. Audiences can now judge the interpretations that they are being given, and thus they hear Berlioz performances with a knowledge and

critical attention comparable to those with which they hear the other classical composers.

MAJOR WORKS

OPERAS: *Benvenuto Cellini* (first performed 1838); *Les Troyens*, comprising *La Prise de Troie* and *Les Troyens à Carthage* (composed 1855–58); and *Béatrice et Bénédict* (1862).

CHORAL WORKS: *Huits Scènes de Faust* (1829); *Lélio ou Le Retour à la vie* (1831, sequel to *Symphonie fantastique*); *Requiem—Grand Messe des Morts* (1837); *Roméo et Juliette* (1839); *Symphonie funèbre et triomphale* (1840); *La Damnation de Faust* (1846); *L'Enfance du Christ* (1854), oratorio; *Te Deum*, p. 22 (1855).

ORCHESTRAL WORKS: *Waverly* (1823), overture; *Les Francs-Juges* (composed *c.* 1827), overture; *Symphonie fantastique* (1830–31); *Le Roi Lear* (1831), overture; *Le Corsaire* (1831–52), overture; *Harold en Italie* (1834), symphony with solo viola; *Le Carnaval romain* (1844), overture.

VOICE AND ORCHESTRA: *La Mort de Cléopâtre* (1829); *La Captive* (1834); *Les Nuits d'été* (1843 and 1856); *La Mort d'Ophélie* (1850).

SONGS WITH PIANO: *Irlande* (1829–30), five songs; *Les Nuits d'été* (original version, 1834–41); *Trente-trois Melodies* (1863).

BIBLIOGRAPHY. An intensive study of the musical aesthetic elements in Berlioz's work as a whole is being facilitated by the publication of a scholarly edition of both his scores and his literary works. The former, gradually replacing the earlier and untrustworthy "German" edition, are being prepared in England and published in West Germany; the latter are being edited in France. Their joint effect, besides disposing of 19th-century errors and prejudices, should be to enlarge materially the pleasure of untold readers and listeners.

HECTOR BERLIOZ, *Memoirs*, trans. by DAVID CAIRNS (1969), gives the story, the atmosphere, and the purpose of the artist's life; his *Evenings with the Orchestra*, trans. by JACQUES BARZUN (1956), is a sampling of Berlioz's brilliant music criticism. *The New Letters of Berlioz, 1830–1868*, ed. by JACQUES BARZUN (1954), provides a selection illustrative of the composer's day-to-day activities. For a judgment typical of the Wagnerian era, see ROMAIN ROLLAND, "Berlioz," *Musiciens d'aujourd'hui*, 2nd ed. (1908), trans. in BARRETT H. CLARK (comp.), *Great Short Biographies of the World* (1929). J.G. PROD'HOMME, *Hector Berlioz (1803–1869) sa vie et ses oeuvres*, 3rd ed. (1927), marks the beginning of the rehabilitation. WALTER J. TURNER, *Berlioz, the Man and His Work* (1934); and TOM S. WOTTON, *Hector Berlioz* (1935, reprinted 1969), affirm, with much musical detail, the now accepted estimate of the man and his works. JACQUES BARZUN, *Berlioz and the Romantic Century*, 3rd ed., 2 vol. (1969), gives the fullest account of Berlioz's life and times.

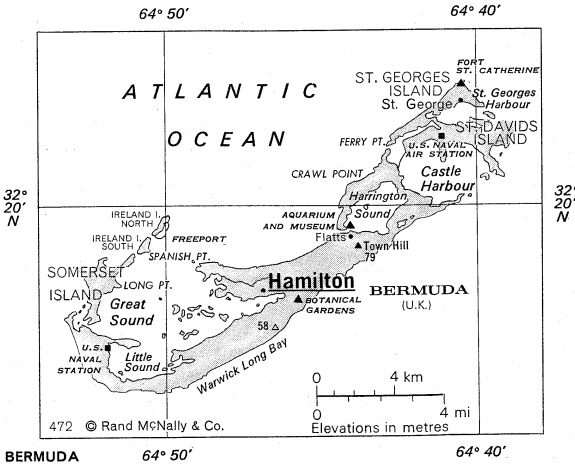
(J.Ba.)

Bermuda

Bermuda is an internally self-governing British colony in the western North Atlantic, consisting of a group of islands about 570 miles off Cape Hatteras on the east coast of the United States. The fishhook-shaped chain of islands stretches for about 22 miles but has an area of only about 21 square miles (54 square kilometres), of which about two square miles are leased to the United States government for air and naval bases. Of the 145 islands and groups of rock, 120 are named but only about 20 are inhabited. The largest is Great Bermuda, known as the Main Island, about 14 miles in length.

The colony's name was derived from that of the Spanish navigator Juan Bermúdez, who is generally credited with discovering the islands, perhaps about 1503. Bermuda's mild climate year round, the surrounding waters, and the historic charm of Hamilton, the capital, and of Saint George, the former capital, are among the factors that make tourism the chief industry. In 1970 more than half of the 53,000 inhabitants were of African descent. Bermuda's Parliament is the oldest among the British Commonwealth countries, representative government having been introduced into the colony in 1620.

Landscape and environment. *Physical features.* One of the most northerly groups of coral islands in the world, Bermuda consists mainly of chalky deposits capping a volcanic cone rising more than 14,000 feet from the ocean floor. The islands are generally hilly, having a maximum elevation of about 260 feet above sea level;



there are a number of fertile depressions. There are a few marshes and brackish ponds, but no rivers or lakes.

Climate. The climate is mild, humid, equable, and frost-free. August is the hottest month and February the coldest. The annual maximum temperature averages about 90° F (32° C), and the annual minimum 47° F (8° C). The annual mean temperature averages 70° F (21° C). The average mean humidity is about 77 percent, and an average annual precipitation of about 57 inches is distributed fairly evenly throughout the year.

Vegetation and animal life. Vegetation growth is luxuriant, and there are more than 950 kinds of flowering trees, plants, and vines. Originally, an indigenous Bermuda cedar covered the islands densely, but 80 percent of these trees were killed by disease. Bermuda has no native mammals—wild pigs left by shipwrecks have now died out—and it has only one native reptile, the lizard. There are no dangerous insects or reptiles. Migratory birds visit the islands. The cahow, or Bermuda petrel, once considered extinct, breeds here, as do a few other native land birds. The surrounding waters abound with fish and lobsters.

History. Discovery and settlement

Fernández de Oviedo, a Spanish navigator and historian who sailed close to the Bermudas in 1515, first attributed their discovery to his countryman, Juan Bermúdez. The exact discovery date is unknown, but an Italian map published in 1511 shows the islands in an approximately correct position. A 17th-century French cartographer gave the date of discovery as 1503. The islands remained uninhabited until 1609, when the English admiral Sir George Somers beached his flagship "Sea Venture" after hitting one of the reefs. Sir George died in Bermuda the following year. His companions, ignorant of Juan Bermúdez, named the islands Somers Islands, which is still a secondary designation.

In 1612 Bermuda was included in the third charter of the Virginia Company, and 60 English settlers were sent out to it. In 1616 an Indian and a Negro were brought from the Bahamas as slaves. So many others followed that the slave population came to outnumber the whites. They were not freed until slavery was abolished throughout the British Empire in 1834.

The islands remained under company administration until 1684, when the government passed jointly to the English Crown and the company; the rights of the inhabitants remained undisturbed, and Bermuda never became a Crown Colony under the total jurisdiction of the sovereign. In 1815 the capital was transferred from Saint George on Saint Georges Island to Hamilton on Great Bermuda. Much of Bermuda's subsequent history took the form of a succession of economic booms and depressions. Blockade running during the United States Civil War and rum smuggling during Prohibition (the legal prevention of the manufacture, sale, or transportation of alcoholic beverages in the United States from 1919 to 1933), as well as privateering (illegal only after about 1700), made Bermuda a rich country. In the 20th century Bermuda developed a flourishing tourist industry.

In 1941, naval and air bases in Bermuda were leased for 99 years to the United States government. The British garrison in Bermuda was finally withdrawn in 1957. The British naval dockyard was closed in the 1950s, but a small naval establishment is still maintained.

In 1963 Bermuda's first political party, the Progressive Labour Party (PLP), was formed. Its members—nine black and one white—claimed generally to represent the majority of the nonwhite population and advocated total independence from Britain and the introduction of an income tax. In 1964 independent members of the legislature formed the United Bermuda Party (UBP), whose 22 white and eight black members were committed to racial integration. In 1968 a new constitution, placing strong executive powers in the hands of the leader of the party elected to office, came into force,

Before the general elections in 1968 the colony experienced the worst outbreak of civil disorder in its history, largely the result of deep racial tensions. Rioting and arson, much of it by young blacks, caused the governor to seek military assistance from Britain. In the election the United Bermuda Party won 30 seats, the Progressive Labour Party the other ten—a result that was seen as an expression of support for the continued status of Bermuda as a British colony. Tension was heightened in March 1973 when the governor, Sir Richard Sharples, was shot to death; this followed an earlier assassination of Police Commissioner George Duckett.

Bermuda, Area and Population

	area		population*	
	sq mi	sq km	1960 census	1970 census
Municipalities				
Hamilton	0.27	0.70	2,800	2,100
Saint George	0.53	1.37	1,300	1,800
Parishes				
Devonshire	1.89	4.90	4,800	6,300
Hamilton	1.95	5.05	2,700	3,300
Paget	2.03	5.26	3,900	4,600
Pembroke	1.79	4.64	11,400	11,700
Saint Georges	3.72	9.63	2,100	2,300
Sandys	1.94	5.02	4,700	5,800
Smiths	1.87	4.84	2,300	4,200
Southampton	2.21	5.72	2,500	3,900
Warwick	2.21	5.72	4,200	6,500
Total Bermuda	20.41	52.86†	42,600†‡	52,300†§

*Rounded to nearest 100 persons. †Figures do not add to total given because of rounding. ‡De jure. Excludes 11,000 members of various armed forces and 3,000 tourists and transients. Total de facto population was 56,000. Exact de jure total of 42,640 was later adjusted upward to 44,617, but no comparable adjustment was made for separate parishes. §De jure. Excludes armed forces, tourists, and transients, for which no separate figures were available.
Source: Official government figures.

The people. The 1970 census recorded a population of 52,300. Three-fifths of these were black, mainly descendants of slaves brought from Africa before the abolition of the trans-Atlantic slave trade by Britain in 1807. The white population came primarily from the British Isles, although a few were descended from Portuguese labourers brought from Madeira and the Azores in the mid-19th century. In addition, some 3,000 American, Canadian, and British forces were stationed there.

The ethnic pattern

The main language spoken is English, but some Portuguese is heard. Almost half the population are members of the Church of England; more than 5,000 are members of the African Methodist Episcopal Church, and a smaller number are members of the Roman Catholic Church. Altogether 21 religious faiths are represented.

The resident population is fairly evenly distributed over the main island of Great Bermuda; small settlements are established elsewhere. While the majority of visitors stay on the main island, many elect to stay on Saint Georges Island. Hamilton, named for a former governor, is one of the world's smallest cities; it has an area of about 180 acres and a population of about 2,100. Tucked between the hillside and the sea, it is a constant beehive of activity. The port of Hamilton is a deep, land-locked harbour; tourist liners berth only a few yards from the shops, which are large and well stocked with quality goods from around the world.

Bermuda's architecture is distinctive. White terraced roofs, designed to catch rainwater, top pastel-coloured houses. There are narrow lanes, pink-tinted beaches, and a remarkably clear, azure sea. The social atmosphere is varied—Somerset being rural, Hamilton lively, Saint Georges historic, and Tucker's Town exclusive.

The economy. The economic structure of Bermuda is based on tourism, which brings in over $70,000,000 a year, and on income derived from the military bases. (In 1970 the monetary unit was changed from the pound sterling to the Bermuda dollar, which is at par with the United States dollar.) More than 300,000 tourists visit the islands annually. Of the government's annual revenue, about half is obtained from customs receipts. Despite both a cost of living and a population density that are among the world's highest, Bermuda has almost no poverty or unemployment. No income or property taxes are levied, nor does the government receive foreign aid.

Local industries

Local industries are small and are confined to small-boat building and ship repairing, and the manufacture of pharmaceuticals, concentrated essences (such as perfumes, flavourings, and extracts), and handicraft souvenirs. The government, however, began to encourage light industry (such as ship repairing, boat building, and furniture making) in the late 1960s, mainly around the former naval dock areas on Ireland Island. There is a little commercial fishing; agriculture is purely local in scope. Neither is sufficient to satisfy local demand, and foodstuffs from the United States are a major import. Easter lilies and bananas are exported on a small scale.

Tax advantages have encouraged a considerable number of foreign companies to register in Bermuda. This has produced a large amount of banking and other financial activity.

Transportation. There are approximately 132 miles of public roads, most of them paved. Bridges and causeways link Great Bermuda with the other main islands (Somerset, Watford, Boaz, Ireland, St. Georges, St. Davids, and Coney). Ferries provide another interisland link. Cars, prohibited before 1946, are restricted to one per family and are limited in size and power. Many motorized bicycles are in use. The Ports Authority maintains tugs, passenger tenders, and a floating dry dock. In addition to the port of Hamilton, there are also offshore anchorages; both Saint Georges and the duty-free port at Ireland Island have considerable berthing space.

A United States naval air station at Kindley Field on Saint Georges Island is the only airport and is used by both military aircraft and by passenger aircraft of major international airlines. It has facilities for "jumbo" jets and can handle passengers at the rate of 700 an hour. A train service, introduced in 1931, was scrapped in 1947 and has not been replaced. Telephone services are provided by the Bermuda Telephone Company, Ltd.

Administration and social conditions. *Administration.* Under the constitution of 1968, the Queen appoints the governor, who retains responsibility for external affairs, defense, internal security, and the police. In other matters he acts on the advice of the Executive Council, which consists of the leader of the majority party in the legislature and at least six other members. The legislature consists of the Queen, the Legislative Council of 11 members (five nominated by the governor, four by the government leader or head of the Executive Council, and two by the leader of the opposition), and an elected 40-member House of Assembly. The judiciary consists of the Court of Appeal, the Supreme Court, and the Magistracy. There are five penal institutions on the islands.

Each of the nine parishes annually appoints its own vestry (local government authority), which has power to levy taxes and manage local affairs. There are two municipalities, the city of Hamilton and the town of Saint George.

Social conditions. Education is compulsory from five to 16 years of age. Primary education is provided free

Education

at government-maintained and aided schools. About one-fifth of the government's expenditures are devoted to education. Secondary education is provided at both state-aided and private schools. There are no institutions for higher education, although there are government scholarships available for students to study at universities abroad. Vocational courses are offered at a technical institute and at a catering college.

The King Edward VII Memorial Hospital is the only general hospital, although there are also two residential hospitals for geriatric cases and one for all types of mental disorder.

The state provides no general medical scheme. The Department of Health and Welfare, however, provides free services for babies and children of school age.

Cultural life. Cultural societies abound, the most active being devoted to art, drama, and choral singing. There are several cinemas. Kiteflying at Easter is a custom peculiar to Bermuda. Some of the folklore of Africa is preserved by the Gombey Dancers, black troupes who have developed elaborate costumes and on public holidays dance to the strong rhythms of drums. Many British folk traditions, holidays, and ceremonies are observed. In the early 1970s there were four radio stations and two television stations, as well as one daily and three weekly newspapers. There is a subscription library with two small branches. The colonial archives contain records dating back to the 17th century.

Future prospects. Bermuda remains a playground for tourists and a tax haven for many companies. The island's future depends on the continued stability of its political and economic life.

While there is a certain amount of racial unrest, in general white and black Bermudians live together harmoniously. Progress toward full integration is slowly but surely being made. Blacks held the four most important governmental posts in the early 1970s, and hopes were expressed that the black population would gain an increased number of qualified representatives in the future.

BIBLIOGRAPHY. The following books are the most scholarly and comprehensive accounts of Bermuda's early history to 1897: J.H. LEFROY, *Memorials of the Discovery and Early Settlement of the Bermudas . . .*, 2 vol. (1877–79), compiled from the early colonial records and manuscripts by a former governor; JEAN DE CHANTAL KENNEDY, *Biography of a Colonial Town* (1961), a social history of Hamilton from 1790 to 1897; HENRY CAMPBELL WILKINSON, *Adventures of Bermuda: A History of the Island from Its Discovery Until the Dissolution of the Somers Island Company in 1684*, 2nd ed. (1958), and *Bermuda in the Old Empire: A History of the Island from the Dissolution of the Somers Island Company Until the End of the American Revolutionary War, 1684–1784* (1950). General works include: *Fodor's Guide to the Caribbean, Bahamas, and Bermuda*, an annual travel guide containing miscellaneous information; JOHN CROCKER, *The Centaur Guide to Bermuda, the Bahamas, Hispaniola, Puerto Rico and the Virgin Islands* (issued at frequent intervals); and TERRY TUCKER, *Islands of Bermuda* (1970), which lists each island and islet alphabetically and gives brief historical data on each as well as an aerial photograph; the only source for much information, especially on the smaller islands. See also official publications, such as the BERMUDA, CENSUS COMMITTEE, *Bermuda Census, 1960*, and early information from the 1971 census; the *Registrar General Annual Reports;* and the BERMUDA GOVERNMENT, *Annual Report*. Additional information may be found in the *West Indies and Caribbean Yearbook; the Yearbook of the Commonwealth* (HMSO); and pamphlets of the PUBLIC INFORMATION OFFICE.

(Pa.H.)

Bernard, Claude

A leading 19th-century French physiologist, Claude Bernard is known chiefly for discoveries concerning the role of the pancreas in digestion, the glycogenic function of the liver, and the regulation of the blood supply by the vasomotor nerves. His findings became part of the factual content of physiology. On a broader stage, Bernard played a role in establishing the principles of experimentation in the life sciences, advancing beyond the vitalism and indeterminism of earlier physiologists to become one of the founders of experimental medicine. His most seminal contribution was his concept of the internal environment of the organism, which led to the present understanding of homeostasis—*i.e.*, the self-regulation of vital processes.

Bernard was born on July 12, 1813, near the village of Saint-Julien in Beaujolais. His father, Pierre, was a wine-grower; his mother, Jeanne Saulnier, of peasant background. There was one other child, a girl. When Claude was very young, his father failed in a wine-marketing venture and tried to make ends meet by teaching school. Despite his efforts, the family never prospered; and, when he died, the survivors were left in debt. It was his mother whom Claude Bernard was always to cherish and remember. Educational opportunities were scarce for a poor winegrower's son in the France of Louis XVIII. The boy studied Latin with the local priest and then was enrolled in a Jesuit-conducted school at Villefranche, where no natural science was taught. At 18 he ended his secondary schooling at Thoissey without a diploma and was apprenticed to an apothecary in a Lyon suburb.

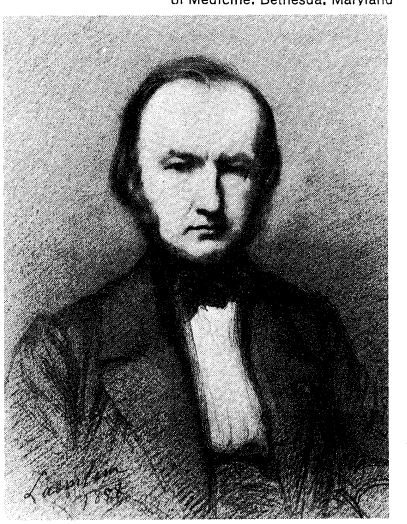
Bernard's days were spent in menial tasks relieved by errands to a veterinary school or, on his rare times off, by visits to a theatre. He wrote a playlet, *La Rose du Rhône*, now lost, and received 100 francs for it. Next, he began writing *Arthur de Bretagne*, a historical drama in five acts. His employer was not pleased, however, and the apprenticeship came to a halt, the youth returning home in July 1833. By November 1834 he was in Paris with the completed manuscript of *Arthur de Bretagne* and a letter of introduction. The literary critic Saint-Marc Girardin read his play and advised him to try medicine instead of playwriting. Many years later the two met again, as members of the French Academy.

Early medical training

Bernard enrolled that same winter in the school of medicine and, in due course, was admitted as an extern in the hospitals. Outwardly reserved and even shy at that time, he had an inner strength that was to overcome poverty and discouragements. Of 29 students passing the examination for the internship, Bernard ranked 26th. Serving in Paris hospitals were the celebrated doctors Pierre Rayer and François Magendie, and Bernard studied under the latter at both the Hôtel-Dieu and the Collège de France. Magendie noticed Bernard's skillful dissections, but his gruff manner disheartened the student, and Bernard almost resigned himself to settling in Beaujolais in a country practice. Fortunately, Rayer gave him new hope, and Magendie took him on as a research assistant.

Bernard became involved in Magendie's research on spinal nerves. His first publication dealt with another nerve, the chorda tympani, while his medical dissertation was devoted to the function of the gastric juice in nutrition (1843). These maiden publications were prophetic, for much of his later research concerned neurology and metabolism. Failing in the examination that would have qualified him to teach in the medical school,

By courtesy of the National Library of Medicine, Bethesda, Maryland



Bernard, lithograph by A. Laemlein, 1858.

he collaborated with others in research on digestion and on the exotic poison curare, thus treading two paths that would lead him to fame. He was rather old at 31 to be content with a research assistantship, however, and resigned late in 1844. Left in financial straits, he turned his thoughts again toward medical practice.

To save his research career, a friend arranged a marriage of convenience for him with Marie-Françoise Martin, daughter of a Paris doctor, who brought him a dowry of 60,000 francs. The marriage was destined to be painfully unhappy. Mme Bernard would later resent having her dowry go to pay her father-in-law's debts, would hate her mother-in-law, would object to Bernard's annual vacations at Saint-Julien, would denounce his practice of vivisection, and would turn their two daughters against him. She wanted him to be a society doctor instead of an illustrious scientist. Their separation was to follow his election to the French Academy and his appointment to the Imperial Senate—a poignant contrast between personal unhappiness and public renown.

All of that, however, was in the future in 1845; his first son was born in 1846 but did not survive. In 1847 his father died, his daughter, Tony, was born, and he became Magendie's substitute at the Collège de France. This period was marked by a veritable explosion of discoveries, beginning in 1846, when Bernard solved the mystery of the carnivorous rabbits. Puzzled one day by the chance observation that some rabbits were passing clear—not cloudy—urine, just like meat-eating animals, he inferred that they had not been fed and were subsisting on their own tissues. He confirmed his hypothesis by feeding meat to the famished animals. An autopsy of the rabbits disclosed an unexpected but remarkable discovery concerning the role of the pancreas in digestion: the secretions of the pancreas broke down fat molecules into fatty acids and glycerin. Next, he took a step toward explaining diabetes by showing that sugar in the blood is not always a symptom of the disease, though his hopes to conquer diabetes were premature because of the still too many unknowns in the problem.

Discovery of the glycogenic function of the liver

His work on the pancreas led to research on the liver, culminating in his second great discovery, the glycogenic function of the liver; and his dissertation on this subject won him the doctorate of science in 1853. Simultaneously, he was nearing his third great achievement—explanation of the regulation of the blood supply by the vasomotor nerves. This experiment overlapped in time with a fourth discovery, one concerning curare. He showed how this dread poison causes paralysis and death by attacking the motor nerves, while having no effect on the sensory nerves. He demonstrated that, because of this selectivity, curare could be used as an experimental tool in differentiating neuromuscular from primary muscular mechanisms.

Within less than a decade, from obscurity in the shadow of Magendie, he had risen to a commanding position in science. In 1854 a chair of general physiology was created for him in the Sorbonne, and he was elected to the Academy of Sciences. When Magendie died in 1855, Bernard succeeded him in the Collège de France.

Magendie's empirical method of conducting experiments without a guiding hypothesis was by then out-of-date, partly as a result of his own discoveries. Bernard's historic role was to demonstrate the experimenter's need for a guiding hypothesis to be either confirmed or refuted by the results.

For various reasons, a shift was occurring in Bernard's scientific interests. The productive researcher was turning into a philosopher of science. Failing health after 1860 led him to spend more time at Saint-Julien, less time in the laboratory. Louis Pasteur would blame the unhealthful conditions here for his colleague's long illness. By odd coincidence, Bernard suffered apparently from chronic enteritis, with symptoms affecting the pancreas and the liver. By way of compensation, the enforced leisure left him time for reflection, out of which would come his masterpiece, *Introduction à la médecine expérimentale* (1865; *An Introduction to the Study of Experimental Medicine*, 1927).

This work was planned as a preface, if a very long one, to a work of greater magnitude, never completed. Bernard's aim in the *Introduction* was to demonstrate that medicine, in order to progress, must be founded on experimental physiology. The other points in his argument are that (1) the physical and chemical sciences provide the foundation for physiology, although it is not reducible to them; (2) the notion of "vital force" does not explain life; (3) vivisection is indispensable for physiological research; and (4) biology depends on the principle of scientific determinism; *i.e.*, the principle that, under identical conditions, the phenomena will be identical. He was skeptical regarding statistics and did not anticipate the later usefulness that statistical techniques would achieve. Still germane for modern science is his presentation of the concept of the *milieu intérieur*, or "internal environment."

Reasons for experimentation

The book brought new honours to Bernard, notably election to the French Academy. After the breakup of his marriage, he found some consolation in a platonic friendship with Marie Raffalovich, a handsome woman of Russian-Jewish origin. She attended his lectures and translated scientific works for him. His friends included such literary figures as Ernest Renan, Hippolyte Taine, and the Goncourts, besides such scientists as Pasteur and Marcelin Berthelot.

The most famous of the students trained by Bernard were Albert Dastre, Paul Bert, and Arsène d'Arsonval. Bert succeeded Bernard in the Sorbonne when the latter transferred to the Musée d'Histoire Naturelle. Bernard's own experiments were taking new directions. The phenomena common to animals and plants formed the subject of lectures published posthumously. He also began research on fermentation. His findings were published after his death by Berthelot; and, because they conflicted with Pasteur's views, cast a cloud over the microbe hunter's memory of his late colleague.

Bernard's health had declined precipitously in the autumn of 1877. On New Year's Day he caught cold, and inflammation of the kidneys set in. Soon he was confined to his bed. Mme Raffalovich and her daughter nursed him, students and friends gathered around; his daughter Tony stopped short of his threshold. A travelling rug was laid over his legs. He is reported to have said: "This time it will serve me for the journey from which there is no return." His sister called in the priest to take his confession, though his closest friends were convinced that there was no deathbed conversion. He died February 10, 1878. There was a national funeral, the first ever for a scientist in France.

BIBLIOGRAPHY. Editions of Bernard's works include *Introduction à la médecine expérimentale* (1865; Eng. trans. by H.C. GREEN, with a new foreword by I. BERNARD COHEN, 1957); *L'Oeuvre de Claude Bernard*, with addresses by ERNEST RENAN *et al.*, and a bibliography by G. MALLOIZEL (1881); and *Cahier de Notes 1850–1860*, a useful edition of a Bernard notebook by M.D. GRMEK (1965). J.M.D. OLMSTED, *Claude Bernard: Physiologist* (1938), is the fullest biography, but may be supplemented by J.M.D. and E.H. OLMSTED, *Claude Bernard and the Experimental Method in Medicine* (1952); and REINO VIRTANEN, *Claude Bernard and His Place in the History of Ideas* (1960).

(R.Vi.)

Bernard of Clairvaux, Saint

Bernard, nobleman from Fontaine-les-Dijon, France, a Cistercian monk and mystic, and abbot of Clairvaux, was a contemplative-activist who exercised a tremendous influence on the political, literary, and religious life of Western civilization.

The sources of this influence were threefold: his profound and intense love affair with godly wisdom; his impulsive temperament manifested in his propensity to accept evidence without properly evaluating it; and his ardent zeal that made him quick to intervene against any threats to the integrity of the faith.

Early preparations

Born in 1090 of Burgundian landowning aristocracy, probably at Fontaines, near Dijon, Bernard grew up in a family of five brothers and one sister. The familial atmosphere engendered in him a deep respect for mercy,



St. Bernard of Clairvaux, detail of an altarpiece by the Florentine School, early 15th century. In the Staatliche Museen zu Berlin.
By courtesy of the Staatliche Museen zu Berlin, D.D.R.

justice, and loyal affection for others. Faith and morals were taken seriously, but without priggishness. Both his parents were exceptional models of virtue. It is said that his mother, Aleth, exerted a virtuous influence upon Bernard only second to what Monica had done for Augustine of Hippo in the 5th century. Her death, in 1107, so affected Bernard that he claimed that this is when his "long path to complete conversion" began. He turned away from his literary education, begun at the school at Châtillon-sur-Seine, and from ecclesiastical advancement, toward a life of renunciation and solitude.

Bernard sought the counsel of the abbot of Cîteaux, Stephen Harding, and decided to enter this struggling, small, new community that had been established by Robert of Molesmes in 1098 as an effort to restore Benedictinism to a more primitive and austere pattern of life. Bernard took his time in terminating his domestic affairs and in persuading his brothers and some 25 companions to join him. He entered the Cîteaux community in 1112, and from then until 1115 he cultivated his spiritual and theological studies.

Bernard's struggles with the flesh during this period may account for his early and rather consistent penchant for physical austerities. He was plagued most of his life by impaired health, which took the form of anemia, migraine, gastritis, hypertension, and an atrophied sense of taste. It is extraordinary that he was so prolific in his writings and in public activity while suffering such physical illness.

Founder and abbot of Clairvaux

In 1115 Stephen Harding appointed him to lead a small group of monks to establish a monastery at Clairvaux, on the borders of Burgundy and Champagne. Four brothers, an uncle, two cousins, an architect, and two seasoned monks under the leadership of Bernard endured extreme deprivations for well over a decade before Clairvaux was self-sufficient. Meanwhile, as Bernard's health worsened, his spirituality deepened. Under pressure from his ecclesiastical superiors and his friends, notably the bishop and scholar William of Champeaux, he retired to a hut near the monastery and to the discipline of a quack physician. It was here that his first writings evolved. They are characterized by repetition of references to the Church Fathers, by the use of analogues, etymologies, alliterations, biblical symbols, and they are imbued with resonance and poetic genius. It was here, also, that he produced a small but complete treatise on Mariology (study of doctrines and dogmas concerning the Virgin Mary), "Praises of the Virgin Mother." Bernard was to become a major champion of a moderate cult of the Virgin, though he did not support the notion of Mary's immaculate conception.

By 1119 the Cistercians had a charter approved by Pope Calixtus II for nine abbeys under the primacy of the abbot of Cîteaux. Bernard struggled and learned to live with the inevitable tension created by his desire to serve others in charity through obedience and his desire to cultivate his inner life by remaining in his monastic enclosure. His more than 300 letters and sermons manifest his quest to combine a mystical life of absorption in God with his friendship for those in misery and his concern for the faithful execution of responsibilities as a guardian of the life of the church.

It was a time when Bernard was experiencing what he apprehended as the divine in a mystical and intuitive manner. He could claim a form of higher knowledge that is the complement and fruition of faith and that reaches completion in prayer and contemplation. He could also commune with nature and say:

> Believe me, for I know, you will find something far greater in the woods than in books. Stones and trees will teach you that which you cannot learn from the masters.

After writing a eulogy for the new military order of the Knights Templars, he would write about the fundamentals of the Christian's spiritual life, namely, the contemplation and imitation of Christ, which he expressed in his sermons "The Steps of Humility" and "The Love of God."

Pillar of the church

The mature and most active phase of Bernard's career occurred between 1130 and 1145. In these years both Clairvaux and Rome, the centre of gravity of medieval Christendom, focussed upon Bernard. Mediator and counsellor for several civil and ecclesiastical councils and for theological debates during seven years of papal disunity, he nevertheless found time to produce an extensive number of sermons on the Song of Solomon. As the confidant of five popes, he considered it his role to assist in healing the church of wounds inflicted by the antipopes (those elected pope contrary to prevailing clerical procedures), to oppose the rationalistic influence of the greatest and most popular dialectician of the age, Peter Abelard, and to cultivate the friendship of the greatest churchmen of the time. He could also rebuke a pope, as he did in his letter to Innocent II:

> There is but one opinion among all the faithful shepherds among us, namely, that justice is vanishing in the Church, that the power of the keys is gone, that episcopal authority is altogether turning rotten while not a bishop is able to avenge the wrongs done to God, nor is allowed to punish any misdeeds whatever, not even in his own diocese (parochia). And the cause of this they put down to you and the Roman Court.

Bernard's confrontations with Abelard ended in inevitable opposition because of their significant differences of temperament and attitudes. In contrast with the tradition of "silent opposition" by those of the school of monastic spirituality, Bernard vigorously denounced dialectical Scholasticism as degrading God's mysteries, as one technique among others, though tending to exalt itself above the alleged limits of faith. One seeks God by learning to live in a school of charity and not through "scandalous curiosity," he held. "We search in a worthier manner, we discover with greater facility through prayer than through disputation." Possession of love is the first condition of the knowledge of God. However, Bernard finally claimed a victory over Abelard, not because of skill or

cogency in argument but because of his homiletical denunciation and his favoured position with the bishops and the papacy.

Pope Eugenius III and King Louis VII of France induced Bernard to promote the cause of a Second Crusade (1147–49) to quell the prospect of a great Muslim surge engulfing both Latin and Greek Orthodox Christians. The crusade ended in failure because of Bernard's inability to account for the quarrelsome nature of politics, peoples, dynasties, and adventurers. He was an idealist with the ascetic ideals of Cîteaux grafted upon those of his father's knightly tradition and his mother's piety, who read into the hearts of the crusaders—many of whom were bloodthirsty fanatics—his own integrity of motive.

In his remaining years he participated in the condemnation of Gilbert de La Porrée—a scholarly dialectician and bishop of Poitiers who held that Christ's divine nature was only a human concept. He exhorted Pope Eugenius to stress his role as spiritual leader of the church over his role as leader of a great temporal power, and he was a major figure in church councils. His greatest literary endeavour, "Sermons on the Canticle of Canticles," was written during this active time. It revealed his teaching, often described as "sweet as honey," as in his later title *doctor mellifluus*. It was a love song supreme: "The Father is never fully known if He is not loved perfectly." Add to this one of Bernard's favourite prayers, "Whence arises the love of God? From God. And what is the measure of this love? To love without measure," and one has a key to his doctrine. Bernard died at Clairvaux on August 20, 1153, was canonized as a saint in 1174, was declared a doctor of the church in 1830, and was extolled in 1953 as *doctor mellifluus* in an encyclical of Pius XII.

BIBLIOGRAPHY. E.C. BUTLER, *Western Mysticism: The Teaching of Saints Augustine, Gregory and Bernard on Contemplation and the Contemplative Life* (1922), on the mystical doctrine of Bernard; HENRI DANIEL-ROPS, *Saint Bernard et ses fils* (1962; Eng. trans., *Bernard of Clairvaux*, 1964), a popular biography and description of the historical development of the Cistercian Order; ETIENNE GILSON, *La Théologie mystique de Saint Bernard* (1934; Eng. trans., *The Mystical Theology of Saint Bernard*, 1940), on the mystical doctrine of Bernard and his contemporaries; BRUNO SCOTT JAMES, *St. Bernard of Clairvaux* (1957), a personality study; JEAN LECLERCQ, *L'Amour des lettres et le désir de Dieu* (1957; Eng. trans., *The Love of Learning and the Desire for God: A Study of Monastic Culture*, 1961), indispensable for an adequate and appreciative understanding of the entire cultural context of Bernard; DENIS MEADOWS, *A Saint and a Half: A New Interpretation of Abelard and St. Bernard of Clairvaux* (1963), a quite thorough and readable treatment of the controversy and personalities involved; THOMAS MERTON, *The Last of the Fathers: Saint Bernard of Clairvaux and the Encyclical Letter, Doctor Mellifluus* (1954); ALBERT VICTOR MURRAY, *Abelard and St. Bernard: A Study in Twelfth-Century Modernism* (1967), a technical and scholarly approach to the issue; *St. Bernard of Clairvaux: The Story of His Life As Recorded in the Vita Prima Bernardi . . .*, trans. by G. WEBB and ADRIAN WALKER (1960), a chief biographical source by several contemporaries of Bernard, which therefore must be read with reserve as it presents a different type of historical biography; E. VACANDARD, *Vie de saint Bernard, abbé de Clairvaux*, 2 vol. (1895), was the definitive chronology and remains quite authoritative; I. VALLERY-RADOT, *Bernard de Fontaines: Abbé de Clairvaux*, 2 vol. (1963–69), highly definitive as a chronological study as well as reviewing all contributions since the 19th century; WATKINS WILLIAMS, *St. Bernard of Clairvaux* (1935), concentrates upon Bernard's political activity.

Works: The *Works* (of St. Bernard of Clairvaux), trans. by a Priest of Mount Melleray, 6 vol. (1920–25); and *The Works of Bernard of Clairvaux*, "Cistercian Fathers Series," vol. 1 (1970–), two English-language collections; *Sancti Bernardi Opera*, ed. by J. LECLERCQ, C.H. TALBOT, and H.M. ROCHAIS, 6 vol. (1957–), the best critical text.

(J.R.M.)

Bernhardt, Sarah

Her wonderful voice and gift for emotional acting, combined with a vivid and unconventional personality, made Sarah Bernhardt one of the best known figures in the history of the stage. She was born in Paris on October 22/23, 1844, the illegitimate daughter of Judith Van Hard, a notorious Dutch courtesan who had established herself in Paris, and Edouard Bernard, a law student. As the presence of a baby interfered with her mother's life, Sarah was brought up at first in a *pension*, and, later, in a convent. A difficult, willful child of delicate health, she wanted to become a nun, but one of her mother's lovers, the duc de Morny, Napoleon III's half-brother, decided that she should be an actress, and, when she was 16, arranged for her to enter the Conservatoire, the government-sponsored school of acting. She was not considered a particularly promising student, and, although she revered some of her teachers, she regarded the Conservatoire's methods as antiquated and too deeply steeped in tradition.

Early life and training

By courtesy of the Library of Congress, Washington, D.C.



Sarah Bernhardt, photograph by Napoleon Sarony, 1880.

Sarah Bernhardt left the Conservatoire in 1862, and, thanks to the duc de Morny's influence, was accepted by the national theatre company, the Comédie-Française, as a beginner on probation. During the obligatory three debuts required of probationers, she was scarcely noticed by the critics. Her contract with the Comédie-Française was cancelled in 1863 after she had slapped the face of a senior actress, who had been rude to her younger sister. For a time she found employment at the Théâtre du Gymnase-Dramatique. After playing the role of a foolish Russian princess, she entered a period of soul-searching, questioning her talent for acting. During these critical months she became the mistress of Henri, prince de Ligne, and gave birth to her only child, Maurice.

In 1866 she signed a contract with the Odéon theatre and, during six years of intensive work with a congenial company there, gradually established her reputation. Her first resounding success was as Anna Damby in the 1868 revival of *Kean*, by the novelist and playwright Alexandre Dumas *père*. Bernhardt's greatest triumph at the Odéon, however, came in 1869, when she played the minstrel Zanetto in the young dramatist François Coppée's one-act verse play *Le Passant* ("The Passerby")—a part that she played again in a command performance before Napoleon III.

During the Franco-German War in 1870, she organized a military hospital in the Odéon theatre. After the war, the reopened Odéon paid tribute to France's great 19th-century writer Victor Hugo with a production of his verse-play *Ruy Blas*. As Queen Maria, Bernhardt charmed her audiences with the lyrical quality of her voice. It was then that Hugo coined the phrase "golden voice," though her critics usually called her voice "silvery," as resembling the tones of a flute.

In 1872 she left the Odéon and returned to the Comédie-Française. One of her remarkable successes there was in the title role of Voltaire's *Zaïre* (1874); generally,

however, she received only minor parts. Eventually, she was given the chance to play the title role in Racine's *Phèdre,* a part for which the critics supposed she lacked the resources needed to portray violent passion. Her performance, however, made them revise their estimate and write enthusiastic reviews. In Hugo's *Hernani,* her portrayal of Doña Sol was said to have brought tears to the author's eyes.

In 1879, when the Comédie-Française appeared in London, Bernhardt played in the second act of *Phèdre* and triumphed in spite of a violent attack of stage fright. An international career lay before her; yet critics still refused to succumb to her spell—the novelist Henry James commented on the "admirable delicacy and grace" with which she had handled the plaintive passages, but added that "in the violent scenes she forces her note beyond all reason, and becomes painfully shrill and modern."

Her success with London audiences made Sarah Bernhardt overbearing and led to fits of bad temper in the sedate halls of the Comédie-Française. The break came in 1880, when, forming her own company, she became a travelling star and international idol. She appeared fairly regularly in England but extended her itinerary to the European continent, the United States, and Canada. New York saw her for the first time on November 8, 1880. Eight visits to the United States followed. In 1891–93 Bernhardt undertook a world tour that included Australia and South America. Aside from her appearances as Phèdre, there were two parts that audiences all over the world clamoured to see her in: Marguérite Gautier, the redeemed courtesan in *La Dame aux Camélias* of Alexandre Dumas *fils* and the title role of the popular playwright Eugène Scribe's *Adrienne Lecouvreur.*

International success

In the 1880s a new element had entered her artistic life with the emergence of Victorien Sardou as chief playwright for melodrama. With Bernhardt in mind, Sardou wrote *Fédora* (1882), *Théodora* (1884), *La Tosca* (1887), and *Cléopâtre* (1890). Sardou, directing his own plays in which she starred, taught her a broad, flamboyant style of acting, relying for effect on lavish decors, exotic costumes, and pantomimic action.

In 1905, during a South American tour, she injured her right knee when jumping off the parapet in the last scene of *La Tosca.* By 1915 gangrene had set in, and her leg had to be amputated. Carried about in a litter chair the patriotic Bernhardt insisted on visiting the soldiers at the front during World War I. In 1916 she began her last tour of the United States. Her indomitable spirit sustained her during 18 gruelling months on the road. In November 1918 she arrived back in France but soon set out on another European tour, playing parts she could act while seated. New roles were provided for her by the playwrights Louis Verneuil, Maurice Rostand, and Sacha Guitry. She collapsed during the dress rehearsal of the Guitry play *Un Sujet de roman,* recovered again sufficiently to take an interest in the Hollywood motion picture *La Voyante,* which was filmed in her own house in Paris, where she died on March 26, 1923.

Her hobbies and writings

For a while Sarah Bernhardt had attempted sculpturing and painting as a hobby. In 1920 she published a novel, *Petite idole,* not without interest since the actress-heroine constitutes an idealization of Bernhardt's own career and ambitions. Facts and fiction are difficult to disentangle in her autobiography, *Ma Double Vie; mémoires de Sarah Bernhardt* (1907), written in part to counteract the slanderous and scabrous *Les Mémoires de Sarah Barnum* (1883) by Marie Colombier, who had capitalized on the actual and alleged eccentricities of the actress' private life. Bernhardt's treatise on acting, *L'Art du théâtre* (1923), is revealing in its sections on voice training: the actress had always considered voice as the key to dramatic character.

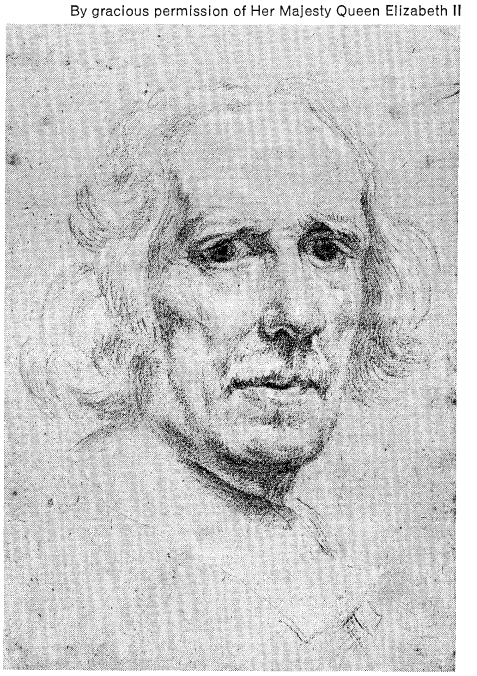
BIBLIOGRAPHY. M. COLOMBIER, *Le Voyage de Sarah Bernhardt en Amérique* (1881), an account written by an actress in the company on Bernhardt's first American tour; M. BARING, *Sarah Bernhardt* (1933), a concise biography based on personal memories; E. PRONIER, *Une Vie au théâtre: Sarah Bernhardt* (1942), the only scholarly and critical work eliminating much legendary material; L. VERNEUIL, *La Vie merveilleuse de Sarah Bernhardt* (1942; Eng. trans., *The Fabulous Life of Sarah Bernhardt,* 1942), an intimate biography by a dramatist who provided Bernhardt with two acting vehicles; MAY AGATE, *Madame Sarah* (1945), valuable information on Bernhardt's acting and teaching methods; A.W. ROW, *Sarah the Divine* (1957), many personal recollections of the author who was Bernhardt's press representative during the last years of her life; J. RICHARDSON, *Sarah Bernhardt* (1959), a highly readable study of Bernhardt as an actress and as a woman; A. CASTELOT, *Sarah Bernhardt* (1961), an excellent book in French; C.O. SKINNER, *Madame Sarah* (1967), a successful, primarily biographical attempt at separating facts from fiction.

(A.M.N.)

Bernini, Gian Lorenzo

Perhaps the greatest sculptor of the 17th century and one of its outstanding architects, Gian Lorenzo Bernini was also a dramatist, stage designer, painter, and courtier. He created the Baroque style of sculpture and developed it to such an extent that other artists are of only minor importance in a discussion of the style.

By gracious permission of Her Majesty Queen Elizabeth II



Bernini, self-portrait drawing, *c.* 1665. In the Royal Library, Windsor Castle.

EARLY CAREER

Bernini was born in Naples on December 7, 1598. His career began under his father, Pietro (1562–1629), a Florentine sculptor of some talent who ultimately moved to Rome. Bernini was remarkably precocious. His first biographer, Filippo Baldinucci (1682), reports that Bernini's first portrait bust, of Bishop Giovanni Battista Santoni (Sta. Prassede, Rome), was carved when Bernini had hardly finished his tenth year. Such works gained him the patronage of Pope Paul V, who prophetically stated, "We hope this youth will become the Michelangelo of this century." The prodigy worked so diligently that his father was forced to restrain him. His hard work, however, earned the praise of the painter Annibale Carracci (1560–1609), who is reported to have said that Bernini had arrived in his childhood where others would glory to be in their old age. The numerous reports of his unique childhood talent have apparently been confirmed by the discovery of a bust, perhaps by Bernini, dated 1612, of Antonio Coppola (S. Giovanni dei Fiorentini, Rome). Although Bernini and his father occasionally collaborated in the early years, as in the "Four Seasons" (Villa Aldobrandini, Frascati), Bernini soon established himself as a wholly independent sculptor. He was strongly influenced by his close study of the antique marbles in the Vatican, and one of his early groups, "The Goat Amalthea with

the Infant Jupiter and a Faun," may have been produced as a forgery of an antique work. It shows his youthful study of the sculpture of the Hellenistic period (323–30 BC) of ancient Greek civilization but already betrays a personal style. Bernini also had an intimate knowledge of High Renaissance painting of the early 16th century. His study of Michelangelo (1475–1564) is revealed in the "St. Sebastian" (*c.* 1617; Thyssen-Bornemisza Collection, Lugano, Switzerland) carved for Cardinal Maffeo Barberini (1568–1644), who was later Pope Urban VIII and Bernini's greatest patron.

Works done for Cardinal Borghese

Bernini's early works attracted the attention of Cardinal Scipione Borghese, a member of the reigning papal family. Under his patronage, Bernini carved his first important life-size sculptural groups. The series shows Bernini's progression from the almost haphazard single view of "Aeneas, Anchises and Ascanius Fleeing Troy," to strong frontality in "Pluto and Proserpina," and then to the hallucinatory vision of "Apollo and Daphne," which was intended to be viewed from one spot as if it were a relief. In his "David" Bernini shows the figure in the act of casting his stone past the viewer at an unseen adversary. Bernini thereby included the spectator in the action, uniting real and artistic space and creating a principle that was to stand at the core of much Baroque art. Several portraits of this period paralleled the Borghese groups. The bust of Monsignor Pedro de Foix Montoya (*c.* 1622; Sta. Maria di Monserrato, Rome) shows a new awareness of the relationship between head and body and a realism that immediately became legendary. The bust of Cardinal Roberto Bellarmino (1623–24; Gesù, Rome; originally part of a larger monument) shows the great theologian in a seemingly transitory act of adoration. Bernini's ability to depict fleeting facial expressions realistically resulted from constant observation and experimentation. On one occasion, he was reported even to have placed his leg in a fire in order to study his expressions of pain. These marble works show Bernini's unparalleled virtuosity in carving the obdurate material to achieve the delicate effects usually found only in bronze sculptures. His sensual awareness of the surface textures of skin and hair and his novel sense of colour broke with the tradition of Michelangelo and marked the emergence of a new period in the history of Western sculpture.

Patronage of Urban VIII. With the pontificate of Urban VIII (1623–44) Bernini entered a period of enormous productivity and artistic development. Urban VIII urged his protégé to paint and to practice architecture. Although Bernini's paintings are of minor importance, a few, notably two self-portraits (*c.* 1620 and *c.* 1640, Borghese Gallery, Rome), show merit. His first architectural work, the remodelled church of Sta. Bibiana in Rome, also houses "Santa Bibiana," his first religious statue incorporating the physiological compositional principles of the Borghese groups. It stands above the high altar and is shown at the moment of martyrdom. Illuminated from above by a concealed natural light, she realistically gazes up at a vision of heavenly glory painted on the vault.

The baldachin for Peter's tomb

At the same time Bernini was commissioned to build a symbolic structure over the tomb of Peter. The result is the famous immense gilt-bronze baldachin executed between 1624 and 1633. The twisted columns derive from the early Christian columns that had been used in the altar screen of Old St. Peter's. The metal "drapery" recalls honorific canopies placed over persons and sites of significance. Bernini's most original contribution to the final work is the upper framework of crowning volutes supporting the orb and cross that appears to be supported by four angels. The baldachin is also a personal papal symbol. Near the top of the baldachin, putti hold aloft the papal tiara, and the entire monument is studded with bees, suns, and laurel tendrils, emblems of the Barberini family. The baldachin freely combines natural, human, architectural, and decorative forms, fusing them into a whole that functions not only as a symbol and as a tomb marker but also as a mediator between the visitor and the large dimensions of St. Peter's. The baldachin is perfectly proportioned to its setting, and one hardly realizes that it is as tall as a four-story building. Its lively outline moving upward to the triumphant crown, its dark colour heightened with burning gold, make it like a living organism. It ultimately formed the centre of a programmatic decoration designed by Bernini for the interior of St. Peter's. An unprecedented fusion of sculpture and architecture, the baldachin is the first truly Baroque monument.

Bernini next supervised the decoration of the four piers supporting the dome of St. Peter's with colossal statues related to relics preserved in them. The statues illustrate dramatic moments in the lives of the subjects. But only one of the four, "St. Longinus," is by Bernini. The soldier who had pierced Christ's side looks up into the light-filled dome as if at the Crucifixion, when he cried out, "Truly, he was the son of God."

Bernini was a fervent Catholic who attended mass every day and took Communion twice a week. Such religious works as "St. Longinus" reveal his unquestioning acceptance of the principles of the Counter-Reformation as formulated by the Council of Trent (1545–63). That council decreed that the purpose of religious art was to teach and inspire the faithful and to serve as propaganda for the Catholic Church. Accordingly, religious art should always be intelligible and realistic, and, above all, it should serve as an emotional stimulus to piety. The development of Bernini's religious art was largely determined by his conscientious efforts to conform to those principles.

Bernini made a series of formal portraits of Urban VIII, but the first bust to achieve the quality of the "Montoya" and "Bellarmino" is a portrait of Bernini's great patron, Cardinal Scipione Borghese. The extraversion and immediacy of the Borghese groups done some ten years before are here transposed for the first time into serious portraiture. The Cardinal is shown in the act of speaking and moving—the action is caught at a moment that seems to reveal all the characteristic qualities of the subject through typical unconscious action. Bernini's most disarming portrait, however, is of his mistress, Costanza Buonarelli (*c.* 1635; Bargello, Florence), the wife of one of his assistants. It is a unique private record. According to his son, Bernini transformed Costanza into stone while he was enamoured of her flesh. Unlike any other bust of the 17th century, it anticipates the psychological precision of the great French master of 18th-century sculpture Jean-Antoine Houdon.

Public buildings. Under Urban VIII Bernini began to produce new and different kinds of monuments—tombs and fountains. The Tomb of Urban VIII shows the Pope seated with his arm raised in a commanding gesture. Below, flanking the gilt-bronze sarcophagus, are two white marble Virtues, Charity and Justice, which respectively were inspired by figures from the paintings of the Flemish artist Peter Paul Rubens (1577–1640) and of the Italian painter Guido Reni (1575–1642). Above the sarcophagus a figure of Death seems to write Urban's name on a leaf that serves as a marker for the tomb, an example of Bernini's use of the *concetto* ("idea," or the archaic "conceit"), an animating thematic idea unifying the diverse elements of the novel and richly coloured tomb. In these years Bernini also designed a revolutionary series of small tomb memorials, of which the most impressive is that of Maria Raggi (1643; Sta. Maria sopra Minerva, Rome).

Fountains

Bernini's fountains are his most obvious contribution to the city of Rome. His first, the "Barcaccia" in the Piazza di Spagna (1627–29), is analogous to the baldachin in its fusion of sculpture and architecture. The "Triton" is a dramatic transformation of a Roman architectonic fountain—the superposed basins of the traditional geometric piazza fountain seem to have come alive. Four dolphins raise a huge shell supporting the sea god, who blows water upward out of a conch.

LATER YEARS

Bernini's most spectacular public monuments date from the later 1640s and 1650s. The Fountain of the Four Rivers in Rome's Piazza Navona supports an ancient

Egyptian obelisk over a hollowed-out rock, surmounted by four marble figures symbolizing the four major rivers of the 17th-century world: the Danube of Europe, the Nile of Africa, the Ganges of Asia, and the Rio de la Plata of the Americas. The papal arms and the dove with an olive branch (the family symbol of Pope Innocent X) above the obelisk make the meaning clear—the restored and invigorated papacy rules the world under God. Despite its obvious symbolism, Bernini's fountain is his most spectacular permanent work. It is closely related to the temporary festival decorations and stage designs in which he also excelled.

The various works heretofore considered all show a dynamic transformation of familiar forms into new combinations. They are fused with meanings that are part of the forms and not superimposed upon them. In these dramatic metamorphoses Bernini revealed a synthetic imagination that is also present in his theatrical works, a few of which are preserved. The essence of Bernini's theatre was the same as that of his art. He broke through familiar conventions to establish an immediate and often surprising contact with the audience. He is known to have used real fire in his theatrical productions, and once he even flooded the stage, making the spectator—at least for a moment—a passionate actor in the spectacle he was watching.

The Cornaro Chapel. The greatest single example of Bernini's mature art is the Cornaro Chapel in Sta. Maria della Vittoria, in Rome, which completes the evolution begun early in his career. The chapel, commissioned by Cardinal Federigo Cornaro, is in a shallow transept in the small church. Its focal point is "The Ecstasy of St. Teresa," a depiction of a mystical experience of the great Spanish Carmelite reformer, Teresa of Ávila (1515–82). In representing Teresa's vision, during which an angel pierced her heart with a fiery arrow of divine love, Bernini followed Teresa's own description of the event. The sculptured group shows the transported saint swooning in the void, covered by cascading drapery. The angel, with clinging, flamelike drapery, stands erect, the arrow poised in his hand, smiling down on the unseeing figure whose ecstasy is interpreted with such disconcerting humanity. The sculptured group is revealed in celestial light within a niche over the altar, where the architectural and decorative elements are richly joined and articulated. The same heavenly light seems to flood the vault, where angels painted on stucco clouds adore the Holy Spirit. At left and right, in spaces resembling opera boxes, numerous members of the Cornaro family are found in spirited postures of conversation, reading, or prayer. The visitor standing before the chapel sees Teresa's ecstasy as if by revelation and looks up to see a heavenly glory.

The three-dimensional picture

The Cornaro Chapel carries Bernini's ideal of a three-dimensional picture to its apex. The figures of St. Teresa and the angel are sculptured in white marble, but the viewer cannot tell whether they are in the round or merely in high relief. The natural daylight that falls on the figures from a hidden source above and behind them is part of the group, as are the gilt rays behind. The whole "miraculous" vision is produced by a light the viewer does not understand. "The Ecstasy of St. Teresa" is not sculpture in the conventional sense. Instead, it is a framed pictorial scene made up of sculpture, painting, and light that also includes the worshipper in a religious drama that is not so much acted as revealed.

The Cornaro Chapel is the crowning synthesis of Bernini's earlier development and in some respects represents the highest achievement of his sculptural art, summing up the revolutionary experiments of the Borghese period and drawing upon the dynamic achievements of the Barberini papacy. In later years, the growing desire to control the environments of his statuary that culminated in the Cornaro Chapel led Bernini to concentrate more and more on architecture.

The principles of hallucinatory religious experience that Bernini first synthesized in the Cornaro Chapel soon found expression in three churches: one at Castel Gandolfo (1658–61; S. Tomaso da Villanova), a second at Ariccia (1662–64), and the third and most richy decorated at Rome (1658–70; S. Andrea al Quirinale). Like the Cornaro Chapel, these churches form architectural frameworks that are a logical extension of Bernini's sculptural ideal, which was always a single view of a climactic religious moment with active communication between spectator and statue. These principles predisposed Bernini to pay extraordinary attention to the environment of his works.

Oval and circular architectural plans

The visitor approaching S. Andrea from the street is attracted by an exedra, a concave wall, that leads to segmental circular steps, and a high, narrow facade that seems to be only a monumental entrance or portal. Once through the door he is immediately confronted by the high altar, since the church is oval in plan with the main axis on the short diameter of the oval. Within the altar space screened from the main body of the church is a richly coloured painting of the martyrdom of St. Andrew, which is seemingly carried by angels and lighted from an invisible source. The viewer's gaze rises to the pediment of the framing altar portal to see a marble St. Andrew floating on a cloud. He appears to have risen from the altar chapel, flying into the "dome of Heaven" above. To reinforce this illusion the lower part of the church is relatively dark. The dome, richly gilt and stuccoed with symbols of St. Andrew, is lighted by windows and a lantern decorated with the dove representing the Holy Spirit.

The interior of the small church of Santa Maria dell' Assunzione at Ariccia is organized in an analogous way. Flanked by palaces, its exterior is Bernini's version of the ancient Roman circular temple now known as the Pantheon (*c.* AD 118–128; Rome). All of his churches show his freedom of choice among traditional geometric forms. Each is given an individual exterior articulation that sets it off from traditional Renaissance structures of the 15th and 16th centuries as a personal and unique expression.

The colonnade at St. Peter's. Bernini's greatest architectural achievement is the colonnade enclosing the piazza before St. Peter's. The chief function of the large space was to hold the crowd that gathered for the papal benediction on Easter and other special occasions. Bernini planned a huge oval attached to the church by a trapezoidal forecourt—forms that he compared to the encircling arms of the mother church. The freestanding colonnades were a novel solution to the need for a penetrable enclosure. The piazza guides the visitor toward the church and counterbalances the overly wide facade of St. Peter's. Bernini's oval encloses a space centred on the Vatican obelisk, which had been moved before the church by Sixtus V in 1586. Bernini moved an older fountain by the architect Carlo Maderno (1555–1629) into the long axis of the piazza and built a twin on the other side to make a scenographic whole. The analogies to the oval plan of S. Andrea al Quirinale are fascinating, as are the differences in meaning and function.

Bernini also rebuilt the Scala Regia, the state staircase that leads from the portico of St. Peter's up to the apartments in the Vatican palace. He made the awkward and irregular site appear regular by optical illusions of great ingenuity. At the foot of the stair he placed a high-relief statue of Constantine, the first Christian emperor and founder of the church of St. Peter's. Constantine is shown at the moment of his conversion—a presentation typical of Bernini in the choice of a dramatic moment—and the statue serves to sublimate an awkward turn at the foot of the stair.

Bernini's architectural projects, however, were not invariably successful. In 1637, he began to erect campanili, or bell towers, over the facade of St. Peter's. But, in 1646, when their weight began to crack the building, they were pulled down and Bernini was temporarily disgraced.

Throne of St. Peter

Bernini's most spectacular religious decoration is the Throne of St. Peter, or the Cathedra Petri (1657–66), a gilt-bronze cover for the medieval wooden throne (cathedra) of the pope. Bernini's task was not only to make a decorative cover for the chair but also to create a meaningful goal in the apse of St. Peter's for the pilgrim's journey through the great church. The seat is seemingly

supported by four imposing bronze figures representing theological doctors of the early church: SS. Ambrose, Athanasius, John Chrysostom, and Augustine. Above, a golden glory of angels on clouds and rays of light emanates from the Dove of the Holy Spirit, which is painted on an oval window. The natural light on which the Dove is borne becomes the visible symbol of the stream of God's grace flowing over the world through the agency of his church.

As in other papal monuments of this period, Bernini's design for the Throne of St. Peter stressed the specifically Counter-Reformatory message of papal legitimacy. At the head of the Roman Catholic Church was St. Peter, buried under the baldachin and symbolically present on the empty throne in the apse. The divine will was done by the church's great preachers, represented by the four doctors who stand for the hierarchy of the church. Seen from the nave, the cathedra is framed by the baldachin over Peter's tomb, and around the base of the dome above is inscribed in Latin, "You are Peter and upon this rock I shall build my church." The foundation of the papacy is symbolized by a relief on the back of the throne representing Christ's command, "Feed my sheep." The cathedra was produced at about the same time as the piazza, and the contrast between these two works shows Bernini's amazing versatility. Both works were done for the Chigi pope, Alexander VII (1655–67), who was one of Bernini's greatest patrons.

Bernini's Tomb of Alexander VII, although largely executed by his pupils, furnishes another example of his ability to overcome a difficult site. The tomb had to be placed in a niche over a door, which gave Bernini the idea that enlivens the whole composition. The door appears to be the entrance to a tomb chamber, above which Bernini placed the praying figure of the Pope, surrounded by the four Virtues—Charity, Prudence, Justice, and Truth. Out from the door below flies a skeletal figure of Death holding an hourglass.

In addition to these large works, Bernini continued to produce a few portrait busts. The first of these, of Francesco I d'Este, duke of Modena (1650–51; Galleria e Museo Estense, Modena), culminates his revolution in portraiture. Much of the freedom and spontaneity of the bust of Cardinal Scipione Borghese is kept, but it is united with a heroic pomp and grandiose movement that portray the ideals of the Baroque age as much as the man. Some of these same features are present in his great bust of Louis XIV, which is the only relic of Bernini's visit to France in the summer of 1665. The "Louis" is more linear, vertical, and stable than the massively flaring "Este." In contrast to the "Este," "Louis" gazes out with godlike authority. The image set a standard for royal portraits that lasted 100 years.

Bust of Louis XIV

Trip to France. Bernini's real reason for journeying to France was to build a new royal residence. By this time, he was so famous that crowds lined the streets of each city along the route to watch him pass. His initial reception in Paris was equally triumphant, but he soon offended his sensitive hosts by imperiously praising the art and architecture of Italy at the expense of that of France. He described the now-destroyed Tuileries Palace, for example, as "a big little thing" reminiscent of "a great squadron of tiny children" and remarked that one painting by the Italian artist Guido Reni was worth more than all of Paris. Such statements made him unpopular at the French court and were to some degree responsible for the rejection of his designs for the Louvre.

When Bernini arrived in Paris, he had already submitted two projects for the Louvre. The first had a bold concave frontispiece, or entry facade, with a convex centre section. The facade was united vertically by a colossal order of pilasters rising several stories and framed by projecting wings. The second project was merely a modification of the first. Such spatial movement, found in his facade of S. Andrea al Quirinale, had never been applied to a city palace in Rome. His final Louvre design of 1664–65 (never built) abandoned curves in favour of a more typical Roman Baroque approach, in which Bernini expanded his designs for an actual Roman palace facade, the Chigi (Odescalchi) in Piazza SS. Apostoli of Rome, begun in 1664. The Chigi facade (later doubled in length, ruining its proportions) was a masterpiece of mounting rhythmic forms, of balance and climax, the end product of over 150 years of Roman facade design. Nowhere else does Bernini so clearly exhibit his community of thought with the Renaissance, and no other work shows how completely he could resolve the abstract problems of architectural design implicit in the work of his predecessors.

Late sculpture. Bernini's late works in sculpture are inevitably overshadowed by his grandiose projects for St. Peter's, but a few of them are of outstanding interest. For the Chigi Chapel in Sta. Maria del Popolo in Rome he carved two groups, "Daniel in the Lions' Den" and "Habakkuk and the Angel" (1655–61). These works show the beginnings of his late style: elongation of the body, expressive gesture, and simplified yet emphatic emotional expression. The same characteristics are already found in the figures supporting the Throne of St. Peter and culminate in the moving "Angels" for the Ponte Sant'Angelo in Rome, which Bernini redecorated with the help of assistants between 1667 and 1671. Pope Clement IX (1667–69) so prized the "Angels" Bernini carved that they were never set up on the bridge and are now in the Roman Church of S. Andrea delle Fratte. They were replaced by copies, of which one, the "Angel with the Superscription INRI," is actually a variation by Bernini himself.

The redecorated Ponte Sant'Angelo leading across the Tiber forms an introduction to the Vatican, and Bernini's other works—the Piazza, Scala Regia, and, within St. Peter's, the baldachin and cathedra—form progressively more powerful expressions of papal power and implied legitimacy to support and inspire pilgrims. Bernini completed one more decoration in St. Peter's in his last years: the altar of the Cappella del SS. Sacramento (1673–74). The pliant, human adoration of the angels contrasts with the timeless architecture of the bronze tabernacle they flank and typifies Bernini's late style. In his last years he seems to have found the inexorable laws of architecture a consoling antithesis to the transitory human state.

Last works. Bernini's greatest late work is the simple Altieri Chapel in S. Francesco a Ripa (*c.* 1674) in Rome. The relatively deep space above the altar reveals a statue representing the death of the Blessed Ludovica Albertoni. Bernini consciously separated architecture, sculpture, and painting for different roles, reversing the process that culminated in the Cornaro Chapel. In that sense, the Altieri Chapel is more traditional, a variation on his church interiors of the preceding years. Instead of filling the arched opening, "Ludovica Albertoni" lies at the bottom of a large volume of space, illuminated in her death by a heavenly light, maintained in her faith by the devotional image painted above, watched over by the Holy Spirit. The play of light and shade on the drapery gathered over her vitals is received as pure emotion. Her hands weakly clutching her breast and her swollen throat and gasping mouth make explicit her painful death. But Bernini chose not to differentiate colouristically between skin, garment, and sheet. Instead, he understated individual colours and textures to achieve a more unified emotional impact.

Bernini's last portrait, that of Gabriele Fonseca, is imbued with the same emotion as the figure of Ludovica Albertoni but expresses it by totally different means. "Fonseca," like the early bust of "Bellarmino," gazes at the altar in devotion and prayer. Perhaps in no other work did Bernini so dramatically externalize a subjective state. The hands seem to clench the rosary with spastic intensity as the beseeching eyes and mouth form a silent prayer. The "Fonseca," like many of Bernini's late works, exploits an emotional vehemence and exaggeration of features that borders on caricature.

Bernini died on November 28, 1680, at the age of 81. He had served eight popes.

ASSESSMENT

Artistic theory

Bernini's theory of art was traditional. He believed in studying the art of Greco-Roman antiquity to counteract

the imperfections of nature, but he may have deliberately overstressed his own relationship to antique works. Although Bernini's sketches and clay models (*bozzetti*) often show a dependence on recognizable antique sources, the finished works are far removed in both form and meaning. His working practice is well known, thanks to preserved drawings and *bozzetti*.

Bernini had unparalleled gifts as an executant, but he was increasingly forced to rely on assistance from others as the number of his commissions grew. He was supremely successful in organizing his studio and planning his work so that sculptures produced by a team actually seem to be all of a piece. Bernini's style was carried on for at least two more generations in various parts of Europe, and Italian sculpture of the 18th century is predicated on his achievement. Through the Roman architects of the High Baroque style, Mattia de' Rossi (1637–95) and Carlo Fontana (1634–1714) and their schools, his influence on architecture extended as far as England. In Austria, J.B. Fischer von Erlach (1656–1723) carried on a restrained Berninesque style of architecture, and Bernini's scenographic *concetti* were transformed into dramatic *tableaux vivants*, or sculptural groupings carved to look like living pictures, executed by the brothers Cosmas (1686–1739) and Egid Quirin (1692–1750) Asam in Bavaria.

Bernini's death marked the end of Italy's artistic hegemony in Europe. He was the last of Italy's remarkable series of universal geniuses, and the Baroque style he helped to create was the last Italian style to become an international standard. Bernini's world was one of religious and political absolutism, and, in spite of his obvious intelligence, he did not question the worth of that world. Perhaps more than any other artist, Bernini was the personification of one dominant aspect of his age, and it was largely his enthusiastic embrace of the status quo that assured his immense success. When he died, he was widely considered not only Europe's greatest artist but also one of its greatest men.

MAJOR WORKS

SCULPTURE: (unless stated located in Rome)—"Aeneas, Anchises and Ascanius Fleeing Troy" (1619; Borghese Gallery); "Neptune and Triton" (*c.* 1620; Victoria and Albert Museum, London); "Pluto and Proserpina" (1621–22; Borghese Gallery); "Apollo and Daphne" (1622–24; Borghese Gallery); "David" (1623–24; Borghese Gallery); "Santa Bibiana" (1624–26; Church of Sta. Bibiana); "St. Longinus" (1629–38; St. Peter's); "Cardinal Scipione Borghese" (1632; Borghese Gallery); Tomb of Urban VIII (1628–47; St. Peter's); Triton Fountain (1642–43; Piazza Barberini); Fountain of the Four Rivers (1648–51; Piazza Navona); "The Ecstasy of St. Teresa" (1645–52; Cornaro Chapel, Sta. Maria della Vittoria); Cathedra Petri (1657–66; St. Peter's); "Constantine" (1654–70; Scala Regia, Vatican Palace); "Louis XIV" (1665; Versailles, France); "Angels" (1667–69; S. Andrea delle Fratte); "Gabriele Fonseca" (*c.* 1668–75; S. Lorenzo in Lucina); "Ludovica Albertoni" (*c.* 1674; S. Francesco a Ripa); Tomb of Alexander VII (1671–78; St. Peter's).

ARCHITECTURE: (unless stated located in Rome)—baldachin (1624–33; St. Peter's); S. Andrea al Quirinale (1658–70); piazza and colonnade, St. Peter's (1656–67); Scala Regia, Vatican (1663–66); unexecuted designs for the Louvre, Paris (1664–65).

BIBLIOGRAPHY. The earliest and most authoritative life of Bernini is FILIPPO BALDINUCCI, *Vita del cavaliere Gio. Lorenzo Bernino* (1682; Eng. trans. by C. ENGGASS, *The Life of Bernini*, 1966). A diary kept by PAUL FREART DE CHANTELOU gives a precious record of Bernini's activities and conversations during his epoch-making visit to France in 1665: *Journal de voyage du chev. Bernin en France*, ed. by L. LALANNE (1885). A short but revealing discussion of all aspects of Bernini's work is found in R. WITTKOWER, *Art and Architecture in Italy, 1600 to 1750*, 2nd rev. ed., pp. 96–129 (1965), which contains the only satisfactory survey of Bernini's architecture. The basic catalog of sculpture, with bibliography, is Wittkower's *Gian Lorenzo Bernini: The Sculptor of the Roman Baroque*, 2nd ed. (1966). A popular introduction to Bernini's work, life, and times is H. HIBBARD, *Bernini* (1965). All of these must be modified by the discovery of new early works, presented (with controversial datings) by I. LAVIN, "Five New Youthful Sculptures by Gianlorenzo Bernini and a Revised Chronology of His Early Works," *Art Bulletin* 50:223–248 (1968). I. LAVIN, *Bernini and the Crossing of St. Peter's* (1968), presents new and important information about the statues and the baldachin.

(H.Hi.)

Bernoulli Family

Eight members of the Bernoulli family in three generations during the 17th and 18th centuries were mature mathematicians, at least three achieving the highest success in mathematics. Driven from their Antwerp home during the Catholic persecution of the Huguenots in 1583, the ancestors of the mathematicians sought refuge first in Frankfurt and then in Basel. The most celebrated of the Bernoullis were the brothers Jakob I and Johann I, who with Daniel, eminent son of Johann I, pioneered the development of and championed the cause of the version of calculus that had been created by the German mathematician Gottfried von Leibniz. By quickly identifying the fundamental problems to which calculus was best adapted, such as particle and fluid dynamics, optics, and probability, they contributed substantially to the advance of the physical sciences in the late 17th century and in much of the 18th. Often aggressive and zealous, occasionally obstinate, but always successful and respected, the Bernoullis represented the emergence of science and mathematics as powerful forces in society.

By courtesy of the Offentliche Bibliothek der Universitat, Basel, Switzerland



(Left) Jakob I Bernoulli (1654–1705), oil painting by Nikolaus Bernoulli, 1687. (Right) Johann I Bernoulli (1667–1748), oil painting by Johann Jakob Meyer, 1720. In a private collection.

Born at Basel on December 27, 1654, Jakob I became interested in mathematics despite his father's opposition. His travels led to a wide correspondence with mathematicians. He accepted a professorial chair of mathematics at the University of Basel in 1687; and, following his mastery of the mathematical works of John Wallis, Isaac Barrow (both English), René Descartes (French), and Leibniz, who first drew his attention to calculus, he embarked upon original contributions.

Johann I, born on July 27, 1667, in Basel, also turned to mathematics despite his father's opposition. In 1691–92 he wrote two texts, not published until later, on differential and integral calculus. In 1692 he taught calculus to the mathematician Guillaume-François-Antoine de L'Hospital, who agreed to pay him for mathematical discoveries. From 1695 to 1705 he taught mathematics at Groningen, The Netherlands, and on the death of his brother Jacob I assumed a professorship at Basel.

Jakob I and Johann I

Jakob I in 1690 was the first to use the term integral in analyzing a curve of descent. His 1691 study of the catenary, or the curve formed by a cord suspended between its two extremities, was soon applied in the building of suspension bridges. In 1695 he also applied calculus to the design of bridges.

Johann I exceeded his brother in the number of contributions he made to mathematics. He applied calculus to the determination of lengths and areas of curves, such as the isochrone, along which a body will fall at constant speed, and the tautochrone, which was found to be important in clock construction. He also made contributions

to the theory of differential equations, the mathematics of ship sails, and optics. Johann I sent to L'Hospital in Paris a method or rule for solving problems involving limits that would apparently be expressed by the ratio of zero to zero, now called L'Hospital's rule on indeterminate forms because it was included in L'Hospital's influential textbook of 1696, *Analyse des infiniment petits* ("Analysis of the Infinitely Small").

The Bernoulli brothers often worked on the same problems, but not without friction. Their most bitter dispute concerned finding the equation for the path followed by a particle from one point to another in the shortest time, if the particle is acted upon by gravity alone, a problem originally discussed by Galileo. In 1697 Jakob I offered a reward for its solution. Accepting the challenge, Johann I proposed the cycloid, the path of a point on a moving wheel, pointing out at the same time the relation this curve bears to the path described by a ray of light passing through strata of variable density. A protracted, bitter dispute then arose when Jakob I challenged the solution and proposed his own. The dispute marked the origin of a new discipline, the calculus of variations.

Jakob I died on August 16, 1705, in Basel. His pioneering work *Ars Conjectandi* (1713, "The Art of Conjecturing") contained his theory of permutations and combinations; the so-called Bernoulli numbers, by which he derived the exponential series; his treatment of mathematical and moral predictability; and the subject of probability—containing what is now called the Bernoulli law of large numbers, basic to all modern sampling theory.

Ardent in his friendships and keen in his resentments, Johann I zealously defended the cause of Leibniz in the dispute with Newton over who had originated calculus. He died on January 1, 1748, in Basel. His text in integral calculus appeared in 1742 and his differential calculus shortly afterwards.

Born on January 29, 1700, in Groningen, Daniel was the second son of Johann I, who first taught him mathematics. After studying philosophy, logic, and medicine at the universities of Heidelberg, Strasbourg, and Basel, in 1721 he received an M.D. degree, and in 1723–24 he wrote *Exercitationes quaedam Mathematicae* on differential equations and the physics of flowing water, which won him a position at the influential Academy of Sciences in St. Petersburg. He lectured there until 1732 in medicine, mechanics, and physics; and he researched the properties of vibrating and rotating bodies and contributed to probability theory. That year he returned to Basel to accept the post in anatomy and botany. By then he was widely esteemed by scholars and admired by the public throughout Europe.

Daniel's work in hydrodynamics

Daniel's reputation was established in 1738 with *Hydrodynamica*, in which he considered the properties of basic importance in fluid flow, particularly pressure, density, and velocity, and set forth their fundamental relationship. He put forward what is called Bernoulli's principle, which states that the pressure in a fluid decreases as its velocity increases. He also established the basis for the kinetic theory of gases and heat by demonstrating that the impact of molecules on a surface would explain pressure and that, assuming the constant, random motion of molecules, pressure and motion increase with temperature. About 1738 his father published *Hydraulica;* this attempt by Johann I to obtain priority for himself was another instance of his antagonism toward his son.

Between 1725 and 1749 Daniel won ten prizes from the Paris Academy of Sciences for work on astronomy, gravity, tides, magnetism, ocean currents, and the behaviour of ships at sea. He also made substantial contributions in probability. He shared the 1735 prize for work on planetary orbits with his father, who, it is said, threw him out of the house for thus obtaining a prize he felt should be his alone. Daniel's prizewinning papers reflected his success on the research frontiers of science and his ability to set forth clearly before an interested public the scientific problems of the day. In 1732 he accepted a post in botany and anatomy at Basel; in 1743, one in physiology; and in 1750, one in physics. Widely esteemed by scholars and admired by the European public, he died on March 17, 1782, in Basel.

Nikolaus I, born October 10, 1687, was the nephew and pupil of Jakob I and Johann I. A law scholar and mathematician he contributed works on theory of probability and on infinite series. He died in Basel on November 29, 1759. Nikolaus II, the first son of Johann I, was born in Basel on November 4, 1695. In 1726 he was appointed professor of mathematics at St. Petersburg, where he died in a tragic drowning accident on October 8, 1726.

Johann II was born May 28, 1710, the third son of Johann I. He won four prizes from the Paris Academy, corresponded widely with European scientists, and in 1742 published his father's *Opera Omnia*. He died on July 17, 1790, in Basel. He had two sons: Johann III, born on November 4, 1744, who administered the Bernoulli estate and was appointed by Frederick II to the astronomical observatory in Berlin, where he died on July 13, 1807. Jakob II, born October 17, 1759, was accomplished in mathematics and physics. He died on August 15, 1789.

BIBLIOGRAPHY. The works of Johann I Bernoulli were published in his *Opera Omnia*, 4 vol. (1743), and of Jakob I in 2 vol. (1744). No comprehensive biography of any member of the Bernoulli family has been published to date. Additional information may be found in brief articles appearing in such books as ERIC T. BELL, *Men of Mathematics*, ch. 8 (1937, reprinted 1961); and TREVOR I. WILLIAMS (ed.), *A Biographical Dictionary of Scientists* (1969).

(Ed.)

Bernstein, Eduard

The German Social Democratic propagandist and political theorist Eduard Bernstein was one of the first Socialists to attempt to revise some of Karl Marx's tenets, such as the imminent collapse of capitalism and the seizure of power by the proletariat. Although not a distinguished theoretician, Bernstein, the "father of revisionism," envisaged a type of social democracy that combined private initiative with social reform.

Bernstein was born in Berlin on January 6, 1850, into a Jewish family that had come to the capital of Prussia from Danzig. His father was a railroad engineer, and his uncle Aaron Bernstein was the editor of the *Berliner Volks-Zeitung*, a newspaper widely read in progressive working-class circles. It was thus not surprising that at a young age Eduard shared the aspirations of many educated Germans for national unity and democracy. Of an engaging and candid disposition, he retained the goodwill of his superiors when, in 1872, as a young bank clerk he announced that he had joined the Social Democratic Party. The turbulent years after Prussia's 1871 defeat of France also contributed to the formation of his political beliefs. Yet the ever-genial Bernstein tended to be attracted more to Socialism of an undogmatic, pragmatic kind than to radical Marxism. He preferred the democratic and pacifist Social Democrats to the somewhat authoritarian Allgemeiner Deutscher Arbeiterverein (General German Workers' Association).

Early Socialist affiliations

In joining the party he became associated with the German Socialist organ, *Die Zukunft* ("The Future"). The economic crisis of 1873, which continued into the 1890s, reinforced his belief in the fragility of capitalism. It was, however, Chancellor Otto von Bismarck's anti-Socialist laws that finally impelled him toward a more radical position. Exiled from Germany, he emigrated to Switzerland, abandoning the "ethical Socialism" of Karl Höchberg, the wealthy patron of *Die Zukunft*. With Marx's consent, he became the editor of the Zürich edition of *Der Sozialdemokrat*, a periodical that was the rallying centre of the underground Socialist party. Expelled from Switzerland at the request of Bismarck in 1888, Bernstein continued the publication of the periodical in London. There he became a close friend of Friedrich Engels, Marx's collaborator and patron, and also came to know intimately the leaders of the influential Fabian Society, which advocated a gradualist development of Socialism. Bernstein set forth his revised views in a series of articles and in a letter to the Social Democratic Party meeting at



Eduard Bernstein, *c.* 1918.
Archiv fur Kunst und Geschichte

Stuttgart in 1898. In the following year he published *Die Voraussetzungen des Sozialismus und die Aufgaben der Sozialdemokratie* (Eng. trans. *Evolutionary Socialism*, 1909).

When Bernstein returned to Germany in 1901, he became the theoretician of the growing revisionist school of the reformist labour movement. He held that Socialism is the final result of the liberalism that is an inherent part of human aspiration and not the mere product of a revolt against the capitalist middle class. He no longer believed in the imminent collapse of capitalism, nor did he any longer regard the bourgeoisie as exclusively parasitic and oppressive. In addition, he believed that the concentration of productive industry was not taking place in all fields as thoroughly or as fast as Marx had predicted. Citing such reforms as factory legislation and the freeing of labour unions from legal restrictions, he pointed out that, under pressure from the Socialist movement, a reaction had set in against the exploitive inclinations of capital. Thus, he argued, the prospects for lasting success lay in a steady advance rather than in a violent upheaval.

In 1902 Bernstein was elected a member of the Reichstag, or Parliament, to which he was re-elected several times. He remained a member of the Reichstag up to 1928. Eventually revisionism became Social Democratic ideology, while the dogmatic Marxism of the Socialist theoretician Karl Kautsky and the eclectic Marxism of the German labour leader August Bebel faded into the background. Bernstein, however, who was opposed to violence between nations as well as between classes, lent his voice to that of the left to fight against militarism. During World War I, although a leading member of his party's right wing, he sided with the Independent Socialists (USPD) to protest his party's support of the war. As soon as peace was restored, however, he returned to his old party and opposed those who wanted to transform the political revolution of November 1918 into a social revolution. He believed that the establishment of the parliamentary republic opened the way to uninterrupted progress, and after the war he served as secretary of state for economy and finance in 1919.

Opposition to World War I

Social Democracy had finally become the great popular and reformist movement he had desired for more than 20 years, and, as an adviser respected by his party, he inspired much of its program. If he helped to discourage the Germans from following the Russian example of 1917, he could not dissuade them from imitating the Italian Fascist model of 1922. He regarded the bloody outrages of the Nazis and their predecessors as the thoughtless actions of unbalanced minds; he was unable to comprehend the nature of National Socialism and remained powerless to prevent its seizure of power. Less than six weeks after his death in Berlin on December 18, 1932, the democratic state on which he had set all his hopes was to give way to the dictatorship of Adolf Hitler.

BIBLIOGRAPHY. EDUARD BERNSTEIN, "Entwicklungsgang eines Sozialisten," in FELIX MEINER (ed.), *Die Volkswirtschaftslehre der Geggenwart in Selbstdarstellungen*, vol. 1 (1924), assists the reader in understanding Bernstein's development as a socialist; it is complemented by his *Sozialdemokratische Lehrjahre* (1928). PIERRE ANGEL, *Eduard Bernstein et l'évolution du socialisme allemand* (1961), is an intellectual biography that attempts to explain the evolution of the socialist movement in the Germany of Bismarck and Wilhelm II. Also interesting is PETER GAY, *The Dilemma of Democratic Socialism: Eduard Bernstein's Challenge to Marx* (1952), which emphasizes Bernstein's anti-Marxist feelings.

(P.R.A.)

Bessel, Friedrich Wilhelm

Friedrich Wilhelm Bessel was a German astronomer whose works laid the foundations for a better determination than any previous method had allowed of the scale of the universe and the sizes of stars, galaxies, and clusters of galaxies. In addition, he made fundamental contributions to accurate positional astronomy, the exact measurement of the positions of celestial bodies; to celestial mechanics, dealing with their movements; and to geodesy, the study of the Earth's size and shape; he enlarged the resources of pure mathematics by his introduction and investigation of what are now known as Bessel functions, which he used first in 1817 to study the problem of determining the motion of three bodies moving under mutual gravitation and, seven years later, he developed Bessel functions more fully for the treatment of planetary perturbations. Much credit for the final establishment of a scale for the universe in terms of solar system and terrestrial distances, which depends vitally on accurate measurement of the distances of the nearest stars from the Earth, must go to Bessel.



Bessel, engraving by E. Mandel after a painting by Franz Wolf (1795–1859).
The Bettmann Archive

In geodesy, Bessel's contributions include a correction in 1826 to the seconds pendulum, the length of which is precisely calculated so that it requires exactly one second for a swing. During 1831–32 he directed geodetical measurements of meridian arcs in East Prussia, and in 1841 he deduced a value of $\frac{1}{299}$ for the ellipticity of the Earth, the amount of elliptical distortion by which the Earth's shape departs from a perfect sphere. He was the first to make effective use of the heliometer, an instrument de-

Contributions to geodesy

signed for measuring the apparent diameter of the Sun. He introduced corrected observations for the so-called personal equation, a statistical bias in measurement characteristic of the observer himself that must be eliminated before results can be considered reliable, and he made a systematic study of the causes of instrumental errors. His own corrected observations were more accurate than any previous ones, and his methods offered the way to great advances in the study of the stars.

Bessel early recognized where his potential genius lay; and he succeeded in realizing it in a profession entirely different from the one that he originally followed. Heinrich Wilhelm Matthäus Olbers (1758–1840), the German astronomer who calculated numerous comet orbits and took a leading part in the discovery of the minor planets, later said that the greatest service he had rendered astronomy was that he recognized and furthered Bessel's genius.

Bessel was born in Minden, Westphalia (now in West Germany), on July 22, 1784, the son of a poor government employee. At the age of 15, he entered an export–import firm. During his apprenticeship, dreaming of travel, he studied languages, geography, the habits of distant peoples, and the principles of navigation, which led him to astronomy and mathematics. Working at night, in 1804 he wrote a paper on Halley's Comet in which he calculated the orbit from observations made in 1607. He sent it to Olbers, who was so impressed that he arranged its publication in *Monatliche Correspondenz* (x, 1804) and proposed Bessel as assistant at the Lilienthal observatory of the celebrated early lunar observer J.H. Schröter. Bessel, who was liked and appreciated by his commercial firm, had to choose between a position of relative affluence if he remained in it and poverty and the stars if he left it. He decided for the latter. After only four years at Lilienthal, the Prussian government charged him with the construction at Königsberg (now Kaliningrad) of the first big German observatory. In 1810 he was appointed professor of astronomy at Königsberg, where he worked assiduously on the reconstruction of the whole science of astronomical observations, directing the observatory from the date of its completion in 1813 until his death, March 17, 1846.

Measurements of star positions

The later achievements of Bessel were possible only because he first established the real framework of the universe by making accurate measurements of the positions and motions of the nearest stars, making corrections for various measuring errors caused by imperfections in his telescopes and by disturbances in the atmosphere. Having established exact positions for about 50,000 stars, he was ready to observe exceedingly small but highly significant motions, relative to one another, among them. Choosing 61 Cygni, a star barely visible to the naked eye and known to possess a relatively high velocity in the plane of the sky, Bessel showed that after correcting for this, the star apparently moved in an ellipse every year. This back and forth motion, called parallax, he said could only be interpreted as being caused by the motion of the Earth around the Sun. Calculation indicated a distance from Earth to 61 Cygni of 10.3 light-years. (The nearest star known is Alpha Centauri, 4.3 light-years away.) Olbers, presented with these conclusions on his 80th birthday, thanked Bessel and said the gift "put our ideas about the universe for the first time on a sound basis." Bessel was honoured for this achievement by the Royal Astronomical Society of London and others.

A major discovery by Bessel was that the two bright stars Sirius and Procyon execute minute motions that could be explained only by assuming that they had invisible companions disturbing their motions. The existence of such bodies, now named Sirius B and Procyon B, was confirmed with more powerful telescopes after Bessel's death. An important share in the discovery of the planet Neptune also belongs to Bessel. In a paper read in 1840, he called attention to exceedingly small irregularities in the orbit of Uranus, which he had observed and concluded were caused by an unknown planet beyond.

He published a number of major articles and introduced a multivolume series, *Astronomische Beobachtungen auf der K. Sternwarte zu Königsberg* (1815–1844); his minor papers number more than 350.

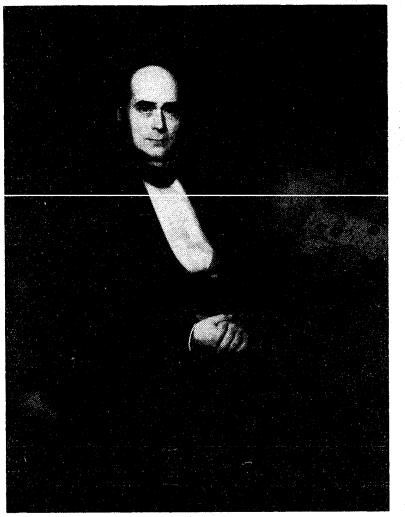
BIBLIOGRAPHY. Additional biographical information may be found in works on astronomy, such as H.S. WILLIAMS, *The Great Astronomers* (1930); A.M. CLERKE, *A Popular History of Astronomy during the Nineteenth Century*, 4th ed. rev. (1902); and F. ZWICKY, *Morphological Astronomy* (1957).

(Ed.)

Bessemer, Sir Henry

English inventor and engineer, Sir Henry Bessemer discovered and developed the first process for manufacturing inexpensive steel, the basic construction material of industry and technology since the mid-19th century.

By courtesy of The Iron and Steel Institute, London; photograph, The Science Museum, London



Bessemer, oil painting by Rudolf Lehmann (1819–1905). In the Iron and Steel Institute, London.

Early inventions

He was born in Charlton in Hertfordshire on January 19, 1813, the son of an engineer and typefounder. He early showed considerable mechanical skill and inventive powers. After the invention of movable stamps for dating deeds and other government documents, and the improvement of a typesetting machine, he went to the manufacture of "gold" powder from brass for use in paints. The florid decoration of the time demanded great quantities of such material, and Bessemer's secret process soon brought him great wealth.

He developed other inventions, notably sugarcane-crushing machinery of advanced design, but he was soon devoted to metallurgy. In his time there were but two iron-based construction materials: cast iron made by the treatment of iron ore with coke in the blast furnace and wrought iron made from cast iron in primitive furnaces by the laborious manual process of "puddling" (stirring the melted iron to remove carbon and raking off the slag). Cast iron was, and still is, excellent for load-bearing purposes such as columns or bridge-piers and for engine parts, but for girders and other spans, and particularly for rails, only wrought iron was suitable. Puddling removed carbon, which makes cast iron brittle, and produced a material that could be rolled or forged, but only in "blooms," or large lumps of 100–200 pounds, and that was full of slag. The blooms had to be laboriously forged together by steam hammers before they could be rolled to any useful length or shape. The only material known as steel was made by adding carbon to pure forms of wrought iron, also by slow and discontinuous methods; the material was hard, would take an edge, and was used almost entirely for cutting tools.

During the Crimean War, Bessemer invented an elongated artillery shell that was caused to rotate by the powder gases. The French authorities with whom he was negotiating, however, pointed out that their cast-iron cannon would not be strong enough for this kind of shell. He

Bessemer process

thereupon attempted to produce a stronger cast iron. In his experiments he discovered that the excess oxygen in the hot gases of his furnace appeared to have removed the carbon from the iron pigs that were being preheated —much as the carbon is removed in a puddling furnace —leaving a skin of pure iron. Bessemer then found that blowing air through melted cast iron not only purified the iron but also heated it further, allowing the purified iron to be easily poured. This heating effect is caused by the combustion of the carbon (and some silicon) in the iron. Utilizing his new techniques, he was soon able to produce large, slag-free ingots as workable as any wrought-iron bloom, and far larger; he invented the tilting converter into which molten pig iron could be poured before air was blown in from below. Eventually, with the aid of an iron–manganese alloy, which was developed at that time by Robert Forester Mushet, Bessemer also found how to remove excess oxygen from the decarburized iron.

His announcement of the process in 1856 before the British Association for the Advancement of Science in Cheltenham, Gloucestershire, brought many ironmasters to his door, and many licenses were granted. Very soon, however, it became clear that two elements harmful to iron, phosphorus and sulfur, were not removed by the process—or at least not by the fireclay lining of Bessemer's converter. It was not until 1878 that the British metallurgist Sidney Gilchrist Thomas developed a lining that removed phosphorus and made possible the use of phosphoric ores of the Continent.

Bessemer had, unknown to himself, been using phosphorus-free iron, but the ironmasters were not so lucky. Their iron was perfectly satisfactory for the puddling process, in which phosphorus is removed because the temperatures are lower, but it could not be used in the Bessemer process. Bessemer was forced to call in his licenses and find a phosphorus-free source of iron in northwestern England; thus he was able to enter the steel market on his own. Once the phosphorus problem was recognized and solved, he became a licensor once again, and vast profits flowed in. It became clear that "mild steel"—as it was known to distinguish it from the hard tool steels—could more clearly and reliably be used in place of wrought iron for ship plate, girders, sheet, rods, wire, rivets, and other items. The invention of the open-hearth (Siemens-Martin) process in the late 1860s eventually outstripped that of the Bessemer process. This has now yielded place, in great measure, to oxygen steelmaking, a development of the Bessemer process.

Later inventions

In his later years—the process had not become a clear success until he was nearing 70—Bessemer continued to invent and make discoveries. The solar furnace he built was more than a successful toy; he designed and built an astronomical telescope for his own amusement; he developed a set of machines for polishing diamonds that helped to re-establish that trade in London; his lifelong susceptibility to seasickness led him to develop a stabilized cabin for steamships, but it was not wholly successful.

Apart from his knighthood, he received many honours, such as the Fellowship of the Royal Society, the Freedom of the City of London, and the presidency of the Iron and Steel Institute, as well as many foreign distinctions. He died in London on March 15, 1898.

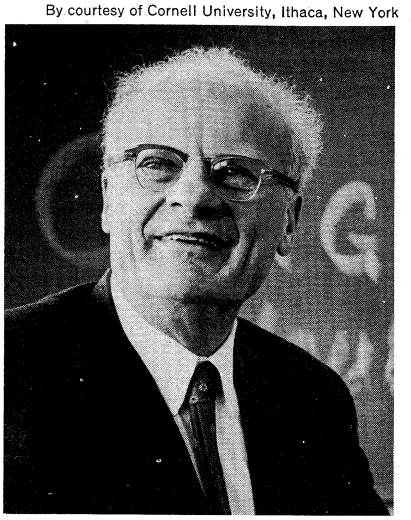
BIBLIOGRAPHY. SIR HENRY BESSEMER, *An Autobiography*, with a concluding chapter by his son, HENRY BESSEMER (1905), the only comprehensive work, and the source of most material written since; "The Manufacture of Iron Without Fuel," *The (London) Times*, p. 10, col. 5–6 (August 14, 1856), the complete address to the British Association at Cheltenham, reprinted from Bessemer's original text.

(J.P.S.)

Bethe, Hans Albrecht

Hans Albrecht Bethe belongs to a group of physicists who among them shaped classical physics into quantum physics. Through pioneering research and influential teaching, he increased substantially the understanding of the atomic processes responsible for the properties of matter and of the nuclear forces governing the structures of atomic nuclei. Moreover, he was a leader in emphasizing the social responsibility of science.

By courtesy of Cornell University, Ithaca, New York



Bethe.

Bethe was born on July 2, 1906, in Strassburg, Germany. He studied physics at the University of Frankfurt and did research in theoretical physics at the University of Munich, where he obtained the doctorate in 1928. His doctoral thesis, on the theory of electron diffraction, remains of fundamental value in understanding observational data. His work on term splitting in crystals in 1929 showed how the symmetrical electric field by which an atom in a crystal is surrounded affects its energy states. In 1931 he worked with Enrico Fermi in Rome. He returned to Germany and served as a lecturer at the University of Tübingen until 1933. After a stay in Manchester he emigrated to the United States and became, in 1934, a lecturer at Cornell University in Ithaca, New York, which remained his home in spite of many offers from other universities. He married Rose Ewald in 1939; they had a son and daughter.

In 1939 Bethe calculated the Sun's energy production. This results from the fusion of four hydrogen atoms (each of mass 1.008) into one helium atom (mass 4.0039). No direct fusion is possible, but Bethe showed that the probabilities of the four steps of the "carbon cycle" can account for the energy output. A carbon isotope of mass 12 reacts successively with three hydrogen nuclei (protons) to form the nitrogen isotope of mass 15; energy is produced through the fusion of a fourth hydrogen nucleus to release a helium nucleus (alpha particle) and the original carbon isotope.

Work in weaponry

At the beginning of World War II, Bethe had no U.S. clearance for military work. But, after reading in the *Encyclopædia Britannica* that the armour-piercing mechanism of grenades was not well understood, he formulated a theory that became the foundation for research on the problem. His work, unpublished except in classified reports, illustrated his faculty for developing highly mathematical theories to the point that their numerical results could be compared with the actual measurements.

After working at the Massachusetts Institute of Technology on the development of radar, Bethe headed the group of theoreticians for the atomic bomb Manhattan Project in Los Alamos, New Mexico. His knowledge of physics and his friendly and quiet personality made him particularly well suited for this demanding job.

The development of the atomic bomb and the dropping of it on Hiroshima and Nagasaki created a strong feeling of social responsibility in Bethe and other Los Alamos physicists. He was one of the organizers and original contributors to the *Bulletin of the Atomic Scientists* (since renamed *Science and Public Affairs*). Moreover, he lectured and wrote on the nuclear threat in order to increase public awareness of it.

Honours

Bethe was awarded the Presidential Medal of Merit in 1946. He was president of the American Physical Society in 1954; he received the Max Planck Medal in 1955 and the Fermi Award in 1961. In 1967 he was awarded the Nobel Prize for his discoveries concerning the energy production of stars. He became, in 1957, a foreign member of the Royal Society of London, as well as a member of the National Academy of Sciences in Washington, D.C.

The discovery of neutron stars led Bethe back to fundamental research in astrophysics in 1970. These stars are held by gravity at such high density that protons, by fusion with electrons, turn into neutrons, which constitute nearly the entire matter. Although his main interest was in the rapidly developing subjects of atomic and nuclear processes, he also applied classical mathematical methods to the calculation of electron densities in crystals, the order–disorder states in alloys, the operational conditions of reactors, the ionization processes in shock waves, and the detection of underground explosions from seismographic records.

BIBLIOGRAPHY. Bethe's later works include *Elementary Nuclear Theory* (1948), a discussion of the experimental evidence concerning the forces acting inside the atomic nucleus, and *Intermediate Quantum Mechanics*, 2nd ed. (1968), a theoretical description of atomic structure. See R.E. MARSHAK (ed.), *Perspectives in Modern Physics* (1964), for further bibliography.

(P.P.E.)

Betulales

The birches and their close allies—the alders, the ironwoods, the hornbeams, and the hazelnuts—are a group of trees and shrubs that comprise an order of flowering plants, the Betulales or birch order. The group is predominantly of the North Temperate and Arctic regions and long ago became deeply embedded in the folklore and ancient forest industries of primitive man. The tree birches are today major sources of cabinet woods and also provide handsome and graceful material for the landscape gardener. Ecologically, the alders especially have been found to be important soil-builders ("pioneers") because of the nitrogen-fixing bacteria in their root nodules; and all of the shrubby members of the order provide important cover for wildlife. Scientifically, as one of the members of the catkin-bearing group of flowering plants (Amentiferae), they have shared the centre of the stage along with the oaks, the walnuts, the willows, and others in stormy debates concerning which of the flowering plants are in reality primitive.

GENERAL FEATURES

Size range and distribution. The birches range from trees to sprawling shrubs; the canoe or paper birch exceptionally reaches a height of 37 metres (120 feet) and a diameter of one metre (three feet) on the northwest coasts of North America. The North American yellow birch and black birch, *Betula maximowicziana* of Japan, and *Betula albo-sinensis* of central and western China achieve heights of 27 to 30 metres (90 to 100 feet). Most of the remaining tree birches are modest in height, such as the gray or old-field birch of the eastern United States, whose clustered stems seldom become much over nine metres (30 feet) high. Unusual flexibility in adapting to its environment is shown by the largest of the birches, the canoe birch, which assumes a depressed and shrubby form when it grows in the exposed habitats of hilltops at its northernmost limits near the tree line, or occurs at high altitudes. Among the dwarf birches, *Betula pumila* may become as much as five metres (15 feet) high; other species that extend into the Arctic achieve a height of one and a half or two metres (five or six feet) when in the lee of a protective rock, but when in the open, they creep on the surface of the tundra. One of these dwarf birches, *Betula michauxii*, is confined to southern Labrador; its slender twigs rise but 30 centimetres (about one foot) or so out of the shallow ponds in which it grows.

Tree heights

The alders are commonly shrubs or small trees of medium size, but the black alder of Europe and northern Africa (*Alnus glutinosa*), the tree alders of western North America (*Alnus rubra* and *A. rhombifolia*), and *Alnus hirsuta* of northeastern Asia and northern Japan, are trees up to 15 to 23 metres (50 to 75 feet) in height and 0.6 metre (two feet) in diameter.

The various species of hornbeams and hop-hornbeams are small trees, some of which become 15 to 18 metres (50 to 60 feet) tall but commonly reach only a height of three to five metres (10 to 15 feet). The very local genus *Ostryopsis*, with two species in China, is a small shrub up to three metres (10 feet) in height.

Among the hazels, a strictly tree-habit characterizes the Turkish hazel, which may reach a height of 24 metres (80 feet), while its close relative, the Chinese hazel of central and western China, is reported to grow over 37 metres (120 feet) high. The great majority of hazels, however, are modest shrubs, often forming coppices.

The members of this order are in cool to cool-temperate areas of the Northern Hemisphere; only in the Americas do they extend along the mountains into the Southern Hemisphere. In the mountains of North America some species occur as endemics. Individual species are usually prominent members of the plant communities of which they form a part, but seldom occur in pure stands.

Geographic distributions

Most of the genera, with the exception of *Ostryopsis*, are widespread in their distribution today. Thus in *Betula* (birch), the white birches and dwarf birches are circumpolar and very nearly coincide in their areas, the dwarf birches extending farthest into the Arctic. They now occupy most areas that were glaciated recently (until about 10,000 years ago). By contrast, other members of the genus are more disrupted in their distribution, tending to be centred on the geologically ancient areas of eastern North America and of eastern Asia, with extensions into the Himalayan and Caucasus regions. Some are even more limited in distribution, occurring only in northern China, Korea, and Japan.

In the genus *Alnus* (alders), many species occur generally throughout the range of the Betulales, except for the Himalayan region, where a small and distinctive group of species occurs. A few other species occur in Japan, across cool temperate North America, and in Europe. One small group of closely related species is limited to Szechwan in southwestern China.

Ostrya (hop-hornbeam) and a few species of the genus *Carpinus* (hornbeam or blue beech) share much the same areas, both groups occurring in central eastern North America, in the southern European region exclusive of the Iberian Peninsula, and in the southeastern portion of Asia. One small group of species of *Carpinus* is limited to temperate east Asia. *Carpinus*, however, has a somewhat wider range than does *Ostrya*, not only in southern Europe and southeastern Asia, but also in the highlands of Central America, from which *Ostrya* is absent. The two genera are notably absent from recently glaciated regions except for limited incursions into the southern fringes of this area.

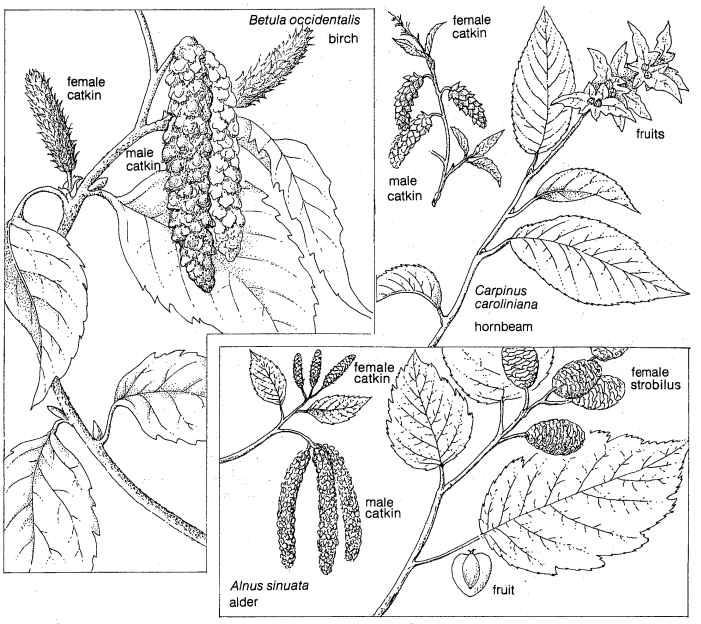
The closely related genus *Ostryopsis* is restricted to portions of the Himalayan region.

Corylus is another genus whose species occur around the Northern Hemisphere, projecting well into formerly glaciated areas, although not as aggressively as do the members of the dwarf birches of the genus *Betula*.

Economic importance. The primary economic members of the family are the tree birches of the Northern Hemisphere, whose fine-grained wood provides lumber used for flooring, trim, furniture, and high quality veneers. The uniform texture of the wood of these birches and of the tree alders makes it excellent for buttons, spools, wooden shoes, shoe lasts, and similar objects. The wood of hop-hornbeam (*Ostrya*) and hornbeam (*Carpinus*) is so difficult to split that it was used in pioneer days for such items as tool handles and levers. All genera are used for fuel and charcoal.

Products and uses of bark

The bark of various genera and also the mature female strobili (seed-bearing structures) of some species of alder are sources of high-quality tanning materials, and have been so used by Eskimos, American Indians, and others. The papery bark of the white birches was former-



Representative plants from two of the three tribes of the birch order.

From (*Betula occidentalis*) *Taxonomy of Flowering Plants*, 2nd ed., by C.L. Porter, W.H. Freeman and Company, Copyright © 1967, and (others) C. Sargent, *Manual of the Trees of North America* (1922), Dover Publications, Inc.

ly used as a substitute for writing paper in Kashmir, and for many kinds of utensils and containers by the American Indians and Ainu people of Japan. In the Lapland region of Norway and Finland and other places it is used, when finely ground, as an emergency food. Some of the birches produce a sugary sap, which is fermented to produce birch beer; the twigs of the cherry birch of North America are steam-distilled to produce an essential oil (oil of wintergreen) mostly composed of methyl salicylate. Leaves of some of the birches in combination with alum produce a green dye, or with chalk produce a yellow colour. There is a multitude of other minor and very local uses, ranging from the hooked sticks derived from *Alnus nitida* (Himalayan black alder) used in primitive rope bridges to extracts of the bark of *Alnus rubra* (of the west coast of North America), which have been recommended for scrofulous conditions, indigestion, and dyspepsia.

Hazel nuts and filberts (*Corylus*) have been harvested for food since the days of primitive man. Hybrids and selections have been produced by horticulturists and are widely used for food and as a source of oils used in food, painting, perfumes, and soap manufacture.

NATURAL HISTORY

Life cycle. The life cycle of these plants is basically like that of other flowering plants, but there are certain peculiarities, largely in the reproductive portion. All members of the family have developed the abundant, small (approximately 0.03 millimetre diameter), nonsticky, unornamented pollen grains that are well-adapted to wind pollination. It has been calculated that one hazel bush produces 600,000,000 pollen grains, shed from the flexible male catkins (pendant clusters of male flowers) before or at the time of the opening of the leaves in spring. The stiff female inflorescences (flower clusters) contain many fewer pistils (female structures), one hazel bush producing only about 1,300 in 200 female inflorescences. Only the styles and stigmatic surfaces of the pistils (the pistil consists of stigma, style, and ovary—the stigma and styles being adapted to receiving the pollen, the ovary enclosing the ovules, which are immature seeds) are mature at pollination time, usually late in May. In most flowering plants the three parts of the pistil develop simultaneously. The stigma is ready to receive the pollen when the style is ready to serve as the pathway for the pollen tube, which then makes its way to the future seeds (ovules) located within the ovary. But in the present group of plants the ovules are not yet formed when pollination occurs. It takes weeks for the ovules to develop. In the meantime, the pollen tube, which in most plants grows directly to the ovule, remains inactive within the tissues of the style. After the development of the ovule, the pollen tube resumes growth. It approaches the embryo sac by a most unusual path, namely through the basal or chalazal portion of the ovule, instead of by way of the micropyle (a small opening at one end of the ovule). This is termed chalazogamy. Subsequent events, including double fertilization, the development of endosperm, embryo, seed, and fruit are typical of flowering plants.

Delayed fertilization of ovule

Ecology. It is unusual for members of the order to occur in pure stands, although this has been reported for birch in the northern U.S.S.R., the black alder in central Europe, and *Carpinus betulus* in the Rhine Valley. Ordinarily, the members of the order Betulales in Europe are associated with pine, spruce, or aspen, especially along waterways; they also occur on the moors of the north European plain. Some shrubby species form part of the understory in oak and beech forests.

In North America, the forest tree species are the red alder of the Pacific Coast region; the paper birch, which occurs from Alaska to New England; and the black birch and the yellow birch, both of which grow in the northeastern U.S. and extend southward in the Appalachian Mountains. Of these, the red alder is ecologically important in its contribution of great quantities of nitrogen to the soil. This characteristic is shared by the other species of the genus, too. The red alder forms forests in association with Sitka spruce, western hemlock, white spruce, Douglas fir, and others, although it may occasionally form pure stands in stream bottoms and moist flats. Its small fruits germinate freely on mineral soil, and it is often the first tree species to become established, growing so rapidly for the first 25 years that it overtops its associates, but thereafter is overtaken by its chief competitor, the Douglas fir. The paper birch characteristically occurs in regions where summers are short and cool and the winters are cold and snowy. It tolerates a wide range of moisture conditions, but grows best in fresh, well-drained, sandy loams. It rarely occurs in pure stands, but grows with jack pine, or red or white spruce and balsam fir, or with aspen. Despite the vast number of winged seeds produced per tree, the fragility of the newly germinated seedlings contributes to a low rate of survival. Paper birch generally lasts but a single generation in natural succession and is replaced by other more shade-tolerant species. While the black birch is a minor species in many forest types, the yellow birch is a major member of several types of forest, being associated with sugar maple and beech, or hemlock, or balsam fir and pin cherry. Its small fruits germinate well after logging, fire, and windthrow but not in the thick litter of the undisturbed forest.

The river birch, the gray birch, the American hornbeam, and the hop-hornbeam are small trees in the forests of cool temperate North America, the two latter thriving as members of the understory in the rich woods of bottomlands, coves, and protected slopes. The gray birch grows in the dry soils of the northeastern United States and Canada, is an outstanding pioneer, and is notoriously intolerant of shade.

Most shrubby members of the order Betulales occur among the alders, birches, hazels, and both species of *Ostryopsis*. Temperate American species such as *Alnus crispa*, *A. incana*, *Betula glandulifera*, *Corylus americana*, and *C. cornuta* are often thicket formers, especially when they occur at the margins of woods or in the open. The arctic and subarctic species, notably among the birches, occur as low shrubs in the barren lands, in boggy or damp heath-covered areas.

Shrubby members of the birch order

FORM AND FUNCTION

The members of this largely temperate to Arctic group are characterized throughout by their woody habit, which carries with it the necessity for surviving the winter above the surface of the soil. Direct concomitants to this re-

quirement are adaptations that meet the need to conserve water during the winter, a period when water is largely in the frozen state and thus physiologically unavailable. Three major morphological adaptations contribute to the survival of Betulales species in such habitats: (1) the shedding of the leaves; (2) the envelopment of the extremely delicate embryonic cells of the growing point and the immature leaves of the next season with a system of tough, waterproof bud scales; and (3) the development of bark initiated by secondary changes under the original epidermis of the twigs and branchlets. In some species, continued growth of the woody portion of the stem is accompanied by the production of bark that is formed in highly distinctive paperlike layers, although the functional superiority of this peculiar type of bark is unknown.

The characteristic cylindrical male catkins, which are so prominent in most members of the family throughout the winter, are precocious inflorescences composed of a multitude of male flowers whose anthers require but a few days of spring warmth to mature and shed their pollen. This precocity of the male flowers is matched by the production of functional styles and stigmas in the female flowers, which are in small stiff catkins or in somewhat modified buds. The large quantities of tiny pollen grains shed before or at the time of leaf expansion in the spring provide the maximal opportunity to utilize the wind as a pollinating agent.

The fruits are also characteristic structures. In those alders and birches with winged fruits that are freely carried about by the wind, the fruits are so small that there may be over 1,000,000 to the pound (*e.g.*, as is the case in paper birch). The process of dissemination in the birches is facilitated by the shattering of the mature female catkin, which leaves only the catkin axis still attached to the parent tree in the fall. In the alders the female catkin, or ament, does not shatter; instead the scales become woody and remain attached to the ament axis, separating sufficiently so that the little fruits may shake out. Some of the alders have substituted corky ridges for the filmy wings, the corky ridges serving as floats that aid in dissemination by water. In *Ostrya*, *Carpinus*, and *Ostryopsis* the fruits are nutlets that are attached to leafy wings or envelopes. These are shed in later summer, leaving the bare female axis on the tree; dissemination is presumably by small rodents or birds. In the hazelnut, the individual fruit is quite large, in *Corylus americana* averaging about 500 per pound, or each one averaging about 3,000 times the weight of an average paper birch fruit. Dissemination of this eminently edible fruit is by animals.

Seed distribution

EVOLUTION

The evolutionary origins of the Betulales, as in the case of flowering plants generally, cannot as yet be determined from the available fossil evidence, even though there is an abundance of leaf impressions, pollens, and woods. These appear suddenly in the sedimentary rocks of the Mesozoic and Cenozoic Eras (*i.e.*, in rocks as old as about 100,000,000 years), very similar in form to the leaves, pollens, and woods of contemporary genera. In the absence of direct evidence, it is necessary to fall back on information obtained from living forms. Based on living species, the order Betulales is considered by most authorities to be a moderately advanced rather than a primitive flowering plant group. In fact, in some respects, with its often nectarless, odourless, wind-pollinated, unisexual, and very tiny flowers, it may be considered advanced as compared with the larger, bisexual flowers of the presumably primitive flowering plants. The order Betulales has stamens differentiated into anther (pollen sac) and filament (stalk), the latter sometimes split all the way to the base of the filament to form two independent stamen halves. The ovary is inferior (*i.e.*, it is enclosed within the basal portions of the other flower parts, which appear to arise at its upper end), composed of two fused carpels, has but one or two functional ovules, and is crowned with a well-defined style (the elongated upper portion of the ovary). The tiny flowers are clustered together and often are fused with adjacent leafy bracts to form very complex inflorescences. The fruits are indehiscent (*i.e.*, they do not split open along definite lines) single-seeded nuts or nutlets. The order Betulales consists of woody trees or shrubs and in this respect is considered to be only moderately advanced. The leaves are pinnately veined (*i.e.*, with a midrib and branch veins in the fashion of a feather), rather than palmate, and are trilacunar (*i.e.*, having three leaf gaps); there are no vestiges of the ancient double origin of the vascular supply such as is found in the floral organs and cotyledons ("seed leaves") of some flowering plants. The puzzling phenomenon of chalazogamy is difficult to assess from the evolutionary point of view, but the presence of only a single integument in the ovule and the long delay between pollination and fertilization are taken as being advanced. In the structure of the wood there are some examples of somewhat advanced features, but other quite primitive wood characters also occur. All of the above information taken together can be interpreted to indicate that the Betulales probably originated before the Cretaceous Period (*i.e.*, before about 136,000,000 years ago). Other families of flowering plants that may have had a hypothetical ancestor similar to that of the Betulales are found in the Fagales (beech order). A common ancestor for the two orders is difficult to determine, but various proposals have been made, the two plant orders Hamamelidales and Sapindales being favoured. The students of wood anatomy favour the Hamamelid hypothesis. Whichever of these is the more tenable, both schools of thought unite in the rejection of the proposal that the birches, oaks, and other wind-pollinated orders represent ancestral types for the insect-pollinated families.

Evolutionary predecessors

CLASSIFICATION

Annotated classification.

ORDER BETULALES

Woody trees or shrubs, with simple, stipulate, tooth-margined, ultimately deciduous leaves, spirally arranged. Flowers cyclic unisexual, both sexes occurring on the same plant but in different inflorescences. The male flowers with six or fewer stamens, completely lacking vestigial pistils, and with or without six or fewer tepals. The male flowers in cymose groups of three (or fewer) and fused with varying numbers of primary, secondary and tertiary bracts. The anthers two-celled, opening longitudinally. The female flowers without vestigial stamens, with or without tepals inserted at the base of the two styles, the inferior ovary composed of two carpels with axile placentation, but the two compartments confluent above the insertion of ovules. Ovules 1–4, anatropous, and with a single integument. The female flowers in cymose groups of three or fewer and fused with adjacent bracts of varying numbers. The aggregations of male flowers attached to a flexible axis, collectively constituting the characteristic cylindrical catkin or ament. The similar female inflorescence with a stiff axis, which is often erect. Pollination by wind in spring, fertilization long delayed; following production of the ovule, the pollen tube (sometimes branching) invades the ovule by way of the chalaza (chalazogamy). Multicellular archesporium. Maturation of the fruit in late summer or fall. The fruit a one-seeded nutlet (often winged) or nut with a large straight embryo, no endosperm. One family with about six genera and 120 species.

Family Betulaceae

The only family of the order, it has the characteristics of the order.

Tribe Betuleae

Trees and shrubs; the male cymule 3-flowered; male florets with tepals; female florets with 2 to 4 tepals each represented externally by an apical gland, or nude through absence of these glands; fruit a nutlet. Two genera, *Alnus* (about 30 species) and *Betula* (about 40 species).

Tribe Carpineae

Small trees; the male cymule 3-flowered; male florets without tepals, with an average of 6 (rarely 8) or fewer stamens per floret; anther halves and distal portion of filament separate; florets of each male cymule adnate to a 3-parted or entire scale (primary bract plus 2 secondary bracts); 2 female florets per cymule, primary bract free from female florets, secondary and tertiary bracts various; fruit a nutlet. Two genera, *Carpinus* (about 26 species) and *Ostrya* (about 7 species).

Tribe Coryleae

Shrubs or rarely small trees; the male cymule 3-flowered; male florets usually without tepals, with an average of 2 or fewer stamens per floret; female cymules 2-flowered and subtended by a free primary bract, other bracts various. Two genera, *Corylus* (about 15 species), and *Ostryopsis* (2 species).

Critical appraisal. Until recently the Betulales, as presented herein, was treated as one of two families (Betulaceae and Fagaceae) in the order Fagales, but many authorities now regard each former family as a distinct order. Some workers have applied the family name Corylaceae to the group here called Betulaceae, and a few have even divided the group into three separate families—Carpinaceae, Corylaceae, and Betulaceae (corresponding for the most part to the tribes named above). The order as presented here, consisting of one family divided into three tribes, is the most recently accepted classification, however.

BIBLIOGRAPHY. The literature on the order Betulales is rather limited, but that on the "Amentiferae," of which the order has long been considered to be a part, is tremendous. In addition to the following works, and others like them, there are many popular guides to trees and shrubs that are helpful in identifying and learning about local species.

HUBERT WINKLER, "Betulaceae," in A. ENGLER, *Das Pflanzenreich IV*, 61 (1904), the most comprehensive treatment of the taxonomy of the Betulales; A. REHDER, *Manual of Cultivated Trees and Shrubs Hardy in North America, Exclusive of the Subtropical and Warmer Temperate Regions*, 2nd ed. rev. (1962), a treatment of those world species hardy in North America; C.S. SARGENT, *Silva of North America*, vol. 9 (1896), a sumptuous treatment of those species native to North America; H.A. FOWELLS, *Silvics of Forest Trees of the United States, United States Department of Agriculture Handbook 271* (1965); J.M. TRAPPE *et al.* (eds.), *Biology of Alder* (1968); *Woody-Plant Seed Manual, Misc. Publs. U.S. Dep. Agric. 654* (1948); the last three titles are a few among many specialized and rather local publications that treat the ecological aspects of the Betulales; E.C. ABBE, "Studies in the Phylogeny of the Betulaceae," *Bot. Gaz.*, 2 pt., 97:1–67 (1935) and 99:431–469 (1938), a detailed and technical article about floral morphology; J.W. HALL, "The Comparative Anatomy and Phylogeny of the Betulaceae," *ibid.*, 113:235–270 (1952), a technical article about secondary xylem; E. ANDERSON and E.C. ABBE, "A Quantitative Comparison of Specific and Generic Differences in the Betulaceae," *J. Arnold Arbor.*, 15:43–49 (1934), a pioneer article on numerical taxonomy; H. HJELMQVIST, "Studies on the Floral Morphology and Phylogeny of the Amentiferae," *Bot. Notiser*, suppl. 2:122–146 (1948), a different point of view about floral morphology in the group; L.A. KUPRIANOVA, "On a Hitherto Undescribed Family Belonging to the Amentiferae," *Taxon*, 12:12–13 (1963), a paper treating the significance of pollen grains within the Betulaceae; C.R. METCALFE and L. CHALK, *Anatomy of the Dicotyledons*, vol. 2, pp. 1302–1309 (1950), a book about the internal structure of many groups of flowering plants, the pages indicated treating the anatomy of the wood and leaves of representative Betulales; J.C.T. UPHOF, *Dictionary of Economic Plants*, 2nd ed. rev. (1968), a presentation of general economic uses of selected members of the order, as well as of other plants; C. PICKERING, *Chronological History of Plants* (1879), a classic book answering the interesting question of when various members of the Betulales were first mentioned in man's history; D.C. PEATTIE, *A Natural History of Trees of Eastern and Central North America* (1950), a popular but accurate presentation of information on the tree birches and their relatives.

(E.C.A.)

Beust, Friedrich von

As a leading Saxon and Austrian minister and statesman, Friedrich Ferdinand Graf von Beust championed the struggle against Bismarck and against Prussian hegemony among the German states. Following Austria's defeat by Prussia in the Seven Weeks' War of 1866, Beust negotiated the *Ausgleich* or "Compromise" (1867), which established the Austro-Hungarian monarchy and helped to restore the Habsburg's international position.

The Saxon phase. A descendant of the Saxon line of an ancient aristocratic family, Beust was born in Dresden on January 13, 1809, the youngest son of Karl Leopold von Beust, a high-ranking member of the law administration. After attending the Kreuzschule in Dresden, he studied law at Göttingen and Leipzig. Relatively progressive, he was inclined toward liberal constitutionalism and devoted himself also to the study of philosophy, history, and politics. Having finished his academic studies in 1830, he began his diplomatic career, which led him first to Saxon missions in Berlin and Paris; he became head of missions in Munich, London, and, in 1848, Berlin. By that time the versatile diplomat had acquired a remarkable practical knowledge of constitutional matters and of important political personalities, among them Bismarck.



Beust, engraving by Joseph Anton Bauer (1820–1904).
By courtesy of the Bild-Archiv, Osterreichische Nationalbibliothek, Vienna

The Saxon king, Frederick Augustus II, favoured the movements for German unity developing out of the revolutions of 1848 but not its movements toward democracy. In 1849 he called for Beust, whom he thought to be a stabilizing influence, and appointed him his new minister for foreign affairs. Beust called in Prussian troops to suppress the popular uprisings in Dresden in May and negotiated a conservative alliance of Saxony, Prussia, and Hanover later the same month. He thus proved from the outset to be the dominant force in the Cabinet. As minister of the interior (from 1853) he sought to expand Saxon economy through a policy of moderate internal reforms. In his foreign policy he aimed at setting up the smaller German states as a third force between Austria and Prussia. Although Saxony's economic interests naturally gravitated toward Prussia, Beust's foreign policy increasingly drew him nearer to Austria, thus forcing him into an ever sharpening opposition to Bismarck.

The Austrian phase. After Prussia's victory over Austria and its Saxon ally in 1866, Beust had to give up his office under pressure from Bismarck. Beust, who in Saxony had felt like a "horse harnessed to a perambulator," unexpectedly found a wider field of activity. In October 1866 the emperor Francis Joseph appointed him Austrian minister for foreign affairs and in February 1867 imperial chancellor of the Habsburg monarchy (minister president in June 1867). Although a foreigner and a Protestant in Roman Catholic Vienna, Beust, optimistic as usual, did not hesitate to accept this difficult office, which challenged his ambition and flattered his vanity. His diplomatic experience, political imagination, and a certain cynicism helped him to master his new tasks quickly. He restored constitutional government and succeeded in temporarily solving the most difficult internal problem by bringing about a compromise (*Ausgleich*) with Hungary, allotting to the Hungarians hegemony in the eastern part of the monarchy. Nevertheless, Beust's attempt throughout Germany to regain popular sympathy for Austria by establishing a model liberal regime was doomed to failure. In the long run he was not able to harmonize the status of the Germans—the dominant group within the western part of the empire—with the claims of the other nationalities subjected to Habsburg rule, particularly the embittered Czechs.

Chancellor of Austria, 1867–71

Beust's foreign policy was finally defeated by national ambitions whose force he underrated. Neither Francis Joseph nor Beust was willing to accept the consequence of Prussia's defeat of Austria in 1866—namely, Austria's enforced exclusion from Germany. Beust was dominated by the idea of revenge, by his rivalry with Bismarck, and by his fear lest Prussia lay hands upon the Habsburgs' German territories. He therefore tried first to prevent the southern German states from uniting with Prussia; he wanted to create the preconditions for later re-establishing Austria's old hegemony, though in an improved form. In pursuing his plan for an alliance with France and Italy to corner Prussia in the East, he proved to be a formidable adversary of Bismarck, much superior to Napoleon III of France in political skill and cleverness. Although his policy moved on the brink of a new war between Austria and Prussia, he was realistic and responsible enough to abstain from any attempt to involve Austria-Hungary in the Franco-German War.

Beust was denied the ultimate realization of his political aims; in 1871 he had to recognize that the German *Reich* under Prussian leadership, excluding Austria, had become a historical reality. Francis Joseph, who had raised him to the rank of count in 1868, dismissed him from his post as chancellor in October 1871. Beust continued to serve the Habsburg monarchy as ambassador in London and, after 1878, in Paris until his retirement in 1882. He died on October 24, 1886, in the castle Altenberg near Vienna.

Dismissal as chancellor

Beust was the last of the Austrian statesmen who felt bound to further the German mission of the Habsburgs. He was rooted more in the tradition of 18th-century cabinet politics than in 19th-century nationalism. His policy, although strongly influenced by emotion, was moderate, not lacking a certain grandeur but never attaining decisive victory. As Bismarck remarked in 1866, "There is no room for both of us in Germany."

BIBLIOGRAPHY. An adequate biography is still lacking. Beust's memoirs, *Aus drei Viertel-jahrhunderten* (Eng. trans. ed. by H. DE WORMS, 2 vol., 1887), are cleverly written but complacent and self-justificatory. The work of F.W. EBELING, *F.F. Graf von Beust*, 2 vol. (1870–71), is an uncritical eulogy of Beust's role in Saxony, which did not even meet with Beust's approval. The article by B. ERDMANNSDORFER in *Allgemeine deutsche Biographie*, vol. 46 (1902), is largely obsolete. E. GROB, *Beusts Kampf gegen Bismarck* (1930), is a short but unbiassed and useful essay. Beust's politics in Austria are presented in detail by H. POTTHOFF, *Die deutsche Politik Beusts* (1968), which includes literature references. General information may be found in A.J. MAY, *The Habsburg Monarchy, 1867–1914* (1951); and A.J.P. TAYLOR, *The Struggle for Mastery in Europe, 1848–1918* (1954).

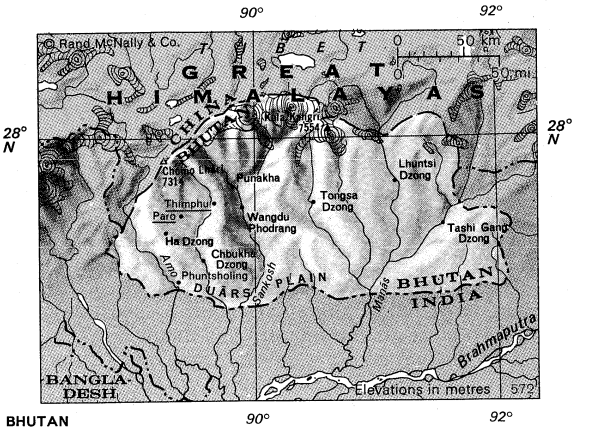
(H.P.)

Bhutan

Along the lofty ridges of the eastern Himalayas lies the sovereign Kingdom of Bhutan. With an area of about 18,000 square miles (47,000 square kilometres), Bhutan is larger than Switzerland; its location between the Assam–Bengal Plain of India to the south and the Chinese-occupied Tibetan Plateau to the north gives it considerable geopolitical significance.

Bhutan's undefined traditional boundary with Tibet follows for the most part the crest of the Great Himalayan Range. Chinese maps show parts of Bhutan's territory within China, and occasional minor incursions of Chinese troops have been reported along the border. Bhutan's boundary with India lies to the south of the Himalayan Range in the Duārs Plain. The Indo-Bhutanese boundary evolved as a result of the British annexation of a major portion of the Duārs region of Bhutan adjoining Assam and Bengal between 1841 and 1864. To the east, Bhutan borders Arunachal Pradesh, formerly the North East Frontier Agency of India, and to the southwest it borders the Indian state of Sikkim, south of the Chumbi Valley in Tibet.

The historic isolation of Bhutan is rapidly disappearing, and forces of change are accelerating. King Jigme Dorji Wangchuk (reigned 1952–72) made drastic alterations in the system of government that may lead to establishment of a constitutional monarchy. Progress was made in the development of transportation and communication after 1960, when the trip from the Indian border to the Bhutanese capital, Thimphu (Thimbu), took six days by mule. In 1965 the journey could be made in six hours by jeep along the 120-mile winding mountain road from Phuntsholing, on the border, to Thimphu. The regular flight between Hashimara, in West Bengal, and Paro, near Thimphu, takes only 30 minutes.



The kingdom has a potentially adequate economic base to support its estimated 1,150,000 people (1974). For the time being, however, it remains less economically developed than the other Himalayan border kingdom, Nepal, or the Indian Himalayan state, Sikkim. The economic core of the nation lies in the fertile valleys of the Lesser Himalayas, which are separated from one another by a series of high and complex interconnecting ridges extending across the country from north to south. The political nucleus of Bhutan lies in the Paro and Thimphu valleys in the Lesser Himalayan region. Thimphu is the official capital.

Geographic regions. Physically, Bhutan may be divided into three regions: the Great Himalayas, the Lesser Himalayas, and the Duārs Plain.

The Great Himalayas. The northern part of Bhutan lies within the Great Himalayas; the snow-capped peaks in this region attain a height of more than 24,000 feet (7,300 metres). High valleys occur at elevations of 12,000 to 18,000 feet, running down from the great northern glaciers. The alpine pastures on the high ranges are used for grazing yaks in the summer months. North of the Great Himalayas are several "marginal" mountains of the Tibetan Plateau, forming the principal watershed between the rivers respectively southward and northward. Dry climate is characteristic of the Great Himalayan region.

Until about 1960, the tempo of life continued in the Great Himalayas much as it had for centuries. Long undisturbed in their ways, Bhutanese traders carried cloth, spices, and grains across the mountain passes into Tibet and brought back salt, wool, and sometimes herds of yaks. The absorption of Tibet by China, however, broke the tranquil isolation and disturbed the traditional way of living in these high regions, where military precautions have been taken to guard against the potential danger of a Chinese incursion from Tibet. Scores of Tibetan refugees swarm through the Great Himalayan region.

The Lesser Himalayas. Spurs from the Great Himalayas radiate southward, forming the ranges of the Lesser (or Inner) Himalayan region. The north–south ranges of the Lesser Himalayas comprise watersheds between the principal rivers of Bhutan. Among these, the Black Mountain Range forms the watershed between the Sankosh and Manas rivers. Differences in elevation and the degree of exposure to moist southwest monsoon winds encourage vegetation, ranging from dense forest on the rainswept windward slopes to alpine vegetation at higher elevations. Several fertile valleys of central Bhutan are in the Lesser Himalayas at elevations vary-

ing from 5,000 to 9,000 feet. These valleys are relatively broad and flat, receiving moderate rainfall (estimated at from 40 to 50 inches or less a year), and are fairly well populated and cultivated. Among them, the Paro (6,950 feet), Punākha (5,170 feet), Thimphu (8,400 feet), and Ha (8,500 feet) are the best known.

The Duars Plain. South of the Inner Himalayas and the foothills lies the narrow Duars Plain, which forms a strip eight to ten miles wide along the southern borders of Bhutan. The Himalayan ranges rise sharply and abruptly from the narrow Duars Plain, which controls access to the strategic passes (known as doors or *dooars*) through the mountains leading into the fertile valleys of the Inner Himalayas. Subject to excessive rainfall (estimated to amount to between 200 and 300 inches a year), the entire Duar tract is unhealthy, hot, and steamy, and is covered with dense semitropical forest as well as undergrowth.

The northern part of the Duars immediately bordering the mountains consists of a rugged, irregular, and sloping surface. At the foot of the mountains small villages are found in forest clearings, but most of the area is covered with dense vegetation inhabited by elephants, deer, tigers, and other wild animals. The southern part of the Duars bordering India is mostly covered with savanna (grassy parkland) and bamboo jungle. In many areas the savanna grasslands are being cleared for rice cultivation. Several Bhutanese market centres and towns, such as Phuntsholing, situated along the India–Bhutan border, are rapidly growing in size. The principal trade routes between central Bhutan and India pass through these centres, following the valleys of the main rivers.

History. The historical origin of Bhutan is obscure. Old manuscripts found in various monasteries contain historical records, but these have not been sufficiently researched to provide an authentic account of Bhutan's early history. It is reported that over three centuries ago an influential lama from Tibet, Sheptoon La-Pha, became the king of Bhutan and acquired the title of Dharma Raja. It seems probable that Bhutan became a distinct political entity about this period. La-Pha was succeeded by Doopgein Sheptoon, who consolidated Bhutan's administrative organization through the appointment of *penlops* (governors of territories) and *jungpens* (governors of forts). Doopgein Sheptoon exercised both temporal and spiritual authority, but his successor confined himself only to the spiritual role and appointed a minister to exercise the temporal power. The minister became the temporal ruler and acquired the title of Deb Raja. This institution of two supreme authorities—a Dharma Raja for spiritual affairs, and a Deb Raja for temporal matters—existed until the death of the last Dharma Raja in the early 20th century. Succession to the spiritual office of Dharma Raja was dependent upon what was considered a verifiable reincarnation of the deceased Dharma Raja. When the last Dharma Raja died no reincarnation was found, and the practice and the office ceased to exist.

Spiritual and temporal authorities

For much of the 19th century Bhutan was plagued by a series of civil wars as the *penlops* of the various territories contended for power and influence. The office of the Deb Raja, in theory filled by election by a council composed of *penlops* and *jungpens*, was in practice held by the strongest of the governors, usually either the *penlop* of Paro or the *penlop* of Tongsa. Similarly, the *penlops*, who were to be appointed by the Deb Raja, in practice fought their way into office. Throughout most of Bhutanese history a continuous series of skirmishes and intrigues took place throughout the land as superseded *jungpens* and *penlops* awaited an opportunity to return to power.

In 1907, the *penlop* of Tongsa—the most powerful of the *penlops* to appear for many years—became the hereditary king (Druk Gyalpo) of Bhutan. The present king, Jigme Singye Wangchuk, is the fourth in this line of hereditary rulers; his great grandfather was recognized by the British as the sole ruler of Bhutan at the beginning of the 20th century.

Jigme Dorji Wangchuk stressed the rapid economic, political, and social development of the country while at the same time seeking to minimize as far as possible the destruction of traditional values. His wife, Queen Kesang, the niece of the king of Sikkim, attended school in Europe.

People and population. With a population of more than 1,000,000, Bhutan is the second most populous Himalayan kingdom after Nepal. Its population density, however, about 46 persons per square mile, is the lowest among the three Himalayan kingdoms. The population is increasing at a rate of perhaps 2 percent a year. The most sparsely populated sections are the rugged Great Himalayan region and the unhealthy malarial ranges bordering the Duars. Adverse physical conditions in both areas limit most of the population to the fertile Inner Himalayan valleys of central Bhutan.

Most of the Bhutanese—over one-half—consist of people of Tibetan extraction known as Bhote or Bhutias, who share a common heritage of Tibetan culture, language, and religion. They are dominant in northern and central Bhutan and speak a variety of Tibetan dialects with a high degree of mutual intelligibility. Their written language is identical with Tibetan, and they practice lamaistic Buddhism.

Bhutias and Nepalese

In southern and southwestern Bhutan an ethnically mixed population with a predominance of Nepalese settlers is found. The Nepalese are members of the Rai, Gurung, and Limbu ethnic groups; they speak Nepali and practice Hinduism. Little assimilation takes place between the Nepalese and Tibetan groups. Since 1959 Nepalese immigration has been stopped, and people of Nepalese origin, although Bhutanese citizens, are prohibited from settling in central Bhutan. Along the whole length of southern Bhutan, where the mountains meet the Duars Plain, the presence of the Nepalese settlers contributes to the development of the region, but also adds to political discontent and instability. Discrimination shown against the Nepalese constitutes a major internal political problem for Bhutan.

In eastern Bhutan the majority of the people are similar to those living in Arunachal Pradesh. They are known by various group names, such as Bhote, Monpa, and Sherdukpen. Although Buddhists, they are less strict in their observance of religious customs, so that fewer monasteries and lamas are to be found in eastern Bhutan.

Despite centuries of living together, different ethnic groups retain their distinguishing characteristics. Differences in language, religion, and ethnic origin affect the political life of the kingdom. The people of Bhutan generally think of themselves in terms of their respective tribes or of their ethnic origins.

The national economy. The natural resources of Bhutan have not been surveyed, and the precise extent of agricultural, forest, mineral, and power resources is not yet known, although an inventory of the country's natural resources is being taken in connection with Bhutan's development plan.

Agricultural resources. The Bhutanese economy is mainly agrarian; most of the population is engaged in agriculture and livestock raising. A land-use reconnaissance carried out in 1961 and 1962 indicated that the amount of land available for agriculture is only a fraction of the total area of the country. An adverse climate, poor soil, and steep slopes have made it necessary to leave a large land area covered with forest growth, Alpine meadows, and grasslands. The relatively low, well-watered, and fertile valleys of central Bhutan have the largest percentage of cultivated land. Because of the great variations in altitude and climate, a variety of crops is grown in Bhutan. Rice, maize, and buckwheat grow well up to an elevation of 4,000 feet. Barley alternates with rice up to about 8,000 feet; wheat grows up to 9,000 feet.

Progressive changes in farming practices are being introduced to increase the productivity of agriculture. Since 1964, 11 state demonstration farms have been established, seven in the temperate zone and four in the subtropical zone. Four government orchards recently established in the temperate zone specialize in the growing of apples, pears, peaches, plums, apricots, and walnuts. A

Increasing agricultural productivity

large number of private orchards also have been established during the past few years. Thousands of fruit plants have been distributed to farmers to popularize fruit growing, as the soil and climate are eminently suitable for the purpose.

A major emphasis on agriculture since 1967 has been on the development of minor irrigation schemes. Extension work in agriculture and horticulture is being carried out to popularize high-yielding varieties of seeds, better implements, increased use of fertilizers, and improved methods of agriculture. High-yielding varieties of paddy, wheat, and maize seeds are being distributed to Bhutanese farmers and are producing encouraging results.

A silkworm farm was recently established at Kanglung in eastern Bhutan to produce raw silk for the local weaving industry. Six livestock farms and two sheep breeding farms also have been established. Wool, which is an essential raw material for rural, or cottage, industry has become extremely scarce since the import of wool from Tibet was stopped by the government. Four veterinary dispensaries and one mobile veterinary-dispensary unit have been set up to improve livestock. One fish hatchery has been established, and the program of stocking rivers and lakes all over the country with brown trout has made satisfactory progress. A cheese-making project has been set up in western Bhutan under the guidance of a Swiss expert.

Forest resources. About 70 percent of the country is covered with forests. A survey and a demarcation of forests covering an area of about 1,000 square miles have been carried out in Paro, Ha, and Thimphu valleys in western Bhutan. A survey of medicinal plants has been completed and the government is considering possibilities of exploiting them commercially. A 62-square-mile wildlife sanctuary has been established at Manas in southern Bhutan; it contains many valuable species of animals, including the golden langur (a slender long-tailed monkey), which is rare elsewhere in the world. The government has established a sawmill with a furniture-making unit at Thimphu. There are also many privately operated sawmills operating all over the country.

The five-year plans. The successful completion of the first five-year plan (1961–66) and second five-year plan (1966–71) has resulted in a marked degree of prosperity in the country. During the second five-year plan, many available services and facilities were more than doubled. The success of the two plans has depended largely on the regular flow of funds from India and upon the availability of Indian technical personnel. Shortages of unskilled labour for completing various projects also have been overcome by importing large numbers of workers from neighbouring India.

Imported labour

With increased activity in all sectors of economy, an acute shortage of road transport for carrying essential materials has developed. The completion of some projects, such as hydroelectric stations, schools, and hospitals, has been delayed because of lack of transport. Two small hydroelectric stations designed to meet local power needs were completed at Thimphu and Paro during 1967, and another station is nearing completion at Wangdü Phodrang. With an abundance of potential water power in various parts of the country, small hydroelectric projects are expected to provide electricity to all major townships in the future.

Management of the economy. The export of vegetables, oranges, cardamon, and timber to India has been steadily increasing. Exports from small-scale industries, such as distilleries, food canneries, and a nut and bolt factory, have also increased. The transition from a barter to a money economy is taking place rapidly. Taxes, which were payable mostly in kind in the past, are now being levied in cash. The increased amount of money in circulation is reflected in the growing transactions of the state-owned Bank of Bhutan. The government's internal revenue in the financial year 1968–69 amounted to approximately 11,870,000 Indian rupees (Rs. 7.50 = $1 U.S.; Rs. 18 = £1 sterling on December 1, 1970), most of which was spent on the maintenance and expansion of administrative services and in building several townships, including the capital, Thimphu. Development projects are being financed mostly by grants from the government of India. Assistance from countries other than India has been negligible.

Bhutan has a deficit budget and new sources of revenue are being sought. The principal revenue sources are taxes on land, houses and cattle, transport, liquor, and forest products. For the benefit of the rural people substantial tax relief was recently granted on land and cattle. Despite the deficit budget the government is reluctant to impose any tax that would increase the burden on villagers. Bhutan is one of the few countries in the world where there is no income tax.

One of the new sources of revenue is tourism. The number of visitors to Thimphu and Paro, not all of whom are Indian, is steadily increasing. With further improvement in communications, and with Bhutan's membership in the United Nations, many more tourists are expected, and the government is building a number of tourist lodges in different parts of the country. Revenue is also obtained from Bhutan's postage stamps; it is estimated that in 1969 Bhutan sold stamps worth $100,000.

Transport and communications. Bhutan's development plans have stressed the improvement of transport and communication. The 120-mile Phuntsholing to Paro–Thimphu national highway is part of a network of roads the Bhutan government plans to build to open up regions that are still largely inaccessible. The total length of the proposed network is 1,500 miles, of which about 550 miles have been constructed through difficult mountain terrain, linking Indian roads to Thimphu and to Paro in western Bhutan, to Tongsa in central Bhutan, and to Tashigang in eastern Bhutan. Work has started on the construction of a lateral east to west road. The road from Paro to Ha has been completed, and the Tashigang to Mongar road is fast nearing completion. Responsibility for the construction of major highways has been assigned to the Indian Border Roads Organization. The Bhutan government has developed its own engineering services to undertake construction of feeder roads.

Highways

A commercial air service operates between Paro in western Bhutan and Calcutta in eastern India. Indian engineers have assisted the Bhutan government in laying telephone lines and exchanges. The principal administrative centres of Bhutan have telecommunication links with India. A postal service between Bhutan and other countries was inaugurated in 1963, when the first series of Bhutanese stamps were issued. A teletypewriter service links Phuntsholing with the outside world; Thimphu, the capital, is also to have a teletypewriter service installed.

The Bhutan State Transport Department has a fleet of 47 trucks and buses. The expansion of economic activity as well as road construction has led to an increase in the movement of people and goods. At present it is estimated that Bhutan requires a minimum of 300 vehicles to maintain transport facilities.

Administration and social conditions. *The structure of government.* The Druk Gyalpo, as King Jigme Dorji Wangchuk was called, wished to develop Bhutan as a modern country without causing it to lose its ancient culture and identity; he took the initiative in adapting the system of government to the changing times. He appointed a council of ministers and shared administrative responsibility, which was formerly his alone. Despite his influence over the people, and the supreme authority he enjoyed, the King was not impervious to suggestions. The 140-member national assembly (Tsongdu) was at times quite critical of the administration, and its deliberations were not disregarded by the government. Half of the members of the national assembly are nominated and the rest are indirectly elected. The King made it clear that the process of democratization would continue, but that changes would not be sudden lest valuable elements in the traditional pattern of life be destroyed. Further administrative changes and reforms are envisaged if the present structure continues to work satisfactorily. There are no political parties.

Justice. A recent development that has carried Bhutan further on the road to modernity is the establishment of

a high court. Bhutan has codified law but no separate judiciary. Judicial duties are vested in the executive, and lawyers are unknown. The high court functions as an appelate authority above the executive. Four judges have already been appointed. The king, however, remains the supreme tribunal.

Education. An expansion of educational facilities is bringing about major social changes. At present there are 105 schools with over 16,000 students. Girls are also taking advantage of educational opportunities. Lest the rapid spread of education without a matching expansion of employment opportunities should create an unemployment problem, which is currently unknown, emphasis is being gradually shifted to vocational education. A national museum has been established at Paro. A collection of valuable manuscripts for the proposed national library at Thimphu has been started. A large number of school textbooks are being produced in Dzongkha (a Tibetan dialect), the official language of Bhutan.

Health. As a result of health measures being taken by the government, the population of the country is increasing. Most of the diseases that formerly caused heavy mortality are being controlled. Four hospitals, 28 dispensaries, and three leprosy hospitals have been established. The reduction in the death rate and the consequent increase in population has led to an expansion of the area under cultivation.

Wages and cost of living. The infusion of large sums of money into the economy as a result of developmental and other activities has created inflationary tendencies. The cost of living has more than doubled during the past few years and continues to show an upward trend. The influx of large numbers of technical personnel and labourers from outside Bhutan has also contributed to the increased cost of living. There has been an appreciable improvement in real income, however, and wages and farm incomes have risen in response to an increased demand.

Prospects for the future. With stable government being maintained and orderly economic and social development taking place under the regime of King Wangchuk, Bhutan faces a secure future. Bhutan has, however, quietly intensified military preparations to guard against any Chinese incursions from neighbouring Tibet. A national militia is being planned, and every male between the ages of 18 and 38 will have five months of compulsory military training. The kingdom's small, Indian-equipped army, consisting of about 5,000 soldiers, is to be expanded by several thousand. Bhutanese soldiers, all volunteers serving minimum five-year-enlistment terms, receive training in guerrilla warfare, which is suited to the mountainous terrain. While Bhutan has a small detachment of Indian military instructors, there are no Indian combat troops in the kingdom. India, however, has a string of army camps along Bhutan's southern border.

Defense

Once isolated from the mainstream of world affairs by steaming jungles in the south and snow-covered mountain ranges in the north, Bhutan now finds itself caught between the old and the new, with its ruler seeking to adapt the nation's way of life to 20th-century opportunities and alternatives. As Bhutan encounters new challenges, its people alter or discard in varying degrees many skills, values, and attitudes that served them well in the past. At the same time, great pride is maintained in traditional culture; there has been a revival of centuries-old lama dances and folk songs, as well as of ancestral skills and handicrafts. As new ways enter national life, a new group of traders, businessmen, and bureaucrats is emerging in the country. In the Thimphu bazaar the number of shops is increasing, and imported goods are displayed side by side with traditional Bhutanese products.

BIBLIOGRAPHY. P.P. KARAN, *Bhutan: A Physical and Cultural Geography* (1967), is a survey of significant aspects of Bhutan's physical and cultural characteristics, including a useful summary of modern research, and several maps and photographs.

(P.P.K.)

Biblical Literature

Biblical literature in this article comprises the Old Testament writings according to the Hebrew canon; intertestamental works, including the Old Testament Apocrypha; the New Testament writings; and the New Testament Apocrypha.

The article is divided into the following sections and subsections:

I. Nature and significance
  Historical and cultural importance
    In Judaism
    In Christianity
  Major themes and characteristics
  Influences
    On Western civilization
    On the modern secular age
II. Old Testament canon, texts, and versions
  The canon
    The Hebrew canon
    The Christian canon
  Texts and versions
    Textual criticism: manuscript problems
    Textual criticism: scholarly problems
    Texts and manuscripts
    Early versions
    Later and modern versions: English
    Later and modern versions: continental
    Non-European versions
III. Old Testament history
  Early developments
    Background and beginnings
    Exodus and conquest
    The tribal league
    The united monarchy
  From the period of the divided monarchy through the restoration
    The divided monarchy: from Jeroboam I to the Assyrian conquest
    The final period of the kingdom of Judah
    The Babylonian Exile and the restoration
IV. Old Testament literature
  The Torah (Law, Pentateuch, or Five Books of Moses)
    Composition and authorship
    Genesis
    Exodus
    Leviticus
    Numbers
    Deuteronomy
  The Nevi'im (the Prophets)
    The canon of the Prophets
    Hebrew prophecy
    Joshua
    Judges
    Samuel
    Kings
    Isaiah
    Jeremiah
    Ezekiel
    The Twelve
  The Ketuvim
    Psalms
    Proverbs
    Job
    The Megillot (the Scrolls)
    Daniel
    Ezra, Nehemiah, and Chronicles
V. Intertestamental literature
  Nature and significance
    Definitions
    Texts and versions
    Persian and Hellenistic influences
    Apocalypticism
  Apocryphal writings
    Apocryphal works indicating Persian influence
    Apocryphal works lacking strong indications of influence
    Additions to Daniel and Esther
    Greek additions to Esther
    I and II Maccabees
    Wisdom literature
  The pseudepigraphal writings
    Works indicating a Greek influence
    Apocalyptic and eschatological works
    Pseudepigrapha connected with the Dead Sea Scrolls
  Qumrān literature (Dead Sea Scrolls)

VI. New Testament canon, texts, and versions
  The New Testament canon
    Conditions aiding the formation of the canon
    The process of canonization
  Texts and versions
    Textual criticism
    Texts and manuscripts
    Versions
VII. New Testament history
  The Jewish and Hellenistic matrix
  Jewish sects and parties
  The religious situation in the Greco-Roman world of the 1st century AD
  Adaptation of the Christian message to the Hellenistic religious situation
  The life of Jesus
  The chronology of Paul
VIII. New Testament literature
  Introduction to the Gospels
    Meaning of the term gospel
    Form criticism
  The Synoptic problem
    Early theories about the Synoptic problem
    The two- and four-source hypotheses
  The Synoptic Gospels
    The Gospel According to Mark
    The Gospel According to Matthew
    The Gospel According to Luke
  The Fourth Gospel: The Gospel According to John
  The Acts of the Apostles
    The purpose and style of Acts
    The content of Acts
  The Pauline Letters
    The Letter of Paul to the Romans
    The First Letter of Paul to the Corinthians
    The Second Letter of Paul to the Corinthians
    The Letter of Paul to the Galatians
    The Letter of Paul to the Ephesians
    The Letter of Paul to the Philippians
    The Letter of Paul to the Colossians
    The First Letter of Paul to the Thessalonians
    The Second Letter of Paul to the Thessalonians
  The Pastoral Letters: I and II Timothy and Titus
    The Pastoral Letters as a unit
    Content and problems
    The Letter of Paul to Philemon
  The Letter to the Hebrews
  The Catholic Letters
    The Letter of James
    The First Letter of Peter
    The Second Letter of Peter
    The Johannine Letters: I, II, and III John
    The Letter of Jude
  The Revelation to John
IX. New Testament Apocrypha
  Nature and significance
  The New Testament apocryphal writings
X. Biblical literature in liturgy
  Biblical literature in the liturgy of Judaism
  Biblical literature in the liturgy of Christianity
    Eastern Orthodoxy
    Roman Catholicism
    Protestantism

I. Nature and significance

The larger part of the Bible, the Old Testament, consists of writings that were first brought together and preserved as the sacred books of the ancient Hebrew people. As the Bible of the Hebrews and their Jewish descendants down to the present, these books have been perhaps the most decisive single factor in the preservation of the Jews as a people and Judaism as a religion. Joined with the New Testament in Christianity's Bible, they have played a special role in the history and culture of the modern world, especially in the West.

HISTORICAL AND CULTURAL IMPORTANCE

In Judaism. After the kingdoms of Israel and Judah had fallen, in 722 BCE (before the Common Era, equivalent to BC) and 587/586 BCE, respectively, the Hebrew people outlived defeat, captivity, and the loss of their national independence, largely because they possessed writings that preserved their history and traditions. Many of them did not return to Palestine after their exile. Those who did return did so to rebuild a temple and reconstruct a society that was more nearly a religious community than an independent nation. The religion found expression in the books of the Old Testament: books of the Law (Torah), history, prophecy, and poetry. The survival of the Jewish religion and its subsequent incalculable influence in the history of Western culture are difficult to explain without acknowledgment of the importance of the biblical writings.

Expression of the Jewish religion in the Old Testament

When the Temple in Jerusalem was destroyed in 70 CE (Common Era, equivalent to AD), the historical, priestly sacrificial worship that centred in it came to an end and was never to be resumed. But the religion of the Jewish people had by that time gone with them into many lands, where it retained its distinctive character and vitality because it could still draw its nurture from biblical literature. The Bible was with them in their local synagogues, where it was read, prayed, and taught. It preserved their identity as a people, inspired their worship, arranged their calendar, permeated their family lives; it shaped their ethical ideals, sustained them in persecution, and touched their intellects with intensity. Whatever Jewish talent and genius have contributed to the enrichment of the life of Western people is due in no small degree to the influence of the Bible.

In Christianity. The Hebrew Bible is as basic to Christianity as it is to Judaism. Without the Old Testament, the New Testament could not have been written and there could have been no man like Jesus; Christianity could not have been what it became. This has to do with cultural values, basic human values, as much as with religious beliefs. The Genesis stories of prehistoric events and people are a conspicuous example. The Hebrew myths of creation have superseded the racial mythologies of Latin, Germanic, Slavonic, and all other Western peoples. This is not because they contain historically factual information or scientifically adequate accounts of the universe, the beginning of life, or any other subject of knowledge, but because they furnish a profoundly theological interpretation of the universe and human existence, an intellectual framework of reality large enough to make room for developing philosophies and sciences.

Centrality of the one and only God

This biblical structure of ideas is shared by Jews and Christians. It centres in the one and only God, the Creator of all that exists. All things have their place in this structure of ideas. All mankind is viewed as a unity, with no race existing for itself alone. The Covenant people (*i.e.*, the Hebrews in the Old Testament and Christians in the New Testament) are chosen not to enjoy special privileges but to serve God's will toward all nations. The individual's sacred rights condemn his abuse, exploitation, or neglect by the rich and powerful or by society itself. Widows, orphans, the stranger, the friendless, and the helpless have a special claim. God's will and purpose are viewed as just, loving, and ultimately prevailing. The future is God's, when his rule will be fully established.

The Bible went with the Christian Church into every land in Europe, bearing its witness to God. The church, driven in part by the power of biblical themes, called men to ethical and social responsibility, to a life answerable to God, to love for all men, to sonship in the family of God, and to citizenship in a kingdom yet to be revealed. The Bible thus points to a way of life never yet perfectly embodied in any society in history. Weighing every existing kingdom, government, church, party, and organization, it finds them wanting in that justice, mercy, and love for which they were intended.

MAJOR THEMES AND CHARACTERISTICS

The Bible is the literature of faith, not of scientific observation or historical demonstration. God's existence as a speculative problem has no interest for the biblical writers. What is problematical for them is the human condition and destiny before God.

The great biblical themes are about God, his revealed works of creation, provision, judgment, deliverance, his covenant, and his promises. The Bible sees what happens to mankind in the light of God's nature, righteousness, faithfulness, mercy, and love. The major themes about mankind relate to man's rebellion, his estrangement and perversion. Man's redemption, forgiveness, reconcilia-

tion, the gifts of grace, the new life, the coming kingdom, and the final consummation of man's hope are all viewed as the gracious works of God.

The Old Testament contains several types of literature: there are narratives combined with rules and instructions (Torah, or Pentateuch) and anecdotes of Hebrew persons, prophets, priests, kings, and their women (Former Prophets). There is an antiracist love story (Ruth), the story of a woman playing a dangerous game (Esther), and one of a preacher who succeeded too well (Jonah). There is a collection of epigrams and prudential wisdom (Proverbs) and a philosophic view of existence with pessimism and poise (Ecclesiastes). There is poetry of the first rank, devotional poetry in the Psalms, and erotic poetry in the Song of Songs. Lamentations is a poetic elegy, mourning over fallen Jerusalem. Job is dramatic theological dialogue. The books of the great prophets consist mainly of oral addresses in poetic form.

The New Testament also consists of different literary forms. Acts is historical narrative, actually a second volume following Luke. A Gospel is not a history in the ordinary sense but an arrangement of remembered acts and sayings of Jesus retold to win faith in him. There is one apocalypse, Revelation (a work describing the intervention of God in history). But the largest class of New Testament writings is epistolary, the letters of Paul and other Apostles. Originally written to local groups of Christians, they were preserved in the New Testament and given the status of doctrinal and ethical treatises.

INFLUENCES

On Western civilization. The Bible brought its view of God, the universe, and mankind into all the leading Western languages and thus into the intellectual processes of Western man. The Greek translation of the Old Testament made it accessible in the Hellenistic period (*c.* 300 BCE–*c.* 300 CE) and provided a language for the New Testament and for the Christian liturgy and theology of the first three centuries. The Bible in Latin shaped the thought and life of Western people for a thousand years. Bible translation led to the study and literary development of many languages. Luther's translation of the Bible in the 16th century has been called the beginning of modern German. The Authorized Version (English) of 1611 (King James Version) and the others that preceded it caught the English language at the blooming of its first maturity. Since the invention of printing (mid-15th century), the Bible has become more than the translation of an ancient Oriental literature. It has not seemed a foreign book, and it has been the most available, familiar, and dependable source and arbiter of intellectual, moral, and spiritual ideals in the West.

Influence on Western thought and language

Millions of modern people who do not think of themselves as religious live nevertheless with basic presuppositions that underlie the biblical literature. It would be impossible to calculate the effect of such presuppositions on the changing ideas and attitudes of Western people with regard to the nature and purpose of government, social institutions, and economic theories. Theories and ideals usually rest on prior moral assumptions—*i.e.*, on basic judgments of value. In theory, the West has moved from the divine right of kings to the divinely given rights of every citizen, from slavery through serfdom to the intrinsic worth of every person, from freedom to own property to freedom for everyone from the penalties of hopeless poverty. Though there is a wide difference between the ideal and the actual, biblical literature continues to pronounce its judgment and assert that what ought to be can still be.

On the modern secular age. The assumption of many people is that the Bible has lost much of its importance in a secularized world; that is implied whenever the modern period is called the post-Judeo-Christian era. In most ways the label fits. The modern period seems to be a time in which unprecedented numbers of people have discarded traditional beliefs and practices of both Judaism and Christianity. But the influence of biblical literature neither began nor ended with doctrinal propositions or codes of behaviour. Its importance lies not merely in its overtly religious influence but also, and perhaps more decisively, in its pervasive effect on the thinking and feeling processes, the attitudes and sense of values that, whether recognized as biblical or not, still help to make people what they are.

Influence on the arts

The deepest influence of biblical literature may be found in the arts of Western people, their music and, especially, in their best poetry, drama, and creative fiction. Many of the most moving and illuminating interpretations of biblical material—stories, themes, and characters—are made today by novelists, playwrights, and poets who write simply as human beings, not as adherents of any religion. There are two views of the human condition that scholars have attributed to biblical influence and that have become dominant in Western literature.

The first of these is the view that the mystery of existence and destiny is implicit in every man and woman. In contrast to the canons of classical tragedy, a person of any rank or station may experience the extremes of happiness or misery, exaltation or tragedy. An aged Jew of Rembrandt's paintings or an illiterate black woman of Faulkner's novels can reach the height of human dignity. The arts also put down the mighty from their seats and exalt those of low degree. Any man may be Everyman, the symbol of all human possibility.

The second view of the human condition is that the time of encountering all reality is now, and the place is here, in man's workaday activities and contingencies, whatever they may be. To be human is to know one short life in mortal flesh, in which the past and future are dimensions of the present. It is now or never that the choice is made, the offer of the gift of life accepted or declined. Any kingdom there is must be entered at once or lost forever. It is here in the actual situation of work and play, of love and need, and not in some far-off better time and place, that the crisis is reached and passed, the issue settled, and the record closed.

These views, though here stated in language that has theological overtones, are not confined to adherents of Judaism or Christianity. They are characteristically Western views of the human condition. That they can be put in words reminiscent of the Bible indicates that the representation of man in Western literature is indeed conditioned by biblical literature. (H.G.D.)

II. Old Testament canon, texts, and versions

THE CANON

The term canon, from a Hebrew-Greek word meaning a cane or measuring rod, passed into Christian usage as a norm or a rule of faith. The Church Fathers of the 4th century CE first employed it in reference to the definitive, authoritative nature of the body of sacred Scripture.

The Hebrew canon. The Hebrew Bible is often known among Jews as TaNaKh, an acronym derived from the names of its three divisions: Torah (Instruction, or Law, the Pentateuch), Nevi'im (the Prophets), and Ketuvim (the Writings).

Divisions of the Hebrew Bible

The Torah contains five books: Genesis, Exodus, Leviticus, Numbers, and Deuteronomy. The Nevi'im comprise eight books subdivided into the Former Prophets, containing the four historical works Joshua, Judges, Samuel, and Kings, and the Latter Prophets, the oracular discourses of Isaiah, Jeremiah, Ezekiel, and the Twelve (Minor—*i.e.*, smaller) Prophets—Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, and Malachi. The Twelve were all formerly written on a single scroll and thus reckoned as one book. The Ketuvim consist of religious poetry and wisdom literature—Psalms, Proverbs, and Job, a collection known as the "Five Megillot" ("scrolls"; *i.e.*, Song of Songs, Ruth, Lamentations, Ecclesiastes, and Esther, which have been grouped together according to the annual cycle of their public reading in the synagogue)—and the books of Daniel, Ezra and Nehemiah, and Chronicles.

The number of books. The number of books in the Hebrew canon is thus 24, referring to the sum of the separate scrolls on which these works were traditionally written in ancient times. This figure is first cited in II

Esdras in a passage usually dated *c.* 100 CE and is frequently mentioned in rabbinic (postbiblical) literature, but no authentic tradition exists to explain it. Josephus, a 1st century CE Jewish historian, and some of the Church Fathers, such as Origen (the great 3rd-century Alexandrian theologian), appear to have had a 22-book canon.

English Bibles list 39 books for the Old Testament because of the practice of bisecting Samuel, Kings, and Chronicles, and of counting Ezra, Nehemiah, and the 12 Minor Prophets as separate books.

The tripartite canon. The threefold nature of the Hebrew Bible (the Law, the Prophets, and the Writings) is reflected in the literature of the period of the Second Temple (6th–1st centuries BCE) and soon after it. The earliest reference is that of the Jewish wisdom writer Ben Sirach (*c.* 180 BCE), who speaks of "the law of the Most High . . . the wisdom of all the ancients and . . . prophecies." His grandson (*c.* 132 BCE) in the prologue to Ben Sirach's work mentions "the law and the prophets and the others that followed them," the latter also called "the other books of our fathers." The same tripartite division finds expression in II Maccabees, the writings of Philo, a Hellenistic Jewish philosopher, and Josephus, a Hellenistic Jewish historian, as well as in the Gospel According to Luke. The tripartite canon represents the three historic stages in the growth of the canon.

Criteria of canonicity

The history of canonization. Because no explicit or reliable traditions concerning the criteria of canonicity, the canonizing authorities, the periods in which they lived, or the procedure adopted have been preserved, no more than a plausible reconstruction of the successive stages involved can be provided. First, it must be observed that sanctity and canonization are not synonymous terms. The first condition must have existed before the second could have been formally conferred. Next, the collection and organization of a number of sacred texts into a canonized corpus (body of writings) is quite a different problem from that of the growth and formation of the individual books themselves.

No longer are there compelling reasons to assume that the history of the canon must have commenced very late in Israel's history, as was once accepted. The emergence in Mesopotamia, already in the second half of the 2nd millennium BCE, of a standardized body of literature arranged in a more or less fixed order and with some kind of official text, expresses the notion of a canon in its secular sense. Because Babylonian and Assyrian patterns frequently served as the models for imitation throughout the Near East, sacred documents in Israel may well have been carefully stored in temples and palaces, particularly if they were used in connection with the cult or studied in the priestly or wisdom schools. The injunction to deposit the two tables of the Decalogue (Ten Commandments) inside the ark of the covenant and the book of the Torah beside it and the chance find of a book of the Torah in the Temple in 622 BCE tend to confirm the existence of such a practice in Israel.

The Torah. The history of the canonization of the Torah as a book must be distinguished from the process by which the heterogeneous components of the literature as such developed and were accepted as sacred.

The Book of the Chronicles, composed *c.* 400 BCE, frequently refers to the "Torah of Moses" and exhibits a familiarity with all the five books of the Pentateuch. The earliest record of the reading of a "Torah book" is provided by the narrative describing the reformation instituted by King Josiah of Judah in 622 BCE following the fortuitous discovery of a "book of the Torah" during the renovation of the Temple. The reading of the book (probably Deuteronomy), followed by a national covenant ceremony, is generally interpreted as having constituted a formal act of canonization.

Between this date and 400 BCE the only other ceremony of Torah reading is that described in Nehemiah as having taken place on the autumnal New Year festival. The "book of the Torah of Moses" is mentioned and the emphasis is on its instruction and exposition. The Samaritans, the descendants of Israelites intermarried with foreigners in the old northern kingdom that fell in 722 BCE, became hostile to the Judaeans in the time of Ezra and Nehemiah (6th–5th centuries BCE). They would not likely have accepted the Torah, which they did, along with the tradition of its Mosaic origin, if it had only recently been canonized under the authority of their arch-enemies. The final redaction and canonization of the Torah book, therefore, most likely took place during the Babylonian Exile (6th–5th centuries BCE).

The Nevi'im. The model of the Pentateuch probably encouraged the assemblage and ordering of the literature of the prophets. The Exile of the Jews to Babylonia in 587/586 and the restoration half a century later enhanced the prestige of the prophets as national figures and aroused interest in the written records of their teachings. The canonization of the Nevi'im could not have taken place before the Samaritan schism that occurred during the time of Ezra and Nehemiah, since nothing of the prophetical literature was known to the Samaritans. On the other hand, the prophetic canon must have been closed by the time the Greeks had displaced the Persians as the rulers of Palestine in the late 4th century BCE. The exclusion of Daniel would otherwise be inexplicable, as would also the omission of Chronicles and Ezra–Nehemiah, even though they supplement and continue the narrative of the Former Prophets. Furthermore, the books of the Latter Prophets contain no hint of the downfall of the Persian Empire and the rise of the Greeks, even though the succession of great powers in the East plays a major role in their theological interpretation of history. Their language, too, is entirely free of Grecisms.

Canonization of prophetic writings

These phenomena accord with the traditions of Josephus and rabbinic sources limiting the activities of the literary prophets to the Persian era.

The Ketuvim. That the formation of the Ketuvim as a corpus was not completed until a very late date is evidenced by the absence of a fixed name, or indeed any real name, for the third division of Scripture. Ben Sirach refers to "the other books of our fathers," "the rest of the books"; Philo speaks simply of "other writings" and Josephus of "the remaining books." A widespread practice of entitling the entire Scriptures "the Torah and the Prophets" indicates a considerable hiatus between the canonization of the Prophets and the Ketuvim. Greek words are to be found in the Song of Songs and in Daniel, which also refers to the disintegration of the Greek Empire. Ben Sirach omits mention of Daniel and Esther. No fragments of Esther have turned up among the biblical scrolls (*e.g.*, the Dead Sea Scrolls) from the Judaean Desert. Rabbinic sources betray some hesitation about Esther and a decided ambivalence about the book of Ben Sirach. A third generation Babylonian *amora* (rabbinical interpretive scholar; pl. *amoraim*) actually cites it as "Ketuvim," as opposed to Torah and Prophets, and in the mid-2nd century CE, the need to deny its canonicity and prohibit its reading was still felt. Differences of opinion also are recorded among the *tannaim* (rabbinical scholars of tradition who compiled the Mishna, or Oral Law) and *amoraim* (who created the Talmud, or Gemara) about the canonical status of Proverbs, Song of Songs, Ecclesiastes, and Esther.

All this indicates a prolonged state of fluidity in respect of the canonization of the Ketuvim. A synod at Jabneh (*c.* 100 CE) seems to have ruled on the matter, but it took a generation or two before their decisions came to be unanimously accepted and the Ketuvim regarded as being definitively closed. The destruction of the Jewish state in 70 CE, the breakdown of central authority, and the ever widening Diaspora (collectively, Jews dispersed to foreign lands) all contributed to the urgent necessity of providing a closed and authoritative corpus of sacred Scriptures.

The Samaritan canon. As has been mentioned, the Samaritans accepted the Pentateuch from the Jews. They know of no other section of the Bible, however, and did not expand their Pentateuchal canon even by the inclusion of any strictly Samaritan compositions.

The Alexandrian canon. The Old Testament as it has come down in Greek translation from the Jews of Alex-

andria via the Christian Church differs in many respects from the Hebrew Scriptures. The books of the second and third divisions have been redistributed and arranged according to categories of literature—history, poetry, wisdom, and prophecy. Esther and Daniel contain supplementary materials, and many noncanonical books, whether of Hebrew or Greek origin, have been interspersed with the canonical works. These extracanonical writings comprise I Esdras, the Wisdom of Solomon, Ecclesiasticus (Ben Sirach), Additions to Esther, Judith, Tobit, Baruch, the Epistle of Jeremiah, and additions to Daniel, as listed in the manuscript known as Codex Vaticanus (*c.* 350 CE). The sequence of the books varies, however, in the manuscripts and in the patristic and synodic lists of the Eastern and Western churches, some of which include other books as well, such as I and II Maccabees.

It should be noted that the contents and form of the inferred original Alexandrian Jewish canon cannot be ascertained with certainty because all extant Greek Bibles are of Christian origin. The Jews of Alexandria may themselves have extended the canon they received from Palestine, or they may have inherited their traditions from Palestinian circles in which the additional books had already been regarded as canonical. It is equally possible that the additions to the Hebrew Scriptures in the Greek Bible are of Christian origin.

The canon at Qumrān. In the collection of manuscripts from the Judaean Desert—discovered from the 1940s on—there are no lists of canonical works and no codices (manuscript volumes), only individual scrolls. For these reasons nothing can be known with certainty about the contents and sequence of the canon of the Qumrān sectarians. Since fragments of all the books of the Hebrew Bible (except Esther) have been found, it may be assumed that this reflects the minimum extent of its canon. The situation is complicated by the presence in Qumrān of extracanonical works—some already known from the Apocrypha (so-called hidden books not accepted as canonical by Judaism and the church) and pseudepigrapha (books falsely ascribed to biblical authors) or from the Cairo Geniza (synagogue store room), and others entirely new. Some or all of these additional works may have been considered canonical by the members of the sect. It is significant, however, that so far *pesharim* (interpretations) have been found only on books of the traditional Hebrew canon. Still, the great Psalms scroll departs from the received Hebrew text in both sequence and contents. If the Psalms scroll were a canonical Psalter and not a liturgy, then evidence would indeed be forthcoming for the existence of a rival canon at Qumrān.

Canonical and extracanonical writings

The Christian canon. The Christian Church received its Bible from Greek-speaking Jews and found the majority of its early converts in the Hellenistic world. The Greek Bible of Alexandria thus became the official Bible of the Christian community, and the overwhelming number of quotations from the Hebrew Scriptures in the New Testament are derived from it. Whatever the origin of the Apocryphal books in the canon of Alexandria, these became part of the Christian Scriptures, but there seems to have been no unanimity as to their exact canonical status. The New Testament itself does not cite the Apocryphal books directly, but occasional traces of a knowledge of them are to be found. The Apostolic Fathers (late 1st–early 2nd centuries) show extensive familiarity with this literature, but a list of the Old Testament books by Melito, bishop of Sardis in Asia Minor (2nd century), does not include the additional writings of the Greek Bible, and Origen (*c.* 185–*c.* 254) explicitly describes the Old Testament canon as comprising only 22 books.

From the time of Origen on, the Church Fathers who were familiar with Hebrew differentiated, theoretically at least, the Apocryphal books from those of the Old Testament, though they used them freely. In the Syrian East, until the 7th century the Church had only the books of the Hebrew canon with the addition of Ecclesiasticus, or the Wisdom of Jesus the son of Sirach (but without Chronicles, Ezra, and Nehemiah). It also incorporated the Wisdom of Solomon, Baruch, the Letter of Jeremiah, and the additions to Daniel. The 6th-century manuscript of the Peshitta known as Codex Ambrosianus also has III and IV Maccabees, II (sometimes IV) Esdras, and Josephus' *Wars* VII.

Early councils of the African Church held at Hippo (393) and Carthage (397, 419) affirmed the use of the Apocryphal books as Scripture. In the 4th century also, Athanasius, chief theologian of Christian orthodoxy, differentiated "canonical books" from both "those that are read" by Christians only and the "Apocryphal books" rejected alike by Jews and Christians. In the preparation of a standard Latin version, the biblical scholar Jerome (*c.* 347–419/420) separated "canonical books" from "ecclesiastical books" (*i.e.*, the Apocryphal writings), which he regarded as good for spiritual edification but not authoritative Scripture. A contrary view of Augustine (354–430), one of the greatest Western theologians, prevailed, however, and the works remained in the Latin Vulgate version. The *Decretum Gelasianum*, a Latin document of uncertain authorship but recognized as reflecting the views of the Roman Church at the beginning of the 6th century, includes Tobit, Judith, the Wisdom of Solomon, Ecclesiasticus, and I and II Maccabees as biblical.

Throughout the Middle Ages, the Apocryphal books were generally regarded as Holy Scripture in the Roman and Greek churches, although theoretical doubts were raised from time to time. Thus, in 1333 Nicholas of Lyra, a French Franciscan theologian, had discussed the differences between the Latin Vulgate and the "Hebrew truth." Christian–Jewish polemics, the increasing attention to Hebrew studies, and, finally, the Reformation kept the issue of the Christian canon alive. Protestants denied canonical status to all books not in the Hebrew Bible. The first modern vernacular Bible to segregate the disputed writings was a Dutch version by Jacob van Liesveldt (Antwerp, 1526). Luther's German edition of 1534 did the same thing and entitled them "Apocrypha" for the first time, noting that while they were not in equal esteem with sacred Scriptures they were edifying.

Use of Apocryphal works in the Middle Ages

In response to Protestant views, the Roman Catholic Church made its position clear at the Council of Trent (1546) when it dogmatically affirmed that the entire Latin Vulgate enjoyed equal canonical status. This doctrine was confirmed by the Vatican Council of 1870. In the Greek Church, the Synod of Jerusalem (1672) had expressly designated as canonical several Apocryphal works. In the 19th century, however, Russian Orthodox theologians agreed to exclude these works from the Holy Scriptures.

The history of the Old Testament canon in the English Church has generally reflected a more restrictive viewpoint. Even though the Wycliffite Bible (14th century) included the Apocrypha, its preface made it clear that it accepted Jerome's judgment. The translation made by the English bishop Miles Coverdale (1535) was the first English version to segregate these books, but it did place Baruch after Jeremiah. Article VI of the Thirty-nine Articles of religion of the Church of England (1562) explicitly denied their value for the establishment of doctrine, although it admitted that they should be read for their didactic worth. The first Bible in English to exclude the Apocrypha was the Geneva Bible of 1599. The King James Version of 1611 placed it between the Old and New Testaments. In 1615 Archbishop George Abbot forbade the issuance of Bibles without the Apocrypha, but editions of the King James Version from 1630 on often omitted it from the bound copies. The Geneva Bible edition of 1640 was probably the first to be intentionally printed in England without the Apocrypha, followed in 1642 by the King James Version. In 1644 the Long Parliament actually forbade the public reading of these books, and three years later the Westminster Confession of the Presbyterians decreed them to be no part of the canon. The British and Foreign Bible Society in 1827 resolved never to print or circulate copies containing the Apocrypha. Most English Protestant Bibles in the 20th century have omitted the disputed books or have them as a separate volume, except in library editions, in which they are included with the Old and New Testaments.

TEXTS AND VERSIONS

Textual criticism: manuscript problems. The text of the Hebrew printed Bible consists of consonants, vowel signs, and cantillation (musical or tonal) marks. The two latter components are the product of the school of Masoretes (Traditionalists) that flourished in Tiberias (in Palestine) between the 7th and 9th centuries CE. The history of the bare consonantal text stretches back into hoary antiquity and can be only partially traced.

Masoretic signs and marks

The earliest printed editions of the Hebrew Bible derive from the last quarter of the 15th century and the first quarter of the 16th century. The oldest Masoretic codices stem from the end of the 9th century and the beginning of the 10th. A comparison of the two shows that no textual developments took place during the intervening 600 years. A single standardized recension enjoyed an absolute monopoly and was transmitted by the scribes with amazing fidelity. Not one of the medieval Hebrew manuscripts and none of the thousands of fragments preserved in the Cairo Geniza (synagogue storeroom) contains departures of any real significance from the received text.

This situation, however, was a relatively late development; there is much evidence for the existence of a period when more than one Hebrew text-form of a given book was current. In fact, both the variety of witnesses and the degree of textual divergence between them increase in proportion to their antiquity.

No single explanation can satisfactorily account for this phenomenon. In the case of some biblical literature, there exists the real possibility, though it cannot be proven, that it must have endured a long period of oral transmission before its committal to writing. In the interval, the material might well have undergone abridgement, amplification, and alteration at the hands of transmitters so that not only would the original have been transformed, but the process of transmission would have engendered more than one recension from the very beginning of its written, literary career.

The problem is complicated further by the great difference in time between the autograph (original writing) of a biblical work, even when it assumed written form from its inception, and its oldest extant exemplars. In some instances, this may amount to well over a thousand years of scribal activity. Whatever the interval, the possibility of inadvertent and deliberate change, something that affects all manuscript copying, was always present.

The evidence that such, indeed, took place is rich and varied. First there are numerous divergences between the many passages duplicated within the Hebrew Bible itself—*e.g.*, the parallels between Samuel-Kings and Chronicles. Then there are the citations of the Old Testament to be found in the books of the Apocrypha and apocalyptic literature (works describing the intervention of God in history in cryptic terms), in the works of Philo and Josephus, in the New Testament, and in rabbinic and patristic (early Church Fathers) literature. There are also rabbinic traditions about the text-critical activities of the scribes (*soferim*) in Second Temple times. These tell of divergent readings in Temple scrolls of the Pentateuch, of official "book correctors" in Jerusalem, of textual emendations on the part of scribes, and of the utilization of sigla (signs or abbreviations) for marking suspect readings and disarranged verses. The Samaritan Pentateuch and the pre-Masoretic versions of the Old Testament made directly from Hebrew originals are all replete with divergences from current Masoretic Bibles. Finally, the scrolls from the Judaean Desert, especially those from the caves of Qumrān, have provided, at least, illustrations of many of the scribal processes by which deviant texts came into being. The variants and their respective causes may be classified as follows: aurally conditioned, visual in origin, exegetical, and deliberate.

Types and causes of variants

Problems resulting from aural conditioning. Aural conditioning would result from a mishearing of similar sounding consonants when a text is dictated to the copyist. A negative particle *lo'*, for example, could be confused with the prepositional *lo*, "to him," or a guttural *ḥet* with spirant *kaf* so that *aḥ* "brother" might be written for *akh* "surely."

Problems visual in origin. The confusion of graphically similar letters, whether in the paleo-Hebrew or Aramaic script, is another cause for variations. Thus, the prepositions *bet* ("in") and *kaf* ("like") are interchanged in the Masoretic and Dead Sea Scroll texts of Isaiah.

The order of letters also might be inverted. Such metathesis, as it is called, appears in Psalms, in which *qirbam* ("their inward thoughts") stands for *qibram* ("their grave").

Dittography, or the inadvertent duplication of one or more letters or words, also occurs, as, for example, in the Dead Sea Scroll text of Isaiah and in the Masoretic text of Ezekiel.

Haplography, or the accidental omission of a letter or word that occurs twice in close proximity, can be found, for example, in the Dead Sea Scroll text of Isaiah.

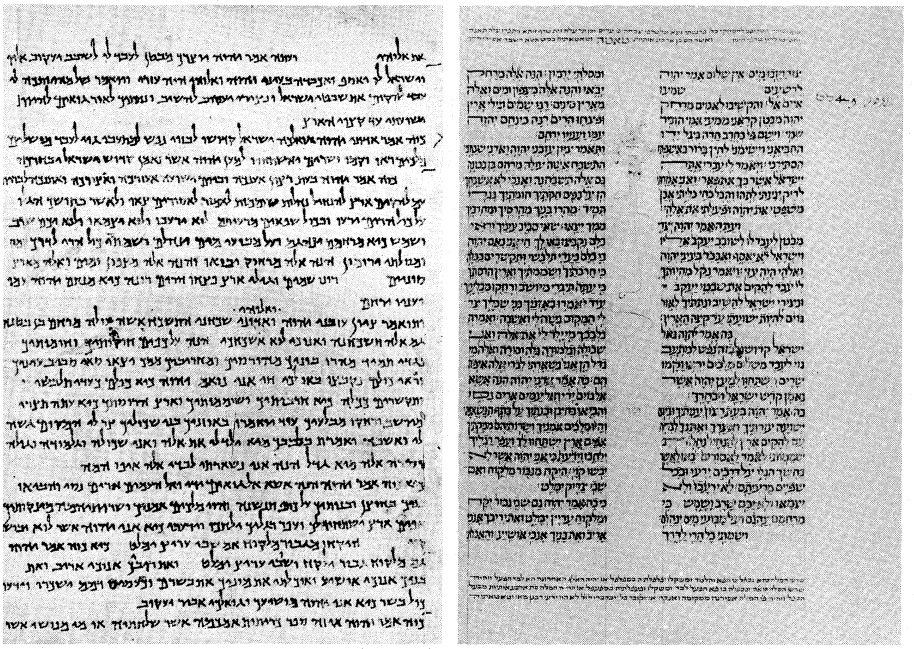
Homoeoteleuton occurs when two separate phrases or lines have identical endings and the copyist's eye slips from one to the other and omits the intervening words. A comparison of the Masoretic text I Samuel, chapter 14 verse 41 with the Septuagint and the Vulgate versions clearly identifies such an aberration.

Exegetical problems. This third category does not involve any consonantal alteration but results solely from the different possibilities inherent in the consonantal spelling. Thus, the lack of vowel signs may permit the word *DBR* to be read as a verb *DiBeR* ("he spoke," as in the Masoretic text of Hosea) or as a noun *DeBaR* ("the word of," as in the Septuagint). The absence of word dividers could lead to different divisions of the consonants. Thus, *BBQRYM* in Amos could be understood as either *BaBeQaRYM* ("with oxen," as in the Masoretic text) or as *BaBaQaR YaM* ("the sea with an ox"). The incorrect solution by later copyists of abbreviations is another source of error. That such occurred is proved by a comparison of the Hebrew text with the Septuagint version in, for example, II Samuel, chapter 1 verse 12; Ezekiel, chapter 12 verse 23; and Amos, chapter 3 verse 9.

Deliberate changes. Apart from mechanical alterations of a text, many variants must have been consciously introduced by scribes, some by way of glossing—*i.e.*, the insertion of a more common word to explain a rare one—and others by explanatory comments incorporated into the text. Furthermore, a scribe who had before him two manuscripts of a single work containing variant readings, and unable to decide between them, might incorporate both readings into his scroll and thus create a conflate text.

Textual criticism: scholarly problems. The situation so far described poses two major scholarly problems. The first involves the history of the Hebrew text, the second deals with attempts to reconstruct its "original" form.

As to when and how a single text-type gained hegemony and then displaced all others, it is clear that the early and widespread public reading of the Scriptures in the synagogues of Palestine, Alexandria, and Babylon was bound to lead to a heightened sensitivity of the idea of a "correct" text and to give prestige to the particular text form selected for reading. Also, the natural conservatism of ritual would tend to perpetuate the form of such a text. The *Letter of Aristeas*, a document derived from the middle of the 2nd century BCE that describes the origin of the Septuagint, recognizes the distinction between carelessly copied scrolls of the Pentateuch and an authoritative Temple scroll in the hands of the high priest in Jerusalem. The Rabbinic traditions (see above) about the textual criticism of Temple-based scribes actually reflect a movement towards the final stabilization of the text in the Second Temple period. Josephus, writing not long after 70 CE, boasts of the existence of a long-standing fixed text of the Jewish Scriptures. The loss of national independence and the destruction of the spiritual centre of Jewry in 70, accompanied by an ever-widening Diaspora and the Christian schism within Judaism, all made the exclusive dissemination of a single authoritative text a vitally needed cohesive force. The text type later known as Masoretic is already well represented at pre-Christian Qumrān. Scrolls from Wādī al-Murab-



(Left) Chapter 49 of the Isaiah Scroll from the Dead Sea Scrolls. In the Shrine of the Book, D. Samuel and Jeane H. Gottesman Centre for Biblical Manuscripts, the Israel Museum, Jerusalem. (Right) Hebrew Masoretic text of chapter 49 from the Book of Isaiah, 1341. In the Jewish National and University Library, Jerusalem.
By courtesy of the (left) Israel Museum, Jerusalem, (right) Jewish National and University Library, Jerusalem

ba'at, Nahal Ze'elim, and Masada from the 2nd century CE are practically identical with the received text that by then had gained overwhelming victory over all its rivals.

Reconstruction of an original text

In regard to an attempt to recover the original text of a biblical passage—especially an unintelligible one—in the light of variants among different versions and manuscripts and known causes of corruption, it should be understood that all reconstruction must necessarily be conjectural and perforce tentative because of the irretrievable loss of the original edition. In the first place, not all textual difficulties need presuppose underlying mutilation. The Hebrew Bible represents but a small portion of the literature of ancient Israel and, hence, a limited segment of the language. A textual problem may be the product of present limited knowledge of ancient Hebrew, because scholars might be dealing with dialectic phenomena or foreign loan-words. Comparative Semitic linguistic studies have yielded hitherto unrecognized features of grammar, syntax, and lexicography that have often eliminated the need for emendation. Furthermore, each version, indeed each biblical book within it, has its own history, and the translation techniques and stylistic characteristics have to be examined and taken into account. Finally, the number of manuscripts that attest to a certain reading is of less importance than the weight to be given to a specific manuscript.

None of this means that a Hebrew manuscript, an ancient version, or a conjectural emendation cannot yield a reading superior to that in the received Hebrew text. It does mean, however, that these tools have to be employed with great caution and proper methodology.

Texts and manuscripts. *Sources of the Septuagint.* A Greek translation of the Old Testament, known as the Septuagint because there allegedly were 70 or 72 translators, six from each of the 12 tribes of Israel, and designated LXX, is a composite of the work of many translators labouring for well over 100 years. It was made directly from Hebrew originals that frequently differed considerably from the present Masoretic text. Apart from other limitations attendant upon the use of a translation for such purposes, the identification of the parent text used by the Greek translators is still an unsettled question. The Pentateuch of the Septuagint manifests a basic coincidence with the Masoretic text. The Qumrān scrolls have now proven that the Septuagint book of Samuel-Kings goes back to an old Palestinian text tradition that must be earlier than the 4th century BCE, and from the same source comes a short Hebrew recension of Jeremiah that probably underlies the Greek.

The Samaritan Pentateuch. The importance of the recension known as the Samaritan Pentateuch lies in the fact that it constitutes an independent Hebrew witness to the text written in a late and developed form of the paleo-Hebrew script. Some of the Exodus fragments from Qumrān demonstrate that it has close affinities with a pre-Christian Palestinian text type and testify to the faithfulness with which it has been preserved. It contains about 6,000 variants from the Masoretic text, of which nearly a third agree with the Septuagint. Only a minority, however, are genuine variants, most being dogmatic, exegetical, grammatical, or merely orthographic in character.

The Samaritan Pentateuch first became known in the West through a manuscript secured in Damascus in 1616 by Pietro della Valle, an Italian traveler. It was published in the Paris (1632) and London Polyglots (1654–57), written in several languages in comparative columns. Many manuscripts of the Samaritan Pentateuch are now available. The Avisha' Scroll, the sacred copy of the Samaritans, has recently been photographed and critically examined. Only Numbers chapter 35 to Deuteronomy chapter 34 appears to be very old, the rest stemming from the 14th century. A new, definitive edition of the Samaritan Pentateuch is being prepared in Madrid by F. Pérez Castro.

The Qumrān texts and other scrolls. Until the discovery of the Judaean Desert scrolls, the only pre-medieval fragment of the Hebrew Bible known to scholars was the Nash Papyrus (*c.* 150 BCE) from Egypt containing the Decalogue and Deuteronomy. Now, however, fragments of about 180 different manuscripts of biblical books are available. Their dates vary between the 3rd century BCE and the 2nd century CE, and all but ten stem from the caves of Qumrān. All are written on either leather or papyrus in columns and on one side only.

Textual significance of the Qumrān scrolls

The most important manuscripts from what is now identified as Cave 1 of Qumrān are a practically complete Isaiah scroll (1QIsa^{a}), dated *c.* 100–75 BCE, and another very fragmentary manuscript (1QIsa^{b}) of the same book. The first contains many variants from the Masoretic text in both orthography and text; the second is very close to the Masoretic type and contains few genuine variants.

The richest hoard comes from Cave 4 and includes fragments of five copies of Genesis, eight of Exodus, one of Leviticus, 14 of Deuteronomy, two of Joshua, three of Samuel, 12 of Isaiah, four of Jeremiah, eight of the Minor Prophets, one of Proverbs, and three of Daniel. Cave 11 yielded a Psalter containing the last third of the book in a form different from that of the Masoretic text, as well as a manuscript of Leviticus.

The importance of the Qumrān scrolls cannot be exaggerated. Their great antiquity brings them close to the Old Testament period itself—from as early as 250–200 BCE. For the first time, Hebrew variant texts are extant and all known major text types are present. Some are close to the Septuagint, others to the Samaritan. On the other hand, many of the scrolls are practically identical with the Masoretic text, which thus takes this recension back in history to pre-Christian times. Several texts in the paleo-Hebrew script show that this script continued to be used side by side with the Aramaic script for a long time.

Of quite a different order are scrolls from other areas of the Judaean Desert. All of these are practically identical with the received text. This applies to fragments of Leviticus, Deuteronomy, Ezekiel, and Psalms discovered at Masada (the Jewish fortress destroyed by the Romans in CE 73), as well as to the finds at Wādī al-Murabba'at, the latest date of which is CE 135. Here were found fragments of Genesis, Exodus, Leviticus, and Isaiah in addition to the substantially preserved Minor Prophets scroll. Variants from the Masoretic text are negligible. The same phenomenon characterizes the fragments of Numbers found at Nahal Ḥever.

Masoretic texts. No biblical manuscripts have survived from the six centuries that separate the latest of the Judaean Desert scrolls from the earliest of the Masoretic period. A "Codex Mugah," frequently referred to as an authority in the early 10th century, and the "Codex Hilleli," said to have been written *c.* 600 by Rabbi Hillel ben Moses ben Hillel, have both vanished.

The earliest extant Hebrew Bible codex is the Cairo Prophets written and punctuated by Moses ben Asher in Tiberias (in Palestine) in 895. Next in age is the Leningrad Codex of the Latter Prophets dated to 916, which was not originally the work of Ben Asher, but its Babylonian pointing—*i.e.*, vowel signs used for pronunciation purposes—was brought into line with the Tiberian Masoretic system.

Production of the Aleppo Codex

The outstanding event in the history of that system was the production of the model so-called Aleppo Codex, now in Jerusalem. Written by Solomon ben Buya'a, it was corrected, punctuated, and furnished with a Masoretic apparatus by Aaron ben Moses ben Asher *c.* 930. Originally containing the entire Old Testament in about 380 folios, of which 294 are extant, the Aleppo Codex remains the only known true representative of Aaron ben Asher's text and the most important witness to that particular Masoretic tradition which achieved hegemony throughout Jewry.

Two other notable manuscripts based on Aaron's system are the manuscript designated as BM or. 4445, which contains most of the Pentateuch and which utilized a Masora (text tradition) *c.* 950, and the Leningrad complete Old Testament designated MSB 19a of 1008. Codex Reuchliana of the Prophets, written in 1105, now in Karlsruhe (Germany), represents the system of Moses ben David ben Naphtali, which was more faithful to that of Moses ben Asher.

Collations of the Masoretic materials. The earliest extant attempt at collating the differences between the Ben Asher and Ben Naphtali Masoretic traditions was made by Mishael ben Uzziel in his *Kitāb al-Ḫulaf* (before 1050). A vast amount of Masoretic information, drawn chiefly from Spanish manuscripts, is to be found in the text critical commentary known as *Minhath Shai*, by Solomon Jedidiah Norzi, completed in 1626 and printed in the Mantua Bible of 1742. Benjamin Kennicott collected the variants of 615 manuscripts and 52 printed editions (2 vol., 1776–80, Oxford). Giovanni Bernado De Rossi published his additional collections of 731 manuscripts and 300 prints (4 vol., 1784–88, Parma), and C.D. Ginsburg did the same for 70 manuscripts, largely from the British Museum, and 17 early printed editions (3 vol. in 4, 1908–26, London).

Printed editions. Until 1488, only separate parts of the Hebrew Bible had been printed, all with rabbinic commentaries. The earliest was the Psalms (1477), followed by the Pentateuch (1482), the Prophets (1485/86), and the Hagiographa (1486/87), all printed in Italy.

The first edition of the entire Hebrew Bible was printed at Soncino (in Italy) in 1488 with punctuation and accents, but without any commentary. The second complete Bible was printed in Naples in 1491/93 and the third in Brescia in 1494. All these editions were the work of Jews. The first Christian production was a magnificent Complutensian Polyglot (under the direction of Cardinal Ximenes of Spain) in six volumes, four of which contained the Hebrew Bible and Greek and Latin translations together with the Aramaic rendering (Targum) of the Pentateuch that has been ascribed to Onkelos. Printed at Alcala (1514–17) and circulated in 1522, this Bible proved to be a turning point in the study of the Hebrew text in western Europe.

The first rabbinic Bible—*i.e.*, the Hebrew text furnished with full vowel points and accents, accompanied by the Aramaic Targums and the major medieval Jewish commentaries—was edited by Felix Pratensis and published by Daniel Bomberg (Venice, 1516/17). The second edition, edited by Jacob ben Hayim ibn Adonijah and issued by Bomberg in four volumes (Venice, 1524/25), became the prototype of future Hebrew Bibles down to the 20th century. It contained a vast text-critical apparatus of Masoretic notes never since equaled in any edition. Unfortunately, Ben Hayyim had made use of late manuscripts and the text and notes are eclectic.

In London, C.D. Ginsburg produced a critical edition of the complete Hebrew Bible (1894, 1908, 1926) revised according to the Masora and early prints with variant readings from manuscripts and ancient versions. It was soon displaced by the *Biblica Hebraica* (1906, 1912) by Rudolf Kittel and Paul Kahle, two German biblical scholars. The third edition of this work, completed by Albrecht Alt and Otto Eissfeldt (Stuttgart, 1937), finally abandoned Ben Hayyim's text, substituting that of the Leningrad Codex (B 19a). It has a dual critical apparatus with textual emendations separated from the manuscript and versional variants. Since 1957 variants from the so-called Judaean Desert scrolls have been included. In progress at the Hebrew University of Jerusalem in the early 1970s was the preparation of a new text of the entire Hebrew Bible based on the Aleppo Codex to include all its own Masoretic notes together with textual differences found in all pertinent sources. A sample edition of the Book of Isaiah appeared in 1965.

Early versions. *The Aramaic Targums.* In the course of the 5th and 6th centuries BCE, Aramaic became the official language of the Persian Empire. In the succeeding centuries it was used as the vernacular over a wide area and was increasingly spoken by the postexilic Jewish communities of Palestine and elsewhere in the Diaspora. In response to liturgical needs, the institution of a *turgeman* (or *meturgeman*, "translator"), arose in the synagogues. These men translated the Torah and prophetic lectionaries into Aramaic. The rendering remained for long solely an oral, impromptu exercise, but gradually, by dint of repetition, certain verbal forms and phrases became fixed and eventually committed to writing.

The Babylonian and Palestinian Targums

There are several Targums (translations) of the Pentateuch. The Babylonian Targum is known as "Onkelos," named after its reputed author. The Targum is Palestinian in origin, but it was early transferred to Babylon where it was revised and achieved great authority. At a later date, probably not before the 9th century CE, it was re-exported to Palestine to displace other, local, Targums. On the whole, Onkelos is quite literal, but it shows a tendency to obscure expressions attributing human form and feelings to God. It also usually faithfully reflects rabbinic exegesis.

The most famous of the Palestinian Targums is that popularly known as "Jonathan," a name derived from a

14th-century scribal mistake that solved a manuscript abbreviation "TJ" as "Targum Jonathan" instead of "Targum Jerusalem." In contrast with two other Targums, which are highly fragmentary (Jerusalem II and III), Pseudo-Jonathan (or Jerusalem I) is virtually complete. It is a composite of the Old Palestinian Targum and an early version of Onkelos with an admixture of material from diverse periods. It contains much rabbinic material as well as homiletic and didactic amplifications. There is evidence of great antiquity, but also much late material, indicating that Pseudo-Jonathan could not have received its present form before the Islāmic period.

Another extant Aramaic version is the Targum to the Samaritan Pentateuch. It is less literal than the Jewish Targums and its text was never officially fixed.

The Targum to the Prophets also originated in Palestine and received its final editing in Babylonia. It is ascribed to Jonathan ben Uzziel, a pupil of Hillel, the famous 1st century BCE–1st century CE rabbinic sage, though it is in fact a composite work of varying ages. In its present form it discloses a dependence on Onkelos, though it is less literal.

The Aramaic renderings of the Hagiographa are relatively late productions, none of them antedating the 5th century CE.

The Septuagint (LXX). The story of the Greek translation of the Pentateuch is told in the *Letter of Aristeas*, which purports to be a contemporary document written by Aristeas, a Greek official at the Egyptian court of Ptolemy II Philadelphus (285–246 BCE). It recounts how the law of the Jews was translated into Greek by Jewish scholars sent from Jerusalem at the request of the king.

This narrative, repeated in one form or another by Philo and rabbinic sources, is full of inaccuracies that prove that the author was an Alexandrian Jew writing well after the events he described had taken place. The Septuagint Pentateuch, which is all that is discussed, does, however, constitute an independent corpus within the Greek Bible, and it was probably first translated as a unit by a company of scholars in Alexandria about the middle of the 3rd century BCE.

The Septuagint, as the entire Greek Bible came to be called, has a long and complex history and took well over a century to be completed. It is for this reason not a unified or consistent translation. The Septuagint became the instrument whereby the basic teachings of Judaism were mediated to the pagan world and it became an indispensable factor in the spread of Christianity.

Influence of the Septuagint

The adoption of the Septuagint as the Bible of the Christians naturally engendered suspicion on the part of Jews. In addition, the emergence of a single authoritative text type after the destruction of the Temple made the great differences between it and the Septuagint increasingly intolerable, and the need for a Greek translation based upon the current Hebrew text in circulation was felt.

The version of Aquila. About 130 CE, Aquila, a convert to Judaism from Pontus in Asia Minor, translated the Hebrew Bible into Greek under the supervision of Rabbi Akiba. Executed with slavish literalness, it attempted to reproduce the most minute detail of the original, even to the extent of coining derivations from Greek roots to correspond to Hebrew usage. Little of it has survived, however, except in quotations, fragments of the Hexapla (see *Origen's Hexapla*, below), and palimpsests (parchments erased and used again) from the Cairo Geniza.

The revision of Theodotion. A second revision of the Greek text was made by Theodotion (of unknown origins) late in the 2nd century, though it is not entirely clear whether it was the Septuagint or some other Greek version that underlay his revision. The new rendering was characterized by a tendency toward verbal consistency and much transliteration of Hebrew words.

The translation of Symmachus. Still another Greek translation was made toward the end of the same century by Symmachus, an otherwise unknown scholar, who made use of his predecessors. His influence was small despite the superior elegance of his work. Jerome did utilize Symmachus for his Vulgate, but other than that, his translation is known largely through fragments of the Hexapla.

Origen's Hexapla. The multiplication of versions doubtless proved to be a source of increasing confusion in the 3rd century. This situation the Alexandrian theologian Origen, working at Caesarea between 230 CE and 240, sought to remedy. In his Hexapla ("six-fold") he presented, in parallel vertical columns, the Hebrew text, the same in Greek letters and the versions of Aquila, Symmachus, the Septuagint and Theodotion in that order. In the case of some books, Psalms for instance, three more columns were added. The Hexapla serves as an important guide to Palestinian pre-Masoretic pronunciation of the language. The main interest of Origen lay in the fifth column, the Septuagint, which he edited on the basis of the Hebrew. He used the obelos (— or ÷) and asterisk (*) to mark respectively words found in the Greek text but not in the Hebrew and vice versa.

Purpose of the Hexapla

The Hexapla was a work of such magnitude that it is unlikely to have been copied as a whole. Origen himself produced an abbreviated edition, the Tetrapla, containing only the last four columns. The original manuscript of the Hexapla is known to have been extant as late as *c.* 600 CE. Today it survives only in fragments.

Manuscripts and printed editions of the Septuagint. The manuscripts are conveniently classified by papyri uncials (capital letters) and minuscules (cursive script). The papyri fragments run into the hundreds, of varying sizes and importance, ranging from the formative period of the Septuagint through the middle of the 7th century. Two pre-Christian fragments of Deuteronomy from Egypt are of outstanding significance. Although not written on papyrus but on parchment or leather, the fragments from Qumrān of Exodus, Leviticus, and Numbers, and the leather scroll of the Minor Prophets from Naḥal Ḥever from the first pre-Christian and post-Christian centuries, deserve special mention among the earliest extant. The most important papyri are those of the Chester Beatty collection, which contains parts of 11 codices preserving fragments of nine Old Testament books. Their dates vary between the 2nd and 4th centuries. During the next 300 years papyri texts multiplied rapidly, and remnants of about 200 are known.

The uncials are all codices written on vellum between the 4th and 10th centuries. The most outstanding are Vaticanus, which is an almost complete 4th-century Old Testament, Sinaiticus, of the same period but less complete, and the practically complete 5th-century Alexandrinus. These three originally contained both Testaments. Many others were partial manuscripts from the beginning. One of the most valuable of these is the Codex Marchalianus of the Prophets written in the 6th century.

Uncial and minuscul codices

The minuscule codices begin to appear in the 9th century. From the 11th to the 16th century they are the only ones found, and nearly 1,500 have been recorded.

The first printed Septuagint was that of the Complutensian Polyglot (1514–17). Since it was not released until 1522, however, the 1518 Aldine Venice edition actually was available first. The standard edition until modern times was that of Pope Sixtus V, 1587. In the 19th and 20th centuries several critical editions have been printed.

Coptic versions. The spread of Christianity among the non-Greek speaking peasant communities of Egypt necessitated the translation of the Scriptures into the native tongue (Coptic). These versions may be considered to be wholly Christian in origin and largely based on the Greek Bible. They also display certain affinities with the Old Latin. Nothing certain is known about the Coptic translations except that they probably antedate the earliest known manuscripts from the end of the 3rd and the beginning of the 4th centuries CE.

The Armenian version. The Armenian version is an expression of a nationalist movement that brought about a separation from the rest of the Church (mid-5th century), the discontinuance of Syriac in Greek worship, and the invention of a national alphabet by St. Mesrob, also called Mashtots (*c.* 361–439/440). According to tradition, St. Mesrob first translated Proverbs from the Syri-

ac. Existing manuscripts of the official Armenian recension, however, are based on the Hexaplaric Septuagint, though they show some Peshitta (Syriac version) influence. The Armenian Bible is noted for its beauty and accuracy.

The Georgian version. According to Armenian tradition, the Georgian version was also the work of Mesrob, but the Psalter, the oldest part of the Georgian Old Testament, is probably not earlier than the 5th century. Some manuscripts were based upon Greek versions, others upon the Armenian.

Special problems of the Ethiopic version

The Ethiopic version. The Ethiopic version poses special problems. The earliest Bible probably was based on Greek versions, after Ethiopia had been converted to Christianity during the 4th and 5th centuries. The earliest existing manuscripts, however, belong to the 13th century. Most manuscripts from the 14th century on seem to reflect Arabic or Coptic influence, and it is not certain whether these represent the original translation or later ones. Many readings agree with the Hebrew against the Septuagint, which may have been caused by a Hexaplaric influence.

The Gothic version. The Gothic version was produced in the mid-4th century by Ulfilas, a Christian missionary who also invented the Gothic alphabet. It constitutes practically all that is left of Gothic literature. The translation of the Old Testament has entirely disappeared except for fragments of Ezra and Nehemiah. Though a Greek base is certain, some scholars deny the attribution of these remnants to Ulfilas.

The Old Latin version. The existence of a Latin translation can be attested in North Africa and southern Gaul as early as the second half of the 2nd century CE, and in Rome at the beginning of the following century. Its origins may possibly be attributed to a Christian adoption of biblical versions made by Jews in the Roman province of Africa, where the vernacular was exclusively Latin. Only portions or quotations from it, however, have been preserved, and from these it can be assumed that the translation was made not from Hebrew but from Greek. For this reason, the Old Latin version is especially valuable because it reflects the state of the Septuagint before Origen's revision. By the 3rd century, several Latin versions circulated, and African and European recensions can be differentiated. Whether they all diverged from an original single translation or existed from the beginning independently cannot be determined. The textual confusion and the vulgar and colloquial nature of the Old Latin recension had become intolerable to the church authorities by the last decade of the 4th century, and *c.* 382 Pope Damasus decided to remedy the situation.

The Vulgate. The task of revision fell to Eusebius Hieronymus, generally known as St. Jerome (died 419/420), whose knowledge of Latin, Greek, and Hebrew made him the outstanding Christian biblical scholar of his time.

Jerome produced three revisions of the Psalms, all extant. The first was based on the Septuagint and is known as the Roman Psalter because it was incorporated into the liturgy at Rome. The second, produced in Palestine from the Hexaplaric Septuagint, tended to bring the Latin closer to the Hebrew. Its popularity in Gaul was such that it came to be known as the Gallican Psalter. This version was later adopted into the Vulgate. The third revision, actually a fresh translation, was made directly from the Hebrew, but it never enjoyed wide circulation. In the course of preparing the latter, Jerome realized the futility of revising the Old Latin solely on the basis of the Greek and apparently left that task unfinished. By the end of 405 he had executed his own Latin translation of the entire Old Testament based on the "Hebrew truth" (*Hebraica veritas*).

Because of the canonical status of the Greek version within the church, Jerome's version was received at first with much suspicion, for it seemed to cast doubt on the

By courtesy of the (left) British and Foreign Bible Society, London, (right) Bibliotheque Nationale, Paris



(Left) Coptic papyrus of the Gospel According to John, 4th century. In the library of the British and Foreign Bible Society, London. (Right) Illustrated text from the earliest known Ethiopic Bible. In the Bibliothèque Nationale, Paris.

Acceptance of the Vulgate within the Western Church

authenticity of the Septuagint and exhibited divergences from the Old Latin that sounded discordant to those familiar with the traditional renderings. Augustine feared a consequent split between the Greek and Latin churches. The innate superiority of Jerome's version, however, assured its ultimate victory, and by the 8th century it had become the Latin Vulgate ("the common version") throughout the churches of Western Christendom, where it remained the chief Bible until the Reformation.

In the course of centuries of rival coexistence, the Old Latin and Jerome's Vulgate tended to react upon each other so that the Vulgate text became a composite. Other corruptions—noted in over 8,000 surviving manuscripts—crept in as a result of scribal transmission. Several medieval attempts were made to purify the Vulgate, but with little success. In 1546 the reforming Council of Trent accorded this version "authentic" status, and the need for a corrected text became immediate, especially because printing (introduced in the mid-15th century) could ensure, at last, a stabilized text. Because the Sixtine edition of Pope Sixtus V (1590) did not receive widespread support, Pope Clement VIII produced a fresh revision in 1592. This Clementine text remained the official edition of the Roman Church. Since 1907, the Benedictine Order, on the initiative of Pope Pius X, has been preparing a comprehensive edition. By 1969 only the Prophets still awaited publication to complete the Old Testament. A year later, a papal commission under Cardinal Augustinus Bea of Germany was charged with the task of preparing a new "revision of the Vulgate," taking the Benedictine edition as its working base.

Syriac versions. The Bible of the Syriac Churches is known as the Peshitta ("simple" translation). Though neither the reason for the title nor the origins of the versions are known, the earliest translations most likely served the needs of the Jewish communities in the region of Adiabene (in Mesopotamia), which are known to have existed as early as the 1st century CE. This probably explains the archaic stratum unquestionably present in the Pentateuch, Prophets, and Psalms of the Peshitta, as well as the undoubtedly Jewish influences generally, though Jewish–Christians also may have been involved in the rendering.

The Peshitta displays great variety in its style and in the translation techniques adopted. The Pentateuch is closest to the Masoretic text, but elsewhere there is much affinity with the Septuagint. This latter phenomenon might have resulted from later Christian revision.

Following the split in the Syriac Church in the 5th century into Nestorian (East Syrian) and Jacobite (West Syrian) traditions, the textual history of the Peshitta became bifurcated. Because the Nestorian Church was relatively isolated, its manuscripts are considered to be superior.

A revision of the Syriac translation was made in the early 6th century by Philoxenos, bishop of Mabbug, based on the Lucianic recension of the Septuagint. Another (the Syro-Hexaplaric version) was made by Bishop Paul of Tella in 617 from the Hexaplaric text of the Septuagint. A Palestinian Syriac version, extant in fragments, is known to go back to at least 700, and a fresh recension was made by Jacob of Edessa (died 708).

There are many manuscripts of the Peshitta, of which the oldest bears the date 442. Only four complete codices are extant from between the 5th and 12th centuries. No critical edition yet exists, but one is being prepared by the Peshitta Commission of the International Organization for the Study of the Old Testament.

Arabic versions. There is no reliable evidence of any pre-Islāmic Arabic translation. Only when large Jewish and Christian communities found themselves under Muslim rule after the Arab conquests of the 7th century did the need for an Arabic vernacular Scripture arise. The first and most important was that of Saʿadia ben Joseph (892–942), made directly from Hebrew and written in Hebrew script, which became the standard version for all Jews in Muslim countries. The version also exercised its influence upon Egyptian Christians and its rendering of the Pentateuch was adapted by Abū al-Ḥasan to the Samaritan Torah in the 11th–12th centuries. Another Samaritan Arabic version of the Pentateuch was made by Abū Saʿīd (Abū al-Barakāt) in the 13th century. Among other translations from the Hebrew, that of the 10th-century Karaite Yāphith ibn ʿAlī is the most noteworthy.

Influence of Arabic versions

In 946 a Spanish Christian of Córdova, Isaac son of Velásquez, made a version of the Gospels from Latin. Manuscripts of 16th-century Arabic translations of both testaments exist in Leningrad, and both the Paris and London polyglots of the 17th century included Arabic versions. In general, the Arabic manuscripts reveal a bewildering variety of renderings dependent on Hebrew, Greek, Samaritan, Syriac, Coptic, and Latin translations. As such they have no value for critical studies. Several modern Arabic translations by both Protestants and Catholics were made in the 19th and 20th centuries.

Later and modern versions: English. Knowledge of the pre-Wycliffite English renditions stems from the many actual manuscripts that have survived and from secondary literature, such as booklists, wills, citations by later authors, and references in polemical works that have preserved the memory of many a translation effort.

Anglo-Saxon versions. For about seven centuries after the conversion of England to Christianity (beginning in the 3rd century), the common man had no direct access to the text of the Scriptures. Ignorant of Latin, his knowledge was derived principally from sermons and metrical prose paraphrases and summaries. The earliest poetic rendering of any part of the Bible is credited to Caedmon (flourished 658–680), but only the opening lines of his poem on the Creation in the Northumbrian dialect have been preserved.

An actual translation of the Psalter into Anglo-Saxon is ascribed to Aldhelm, bishop of Sherborne (died 709), but nothing has survived by which its true character, if it actually existed, might be determined. Linguistic considerations alone rule out the possibility that the prose translation of Psalms 1–50 extant in the Bibliothèque Nationale at Paris is a 7th-century production. In the next century, Bede (died 735) is said to have translated parts of the Gospels, and though he knew Greek and possibly even some Hebrew, he does not appear to have applied himself to the Old Testament.

The outstanding name of the 9th century is that of King Alfred the Great. He appended to his laws a free translation of the Ten Commandments and an abridgment of the enactments of Exodus 21–23. These actually constitute the earliest surviving examples of a portion of the Old Testament in Anglo-Saxon prose.

An important step towards the emergence of a true English translation was the development of the interlinear gloss, a valuable pedagogic device for the introduction of youthful members of monastic schools to the study of the Bible. The Vespasian Psalter is the outstanding surviving example of the technique from the 9th century. In the next century the Lindisfarne Gospels, written in Latin *c.* 700, were glossed in Anglo-Saxon *c.* 950.

The last significant figure associated with the vernacular Bible before the Norman Conquest was the so-called Aelfric the Grammarian (*c.* 955–1020). Though he claimed to have rendered several books into English, his work is more a paraphrase and abridgment than a continuous translation.

Anglo-Norman versions. The displacement of the English upper class, with the consequent decline of the Anglo-Saxon tradition attendant upon the Norman invasion, arrested for a while the movement toward the production of the English Bible. Within about 50 years (*c.* 1120) of the Conquest Eadwine's *Psalterium triplex*, which contained the Latin version accompanied by Anglo-Norman and Anglo-Saxon renderings, appeared. The contemporary Oxford Psalter achieved such influence that it became the basis of all subsequent Anglo-Norman versions. By 1361 a prose translation of most of Scripture in this dialect had been executed.

The Wycliffite versions. By the middle of the 13th century the English component in the Anglo-Norman amalgam had begun to assert itself and the close of the

century witnessed a Northumbrian version of the Psalter made directly from Latin, which, because it survived in several manuscripts, must have achieved relatively wide circulation. By the next century, English had gradually superseded French among the upper classes. When the first complete translation of the Bible into English emerged, it became the object of violent controversy because it was inspired by the heretical teachings of John Wycliffe. Intended for the common man, it became the instrument of opposition to ecclesiastical authority.

Effect of the Wycliffite Bible

The exact degree of Wycliffe's personal involvement in the Scriptures that came to bear his name is not clear. Because a note containing the words "Here ends the translation of Nicholas of Hereford," is found in a manuscript copy of the original (and incomplete) translation, it may be presumed that, though there must have been other assistants, Hereford can be credited with overall responsibility for most of the translation and that his summons before a synod in London and his subsequent departure for Rome in 1382 terminated his participation in the work. Who completed it is uncertain.

The Wycliffite translations encountered increasing ecclesiastical opposition. In 1408 a synod of clergy summoned to Oxford by Archbishop Arundel forbade the translation and use of Scripture in the vernacular. The proscription was rigorously enforced, but remained ineffectual. In the course of the next century the Wycliffite Bible, the only existing English version, achieved wide popularity as is evidenced by the nearly 200 manuscripts extant, most of them copied between 1420–50.

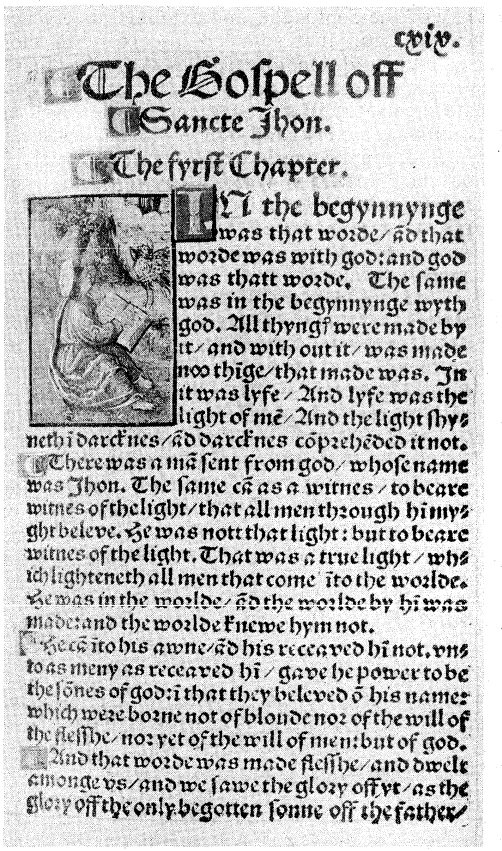
The translation of William Tyndale. Because of the influence of printing and a demand for scriptures in the vernacular, William Tyndale began working on a New Testament translation directly from the Greek in 1523. The work could not be continued in England because of political and ecclesiastical pressures, and the printing of his translation began in Cologne (in Germany) in 1525. Again under pressure, this time from the city authorities, Tyndale had to flee to Worms, where two complete editions were published in 1525. Copies were smuggled into England where they were at once proscribed. Of 18,000 copies printed (1525–28), two complete volumes and a fragment are all that remain.

When the New Testament was finished Tyndale began work on the Old Testament. The Pentateuch was issued in Marburg in 1530, each of the five books being separately published and circulated. Tyndale's greatest achievement was the ability to strike a felicitous balance between the needs of scholarship, simplicity of expression and literary gracefulness, all in a uniform dialect. The effect was the creation of an English style of Bible translation, tinged with Hebraisms, that was to serve as the model for all future English versions for nearly 400 years.

Influence of Tyndale's Bible

The translation of Miles Coverdale. A change in atmosphere in England found expression in a translation that, for all its great significance, turned out to be a retrograde step in the manner of its execution, although it proved to be a vindication of Tyndale's work. On October 4, 1535, the first complete English Bible, the work of Miles Coverdale, came off the press either in Zürich or in Cologne. The edition was soon exhausted. A second impression appeared in the same year and a third in 1536. A new edition, "overseen and corrected," was published in England by James Nycholson in Southwark in 1537. Another edition of the same year bore the announcement, "set forth with the king's most gracious license." In 1538 a revised edition of Coverdale's New Testament printed with the Latin Vulgate in parallel columns issued in England was so full of errors that Coverdale promptly arranged for a rival corrected version to appear in Paris.

The Thomas Matthew version. In the same year that Coverdale's authorized version appeared, another English Bible was issued under royal license and with the encouragement of ecclesiastical and political power. It appeared (Antwerp?) under the name of Thomas Matthew, but it is certainly the work of John Rogers, a close friend of Tyndale. Although the version claimed to be "truly and purely translated into English," it was in reality a combination of the labours of Tyndale and Coverdale. Rogers used the former's Pentateuch and 1535 revision of the New Testament and the latter's translation from Ezra to Malachi and his Apocrypha. Rogers' own contribution was primarily editorial.


cxiv.

The Gospell off Sancte Jhon.

The fyrst Chapter.

IN the begynnynge was that worde/ & that worde was with god: and god was thatt worde. The same was in the begynnynge wyth god. All thyngs were made by it/ and with out it/ was made noo thige/ that made was. In it was lyfe/ And lyfe was the light of me/ And the light shyneth i darcknes/ & darcknes cõprehẽded it not. There was a mã sent from god/ whose name was Jhon. The same cã as a witnes/ to beare witnes of the light/ that all men through hi myght beleve. He was nott that light: but to beare witnes of the light. That was a true light/ which lighteneth all men that come ito the worlde. He was in the worlde/ & the worlde by hi was made: and the worlde knewe hym not. He cõ ito his awne/ & his receaved hi not. vnto as meny as receaved hi/ gave he power to be the sõnes of god: i that they beleved õ his name: which were borne not of blonde nor of the will of the flesshe/ nor yet of the will of man: but of god. And that worde was made flesshe/ and dwelt amonge vs/ and we sawe the glory offyt/ as the glory off the only begotten sonne off the father/

The opening page of chapter 1 of the Gospel According to John from Tyndale's Bible, 1525–26. In the library of the Baptist College, Bristol, England.
By courtesy of the Baptist College, Bristol, England

The Great Bible. In an injunction of 1538, Henry VIII commanded the clergy to install in a convenient place in every parish church, "one book of the whole Bible of the largest volume in English." The order seems to refer to an anticipated revision of the Matthew Bible. The first edition was printed in Paris and appeared in London in April 1539 in 2,500 copies. The huge page-size earned it the sobriquet the Great Bible. It was received with immediate and wholehearted enthusiasm.

Henry VIII's support of the Great Bible

The first printing was exhausted within a short while, and it went through six subsequent editions between 1540 and 1541. "Editions" is preferred to "impressions" here since the six successive issues were not identical.

The Geneva Bible. The brief efflorescence of the Protestant movement during the short reign of Edward VI (1547–53) saw the reissue of the Scriptures, but no fresh attempts at revision. The repressive rule of Edward's successor, Mary, a Roman Catholic, put an end to the printing of Bibles in England for several years. Their public reading was proscribed and their presence in the churches discontinued.

The persecutions of Protestants caused the focus of English biblical scholarship to be shifted abroad where it flourished in greater freedom. A colony of Protestant exiles, led by Coverdale and John Knox (the Scottish Reformer), and under the influence of John Calvin, published the New Testament in 1557.

The editors of the Geneva Bible (or "Breeches Bible," so-named because of its rendering of the first garments made for Adam and Eve in chapter three, verse seven of Genesis)—published in 1560—may almost certainly be identified as William Whittingham, the brother-in-law of Calvin's wife, and his assistants Anthony Gilby and Thomas Sampson. The Geneva Bible was not printed in

England until 1576, but it was allowed to be imported without hindrance. The accession of Elizabeth in 1558 put an end to the persecutions and the Great Bible was soon reinstated in the churches. The Geneva Bible, however, gained instantaneous and lasting popularity over against its rival, the Great Bible. Its technical innovations contributed not a little to its becoming for a long time the family Bible of England which, next to Tyndale, exercised the greatest influence upon the King James Version.

The Bishops' Bible. The failure of the Great Bible to win popular acceptance against the obvious superiority of its Geneva rival and the objectionable partisan flavour of the latter's marginal annotations made a new revision a necessity. By about 1563–64 Archbishop Matthew Parker of Canterbury had determined upon its execution and the work was apportioned among many scholars, most of them bishops, from which the popular name was derived.

The Bishops' Bible came off the press in 1568 as a handsome folio volume, the most impressive of all 16th-century English Bibles in respect of the quality of paper, typography, and illustrations. A portrait of the Queen adorned the engraved title page, but it contained no dedication. For some reason Queen Elizabeth never officially authorized the work, but sanction for its public use came from the Convocation (church synod or assembly) of 1571 and it thereby became in effect the second authorized version.

Need for a Roman Catholic version

The Reims-Douai Bible. The Roman Catholics addressed themselves affirmatively to the same problem faced by the Anglican Church: a Bible in the vernacular. The initiator of the first such attempt was Cardinal Allen of Reims (in France), although the burden of the work fell to Gregory Martin, professor of Hebrew at Douai. The New Testament appeared in 1582, but the Old Testament, delayed by lack of funds, did not appear until 1609–10 when it was finally published at Douai under the editorship of Thomas Worthington. In the intervening period it had been brought into line with the new text of the Vulgate authorized by Clement VIII in 1592.

The King James (Authorized) Version. Because of changing conditions, another official revision of the Protestant Bible in English was needed. The reign of Queen Elizabeth had succeeded in imposing a high degree of uniformity upon the church. The failure of the Bishops' Bible to supplant its Geneva rival made for a discordant note in the quest for unity.

A conference of churchmen in 1604 became noteworthy for its request that the English Bible be revised because existing translations "were corrupt and not answerable to the truth of the original." King James I was quick to appreciate the broader value of the proposal and at once made the project his own.

By June 30, 1604, King James had approved a list of 54 revisers, although extant records show that 47 scholars actually participated. They were organized into six companies, two each working separately at Westminster, Oxford, and Cambridge on sections of the Bible assigned to them. It was finally published in 1611.

Significance of the King James Version

Not since the Septuagint had a translation of the Bible been undertaken under royal sponsorship as a cooperative venture on so grandiose a scale. An elaborate set of rules was contrived to curb individual proclivities and to ensure its scholarly and nonpartisan character. In contrast to earlier practice, the new version was to preserve vulgarly used forms of proper names in keeping with its aim to make the Scriptures popular and familiar.

The impact of Jewish sources upon the King James Version is one of its noteworthy features. The wealth of scholarly tools available to the translators made their final choice of rendering an exercise in originality and independent judgment. For this reason, the new version was more faithful to the original languages of the Bible and more scholarly than any of its predecessors. The impact of the Hebrew upon the revisers was so pronounced that they seem to have made a conscious effort to imitate its rhythm and style in the Old Testament. The English of the New Testament actually turned out to be superior to its Greek original.

Two editions were actually printed in 1611, later distinguished as the "He" and "She" Bibles because of the variant reading "he" and "she" in the final clause of chapter 3, verse 15 of Ruth: "and he went into the city." Both printings contained errors. Some errors in subsequent editions have become famous: The so-called Wicked Bible (1631) derives from the omission of "not" in chapter 20 verse 14 of Exodus, "Thou shalt commit adultery," for which the printers were fined £300; the "Vinegar Bible" (1717) stems from a misprinting of "vineyard" in the heading of Luke, chapter 20.

The English Revised Version. The remarkable and total victory of the King James Version could not entirely obscure those inherent weaknesses that were independent of its typographical errors. The manner of its execution had resulted in a certain unequalness and lack of consistency. The translators' understanding of the Hebrew tense system was often limited so that their version contains inaccurate and infelicitous renderings. In particular, the Greek text of the New Testament, which they used as their base, was a poor one. The great early Greek codices were not then known or available, and Hellenistic papyri, which were to shed light on the common Greek dialect, had not yet been discovered.

A committee established by the Convocation of Canterbury in February 1870 reported favourably three months later on the idea of revising the King James Version: two companies were formed, one each for the Old and New Testaments. A novel development was the inclusion of scholars representative of the major Christian denominations, except the Roman Catholics (who declined the invitation to participate). Another innovation was the formation of parallel companies in the United States to whom the work of the English scholars was submitted and who, in turn, sent back their reactions. The instructions to the committees made clear that only a revision and not a new translation was contemplated.

Cooperation in the publication of the Revised Version

The New Testament was published in England on May 17, 1881, and three days later in the United States after 11 years of labour. Over 30,000 changes were made, of which more than 5,000 represent differences in the Greek text from that used as the basis of the King James Version. Most of the others were made in the interests of consistency or modernization.

The publication of the Old Testament in 1885 stirred far less excitement, partly because it was less well known than the New Testament and partly because fewer changes were involved. The poetical and prophetical books, especially Job, Ecclesiastes, and Isaiah, benefitted greatly.

The revision of the Apocrypha, not originally contemplated, came to be included only because of copyright arrangements made with the university presses of Oxford and Cambridge and was first published in 1895.

The American Standard Version. According to the original agreement, the preferred readings and renderings of the American revisers, which their British counterparts had declined to accept, were published in an appendix to the Revised Version. In 1900 the American edition of the New Testament, which incorporated the American scholars' preferences into the body of the text, was produced. A year later the Old Testament was added, but not the Apocrypha. The alterations covered a large number of obsolete words and expressions and replaced Anglicisms by the diction then in vogue in the United States.

The Revised Standard Version. The American Standard Version had been an expression of sensitivity to the needs of the American public. At the same time, several individual and unofficial translations into modern speech made from 1885 on had gained popularity, their appeal reinforced by the discovery that the Greek of the New Testament used the common nonliterary variety of the language spoken throughout the Roman Empire when Christianity was in its formative stage. The notion that a nonliterary modern rendering of the New Testament best expressed the form and spirit of the original was hard to refute. This, plus a new maturity of classical, Hebraic, and theological scholarship in the United States,

led to a desire to produce a native American version of the English Bible.

In 1928 the copyright of the American Standard Version was acquired by the International Council of Religious Education and thereby passed into the ownership of churches representing 40 major denominations in the United States and Canada. A two-year study by a special committee recommended a thorough revision, and in 1937 the council gave its authorization to the proposal. Not until 1946, however, did the revision of the New Testament appear in print, and another six years elapsed before the complete Revised Standard Version (RSV) was published, the work of 32 scholars, one of them Jewish, drawn from the faculties of 20 universities and theological seminaries. A decision to translate the Apocrypha was not made until 1952 and the revision appeared in 1957. Insofar as the RSV was the first to make use of the Dead Sea Scroll of Isaiah, it was revolutionary.

Modernization in the Revised Standard Version

The Revised Standard Version was essentially not a new translation into modern speech, but a revision. It did engage in a good deal of modernization, however. It dispensed with archaic pronouns, retaining "thou" only for the Deity. But its basic conservatism was displayed in the retention of forms or expressions in passages that have special devotional or literary associations even where this practice makes for inconsistency. The primary aim was to produce a version for use in private and public worship.

Jewish versions. Though Jews in English-speaking lands generally utilized the King James Version and the Revised Version, the English versions have presented great difficulties. They contain departures from the traditional Hebrew text; they sometimes embody Christological interpretations; the headings were often doctrinally objectionable and the renderings in the legal portions of the Pentateuch frequently diverged from traditional Jewish exegesis. In addition, where the meaning of the original was obscure Jewish readers preferred to use the well-known medieval Jewish commentators. Finally, the order of the Jewish canon differs from Christian practice and the liturgical needs of Jews make a version that does not mark the scriptural readings for Sabbaths and festivals inconvenient.

Until 1917 all Jewish translations were the efforts of individuals. Planned in 1892, the project of the Jewish Publication Society of America was the first translation for which a group representing Jewish learning among English-speaking Jews assumed joint responsibility.

This version essentially retained the Elizabethan diction. It stuck unswervingly to the received Hebrew text that it interpreted in accordance with Jewish tradition and the best scholarship of the day. For over half a century it remained authoritative, even though it laid no claim to any official ecclesiastical sanction.

With an increasingly felt need for modernization, a committee was established comprising three professional biblical and Semitic scholars and three rabbis. It began its work in 1955 and the Pentateuch was issued in 1962. The Song of Songs, Ruth, Lamentations, Ecclesiastes, Esther, and Jonah, all in a single volume for the convenience of synagogue use, followed in 1969; and Isaiah and Psalms appeared in 1973. A second committee had been set up in 1955 to work separately on the rest of the Hagiographa.

The New English Bible. The idea of a completely new translation into British English was first broached in 1946. Under a joint committee, representative of the major Protestant churches of the British Isles, with Roman Catholics appointed as observers, the New Testament was published in 1961 and a second edition appeared in 1970. The Old Testament and Apocrypha were also published in 1970.

Success of the New English Bible

The New English Bible proved to be an instant commercial success, selling at a rate of 33,000 copies a week in 1970. The translation differed from the English mainstream Bible in that it was not a revision but a completely fresh version from the original tongues. It abandoned the tradition of "biblical English" and, except for the retention of "thou" and "thy" in addressing God, freed itself of all archaisms. It endeavoured to render the original into the idiom of contemporary English and to avoid ephemeral modernisms.

Catholic versions. With the exception of a version by Irish-American archbishop Francis Patrick Kenrick (1849–60), all translations up to the 20th century were merely versions of the Reims-Douai Bible. A celebrated translation was that of Ronald Knox (New Testament, 1945; Old Testament, 1949; complete edition with Old Testament revised, 1955).

The most significant development in modern Catholic translations was initiated by the Confraternity of Christian Doctrine in 1936. A New Testament version of the Latin Clementine Vulgate (1941), intended as a revision, in effect was a new translation into clear and simple English. The Old Testament revision remained unfinished, the work having been interrupted by a decision inspired by the Pontifical Biblical Commission in 1943 to encourage modern vernacular translations from the original languages instead of from the Latin Vulgate. Accordingly, both the Old and New Testaments were respectively retranslated into modern English from the Hebrew and Greek originals. The resultant Confraternity Version (1952–61) was later issued as the New American Bible (1970). Another modern version, more colloquial, is the Jerusalem Bible (1966), translated from the French Catholic Bible de Jérusalem (one-volume edition, 1961).

Later and modern versions: continental. *Dutch versions.* Until the Reformation, Dutch Bible translations were largely free adaptations, paraphrases, or rhymed verse renderings of single books or parts thereof. A popular religious revival at the end of the 12th century accelerated the demand for the vernacular Scriptures, and one of the earliest extant examples is the Liège manuscript (*c.* 1270) translation of the *Diatessaron* (a composite rendering of the four Gospels) by Tatian, a 2nd century Syrian Christian heretical scholar; it is believed to derive from a lost Old Latin original. Best known of all the rhymed versions is the *Rijmbijbel* of Jacob van Maerlant (1271) based on Peter Comestar's *Historia scholastica.* Despite the poor quality of Johan Schutken's translation of the New Testament and Psalms (1384), it became the most widely used of medieval Dutch versions.

Early vernacular versions

With the Reformation came a renewed interest in the study of the Scriptures. Luther's Bible (see *German versions,* below) was repeatedly rendered into Dutch, the most important version being that of Jacob van Liesveldt (1526). It was mainly to counter the popularity of this edition that Roman Catholics produced their own Dutch Bible, executed by Nicolaas van Winghe (Louvain, 1548). A revision printed by Jan Moerentorf (Moretus, 1599) became the standard version until it was superseded by that of the Peter Canisius Association (1929–39), now in general use. A fresh translation of the New Testament in modern Dutch appeared in 1961.

French versions. The deep conflicts that characterized the history of Christianity in France made it difficult for one authoritative version to emerge.

The first complete Bible was produced in the 13th century at the University of Paris and toward the end of that century Guyart des Moulins executed his *Bible Historiale.* Both works served as the basis of future redactions of which the Bible printed in Paris (date given variously as 1487, 1496, 1498) by order of King Charles VIII, is a good example.

The real history of the French Bible began in Paris, in 1523, with the publication of the New Testament, almost certainly the work of the Reformer Jacques Lefèvre d'Étaples (Faber Stapulensis). The Old Testament appeared in Antwerp in 1528 and the two together in 1530 as the Antwerp Bible. The first true Protestant version came out in Serrières, near Neuchâtel five years later, the work of Pierre-Robert Olivétan. This version was frequently revised throughout the 16th century, the most celebrated editions being Calvin's of 1546 and that of Robert Estienne (Stephanus) of 1553. The Roman Catholics produced a new version, the Louvain Bible of 1550, based on both Lefèvre and Olivétan. Modernizations of Olivétan appeared in succeeding centuries. The most im-

portant French version of the 20th century is the Jerusalem Bible prepared by professors at the Dominican École Biblique de Jérusalem (Paris, 1949–54, complete, 1956).

Gothic and pre-Reformation Bibles

German versions. The early Old Testament in Gothic has already been described. The New Testament remains are far more extensive and are preserved mainly in the Codex Argenteus (*c.* 525) and Codex Gissensis. The translation, essentially based on a Byzantine text, is exceedingly literal and not homogeneous. It is difficult to determine the degree of contamination that the original Gospels translation of Ulfilas had undergone by the time it appeared in these codices.

Nothing is known of the vernacular Scriptures in Germany prior to the 8th century when an idiomatic translation of Matthew from Latin into the Bavarian dialect was made. From Fulda (in Germany) *c.* 830 came a more literal East Franconian German translation of the Gospel story. In the same period was produced the *Heliand* ("Saviour"), a versified version of the Gospels. Such poetic renderings cannot, strictly speaking, be regarded as translations. There is evidence, however, for the existence of German Psalters from the 9th century on. By the 13th century, the different sects and movements that characterized the religious situation in Germany had stimulated a demand for popular Bible reading. Since all the early printed Bibles derived from a single family of late 14th-century manuscripts, German translations must have gained wide popularity. Another impetus towards the use of the German Scriptures in this period can be traced to mystics of the Upper Rhine. A complete New Testament, the Augsburg Bible, can be dated to 1350 and another from Bohemia, Codex Teplensis (*c.* 1400), has also survived.

The Wenzel Bible, an Old Testament made between 1389 and 1400, is said to have been ordered by the emperor Wenzel, and large numbers of 15th-century manuscripts have been preserved.

The first printed Bible (the Mentel Bible) appeared at Strassbourg no later than 1466 and ran through 18 editions before 1522. Despite some evidence that ecclesiastical authority did not entirely look with favour upon this vernacular development, the printed Bible appeared in Germany earlier, and in more editions and in greater quantity than anywhere else.

A new era opened up with the work of Martin Luther, to whom a translation from the original languages was a necessary and logical conclusion of his doctrine of justification by faith—to which the Scriptures provided the only true key. His New Testament (Wittenberg, 1522) was made from the second edition of Erasmus' Greek Testament. The Old Testament followed in successive parts, based on the Brescia Hebrew Bible (1494). Luther's knowledge of Hebrew and Aramaic was limited, but his rendering shows much influence of Rashi, the great 11th–12th-century French rabbinical scholar and commentator, through the use of the notes of Nicholas of Lyra. The complete Lutheran Bible emerged from the press in 1534. Luther was constantly revising his work with the assistance of other scholars, and between 1534 and his death in 1546, 11 editions were printed, the last posthumously. His Bible truly fulfilled Luther's objective of serving the needs of the common man, and it, in turn, formed the basis of the first translations in those lands to which Lutheranism spread. It proved to be a landmark in German prose literature and contributed greatly to the development of the modern language.

Influence of Luther's Bible

The phenomenal success of Luther's Bible and the failure of attempts to repress it led to the creation of German Catholic versions, largely adaptations of Luther. Hieronymus Emser's edition simply brought the latter into line with the Vulgate. Johann Dietenberger issued a revision of Emser (Mainz, 1534) and used Luther's Old Testament in conjunction with an Anabaptist (radical Protestant group) version and the Zürich (Switzerland) version of 1529. It became the standard Catholic version. Of the 20th-century translations, the Grünewald Bible, which reached a seventh edition in 1956, is one of the most noteworthy.

Yiddish versions

German glosses in Hebrew script attached to Hebrew Bibles in the 12th and 13th centuries constitute the earliest Jewish attempts to render the Scriptures into that German dialect current among the Jews of middle Europe, the dialect that developed into Judeo-German or Yiddish. The first translation proper has been partially preserved in a manuscript from Mantua dated 1421. The earliest printed translation is that of the Scriptural dictionaries prepared by a baptized Jew, Michael Adam (Constance, 1543–44; Basel, 1583, 1607). The version of Jacob ben Isaac Ashkenazi of Janów, known as the *Tz'enah u-Re'na* (Lublin, 1616), became one of the most popular and widely diffused works of its kind.

The first Jewish translation into pure High German, though in Hebrew characters (1780–83), made by Moses Mendelssohn, opened a new epoch in German-Jewish life. The first Jewish rendering of the entire Hebrew Bible in German characters was made by Gotthold Salomon (Altona, 1837). An attempt to preserve the quality of the Hebrew style in German garb was the joint translation of two Jewish religious philosophers, Martin Buber and Franz Rosenzweig (15 vol., Berlin, 1925–37; revised ed. Cologne, 4 vol. 1954–62).

Greek versions. A 13th-century manuscript of Jonah by a Jew is the earliest known post-Hellenistic Greek biblical work. A rendering of Psalms was published by a Cretan monk Agapiou in 1563. A version in Hebrew characters (a large part of the Old Testament) appeared in the Constantinople Polyglot Pentateuch in 1547.

The first New Testament was done by Maximus of Gallipoli in 1638 (at Geneva?). The British and Foreign Bible Society published the Old Testament in 1840 (London) and the New Testament in 1848 (Athens). Between 1900 and 1924, however, the use of a modern Greek version was prohibited. The theological faculty of the University of Athens is now preparing a fresh translation.

Hungarian versions. The spread of Lutheranism in the Reformation period gave rise to several vernacular versions. János Sylvester (Erdősi) produced the first New Testament made from the Greek (Sárvár, 1541). The Turkish occupation of much of Hungary and the measures of the Counter-Reformation arrested further printing of the vernacular Bible, except in the semi-independent principality of Transylvania. The first complete Hungarian Bible, issued at Vizsoly in 1590, became the Protestant Church Bible.

In the 20th century, a new standard edition for Protestants was published, the New Testament appearing in 1956 and the Old Testament (Genesis to Job) in 1951 and following. A new modernized Catholic edition of the New Testament from the Greek appeared in Rome in 1957.

Italian versions. The vernacular Scriptures made a relatively late appearance in Italy. Existing manuscripts of individual books derive from the 13th century and mainly consist of the Gospels and the Psalms.

These medieval versions were never made from the original languages. They were influenced by French and Provençal renderings as well as by the form of the Latin Vulgate current in the 12th and 13th centuries in southern France. There is evidence for a Jewish translation made directly from the Hebrew as early as the 13th century.

The first printed Italian Bible appeared in Venice in 1471, translated from the Latin Vulgate by Niccolò Malermi. In 1559 Paul IV proscribed all printing and reading of the vernacular Scriptures except by permission of the church. This move, reaffirmed by Pius IV in 1564, effectively stopped further Catholic translation work for the next 200 years.

The first Protestant Bible (Geneva, 1607, revised 1641) was the work of Giovanni Diodati, a Hebrew and Greek scholar. Frequently reprinted, it became the standard Protestant version until the 20th century. Catholic activity was renewed after a modification of the ban by Pope Benedict XIV in 1757. A complete Bible in translation made directly from the Hebrew and Greek has been in progress under the sponsorship of the Pontifical Biblical Institute since the 1920s.

Portuguese versions. The first Portuguese New Testament (Amsterdam), the work of João Ferreira d'Almeida, did not appear until 1681. The first complete Bible (2 vol., 1748–53) was printed in Batavia (in Holland). Not until late in the 18th century did the first locally published vernacular Scriptures appear in Portugal. A revision of d'Almeida was issued in Rio de Janeiro (in Brazil), the New Testament in 1910 and the complete Bible in 1914 and 1926; an authorized edition in modernized orthography was published by the Bible Society of Brazil (New Testament, 1951; Old Testament, 1958). A new translation of the New Testament from Greek by José Falcão came out in Lisbon (1956–65).

Scandinavian versions. In pre-Reformation times, only partial translations were made, all on the basis of the Latin Vulgate and all somewhat free. The earliest and most celebrated is that of Genesis-Kings in the so-called *Stjórn* ("Guidance"; *i.e.*, of God) manuscript in the Old Norwegian language, probably to be dated about 1300. Swedish versions of the Pentateuch and of Acts have survived from the 14th century and a manuscript of Joshua-Judges by Nicholaus Ragnvaldi of Vadstena from *c.* 1500. The oldest Danish version covering Genesis-Kings derives from 1470.

Danish, Norwegian, Icelandic, and Swedish versions

Within two years of publication, Luther's New Testament had already influenced a Danish translation made at the request of the exiled king Christian II by Christiern Vinter and Hans Mikkelsen (Wittenberg, 1524). In 1550 Denmark received a complete Bible commissioned by royal command (the Christian III Bible, Copenhagen). A revision appeared in 1589 (the Frederick II Bible) and another in 1633 (the Christian IV Bible).

A rendering by Hans Paulsen Resen (1605–07) was distinguished by its accuracy and learning and was the first made directly from Hebrew and Greek, but its style was not felicitous and a revision was undertaken by Hans Svane (1647). Nearly 200 years later (1819), a combination of the Svaning Old Testament and the Resen-Svane New Testament was published. In 1931 a royal commission produced a new translation of the Old Testament with the New Testament following in 1948 and the Apocrypha in 1957.

The separation of Norway from Denmark in 1814 stimulated the revival of literature in the native language. The Old Testament of 1842–87 (revised, 1891) and New Testament of 1870–1904 were still intelligible to Danish readers, but the version of E. Blix (New Testament 1889; complete Bible, 1921) is in New Norwegian. A revised Bible in this standardized form of the language, executed by R. Indrebö, was published by the Norwegian Bible Society in 1938.

The first Icelandic New Testament was the work of Oddur Gottskálksson (Roskilde, Denmark, 1540), based on the Latin Vulgate and Luther. It was not until 1584 that the complete Icelandic Scriptures were printed (at Hólar), mainly executed by Gudbrandur Thorlákson. It was very successful and became the Church Bible until displaced by the revision of Thorlákur Skúlason (1627–55), based apparently on Resen's Danish translation. In 1827 the Icelandic Bible Society published a new New Testament and a complete Bible in 1841 (Videyjar; 1859 Reykjavík), revised and reprinted at Oxford in 1866. A completely new edition (Reykjavík, 1912) became the official Church Bible.

Soon after Sweden achieved independence from Denmark in the early 16th century, it acquired its own version of the New Testament published by the royal press (Stockholm, 1526). Luther's New Testament of 1522 served as its foundation, but the Latin Vulgate and Erasmus' Greek were also consulted. The first official complete Bible and the first such in any Scandinavian country was the Gustav Vasa Bible (Upsala; 1541), named for the Swedish king under whose reign it was printed. It utilized earlier Swedish translations as well as Luther's. A corrected version (the Gustavus Adolphus Bible, named for the reigning Swedish king) was issued in 1618, and another with minor alterations by Eric Benzelius in 1703. The altered Bible was called the Charles XII Bible, because it was printed during the reign of Charles XII. In 1917 the church diet of the Lutheran Church published a completely fresh translation directly from modern critical editions of the Hebrew and Greek originals and it received the authorization of Gustaf V to become the Swedish Church Bible.

South, East, and West Slavic versions

Slavic versions. The earliest Old Church Slavonic translations are connected with the arrival of the brothers Cyril and Methodius in Moravia 863, and resulted from the desire to provide vernacular renderings of those parts of the Bible used liturgically. The oldest manuscripts derive from the 11th and 12th centuries. The earliest complete Bible manuscript, dated 1499, was used for the first printed edition (Ostrog, 1581). This was revised in Moscow in 1633 and again in 1712. The standard Slavonic edition is the St. Petersburg revision of 1751, known as the Bible of Elizabeth.

The printing of parts of the Bulgarian Bible did not begin until the mid-19th century. A fresh vernacular version of the whole Bible was published at Sofia in 1925, having been commissioned by the Synod of the Bulgarian Orthodox Church.

The Serbian and Croatian literary languages are identical; they differ only in the alphabet they use. To further the dissemination of Protestantism among the southern Slavs, Count Jan Ungnad set up a press in 1560 at Urach that issued a translation of the New Testament, in both Glagolitic (1562–63) and Cyrillic (1563) characters. The efforts of the Serbian leader Vuk Karadžić to establish the Serbo-Croatian vernacular on a literary basis resulted in a new translation of the New Testament (Vienna, 1847) that went through many revisions.

The spread of the Lutheran Reformation to the Slovene-speaking provinces of Austria stimulated the need for vernacular translations. The first complete Slovene Bible, translated from the original languages but with close reference to Luther's German, was made by Jurij Dalmatin (Wittenberg, 1584). Not until two centuries later did a Slovene Roman Catholic version, rendered from the Latin Vulgate, appear (Laibach, 1784–1802).

Between the 9th and 17th centuries the literary and ecclesiastical language of Russia was Old Slavonic. A vernacular Scriptures was thus late in developing. An incomplete translation into the Belorussian dialect was prepared by Franciscus Skorina (Prague, 1517–19) from the Latin Vulgate and Slavonic and Bohemian versions, but not until 1821 did the first New Testament appear in Russian, an official version printed together with the Slavonic. With the more liberal rule of Alexander II, the Holy Synod sponsored a fresh version of the Gospels in 1860. The Old Testament was issued at St. Petersburg in 1875. A Jewish rendering was undertaken by Leon Mandelstamm, who published the Pentateuch in 1862 (2nd ed., 1871) and the Psalter in 1864. Prohibited in Russia, it was first printed in Berlin. A complete Bible was published in Washington in 1952.

No manuscript in the Czech vernacular translation is known to predate the 14th century, but at least 50 complete or fragmentary Bibles have survived from the 15th. The first complete Bible was published in Prague in 1488 in a text based on earlier, unknown translations connected with the heretical Hussite movement. The most important production of the century, however, was that associated principally with Jan Blahoslav. Based on the original languages, it appeared at Kralice in six volumes (1579–93). The Kralice Bible is regarded as the finest extant specimen of classical Czech and became the standard Protestant version.

Closely allied to the Czech language, but not identical with it, Slovakian became a literary language only in the 18th century. A Roman Catholic Bible made from the Latin Vulgate by Jiři Palkovič was printed in the Gothic script (2 vol. Gran, 1829, 1832) and another, associated with Richard Osvald, appeared at Trnava in 1928. A Protestant New Testament version of Josef Roháček was published at Budapest in 1913 and his completed Bible at Prague in 1936. A new Slovakian version by Stefan Žlatoš and Anton Jan Surjanský was issued at Trnava in 1946.

A manuscript of a late 14th-century Psalter is the earli-

est extant example of the Polish vernacular Scriptures, and several books of the Old Testament have survived from the translation made from the Czech version for Queen Sofia (Sárospatak Bible, 1455). Otherwise, post-Reformation Poland supplied the stimulus for biblical scholarship. The New Testament first appeared in a two-volume rendering from the Greek by the Lutheran Jan Seklucjan (Königsberg, 1553). The "Brest Bible" of 1563, sponsored by Prince Radziwiłł, was a Protestant production made from the original languages. A version of this edition for the use of Socinians (Unitarians) was prepared by the Hebraist Szymon Budny (Nieswicz, 1570–82), and another revision, primarily executed by Daniel Mikołajewski and Jan Turnowski (the "Danzig Bible") in 1632, became the official version of all Evangelical churches in Poland. This edition was burnt by the Catholics and had to be subsequently printed in Germany. The standard Roman Catholic version (1593, 1599) was prepared by Jakób Wujek whose work, revised by the Jesuits, received the approval of the Synod of Piotrkow in 1607. A revised edition was put out in 1935.

Spanish versions. The history of the Spanish Scriptures is unusual in that many of the translations were based, not on the Latin Vulgate, but on the Hebrew, a phenomenon that is to be attributed to the unusual role played by Jews in the vernacular movement.

Nothing is known from earlier than the 13th century when James I of Aragon in 1233 proscribed the possession of the Bible in "romance" (the Spanish vernacular) and ordered such to be burnt. Several partial Old Testament translations by Jews as well as a New Testament from a Visigoth Latin text are known from this century. In 1417 the whole Bible was translated into Valencian Catalan, but the entire edition was destroyed by the Inquisition.

Between 1479 and 1504, royal enactments outlawed the vernacular Bible in Castile, Leon, and Aragon, and the expulsion of the Jews from Spain in 1492 transferred the centre of Spanish translation activity to other lands. In 1557, the first printed *Index of Forbidden Books* of the Spanish Inquisition prohibited the "Bible in Castilian romance or any other vulgar tongue," a ban that was repeated in 1559 and remained in force until the 18th century. In 1916 the Hispano-Americana New Testament appeared in Madrid as an attempt to achieve a common translation for the entire Spanish-speaking world. The first Roman Catholic vernacular Bible from the original languages was made under the direction of the Pontifical University of Salamanca (Madrid, 1944, 9th ed. 1959).

Swiss versions. Four parts of Luther's version were reprinted in the Swyzerdeutsch dialect in Zürich in 1524–25. The Prophets and Apocrypha appeared in 1529. A year later, the first Swiss Bible was issued with the Prophets and Apocrypha independently translated. The Swiss Bible underwent frequent revision between 1660 and 1882. A fresh translation from the original languages was made between 1907 and 1931.

Non-European versions. Translations of parts of the Bible are known to have existed in only seven Asian and four African languages before the 15th century. In the 17th century Dutch merchants began to interest themselves in the missionary enterprise among non-Europeans. A pioneer was Albert Cornelius Ruyl, who is credited with having translated Matthew into High Malay in 1629, with Mark following later. Jan van Hasel translated the two other Gospels in 1646 and added Psalms and Acts in 1652. Other traders began translations into Formosan Chinese (1661) and Sinhalese (1739).

A complete printed Japanese New Testament reputedly existed in Miyako in 1613, the work of Jesuits. The first known printed New Testament in Asia appeared in 1715 in the Tamil language done by Bartholomäus Ziegenbalg, a Lutheran missionary. A complete Bible followed in 1727. Six years later the first Bible in High Malay came out.

The distinction of having produced the first New Testament in any language of the Americas belongs to John Eliot, a Puritan missionary, who made it accessible to the Massachusetts Indians in 1661. Two years later he brought out the Massachusetts Indian Bible, the first Bible to be printed on the American continent.

By 1800 the number of non-European versions did not exceed 13 Asian, four African, three American, and one Oceanian. With the founding of missionary societies after 1800, however, new translations were viewed as essential to the evangelical effort. First came renderings in those languages that already possessed a written literature. A group at Serāmpore (in India) headed by William Carey, a Baptist missionary, produced 28 versions in Indian languages. Robert Morrison, the first Protestant missionary to China, translated the New Testament into Chinese in 1814 and completed the Bible by 1823. Adoniram Judson, an American missionary, rendered the Bible into Burmese in 1834.

With European exploration of the African continent often came the need to invent an alphabet, and in many instances the translated Scriptures constituted the first piece of written literature. In the 19th century the Bible was translated into Amharic, Malagasy, Tswana, Xosa, and Ga.

American Indian versions

In the Americas, James Evans invented a syllabary for the use of Cree Indians, in whose language the Bible was available in 1862, the work of W. Mason, also a Wesleyan missionary. The New Testament appeared in Ojibwa in 1833, and the whole Bible was translated for the Dakota Indians in 1879. The Labrador Eskimos had a New Testament in 1826 and a complete Bible in 1871.

In Oceania, the New Testament was rendered into Tahitian and Javanese in 1829 and into Hawaiian and Low Malay in 1835. By 1854 the whole Bible had appeared in all but the last of these languages as well as in Rarotonga (1851).

The most recent phase in the development of the non-European Bible translations has been characterized by an attempt to produce "union" or "standard" versions in the common language underlying different dialects. One such is the Swahili translation (1950) that makes the Scriptures accessible to most of East Africa. Another trend has been the updating of versions to bring them in line with the spoken language, especially through the use of native Christian scholars. The first example of this is the colloquial Japanese version of 1955.

By 1970 some part, if not the entire Bible, had been translated into more than 100 languages or dialects spoken in India and over 300 in Africa. (N.M.Sa.)

III. Old Testament history

History is a central element of the Old Testament. It is the subject of narration in the specifically historical books and of celebration, commemoration, and remonstration in all of the books. History in the Old Testament is not history in the modern sense; it is the story of events seen as revealing the divine presence and power. Nevertheless, it is the account of an actual people in an actual geographical area at certain specified historical times and in contact with other particular peoples and empires known from other sources. Hence, far more than with other great religious scriptures, a knowledge of the historical background is conducive, if not essential, to an adequate understanding of a major portion of the Old Testament. Recent archaeological discoveries as well as comparative historical research and philological studies, collated with an analysis and interpretation of the Old Testament text (still the major source of information), have made possible a fuller and more reliable picture of biblical history than in previous eras. For another presentation of Old Testament history, see JUDAISM, HISTORY OF: *Biblical Judaism.*

EARLY DEVELOPMENTS

Background and beginnings. The geographical theatre of the Old Testament is the ancient Near East, particularly the Fertile Crescent region, running from the Tigris and Euphrates rivers up to Syria and down through Palestine to the Nile Delta. In this area great civilizations and empires developed and semi-nomadic ethnic groups, such as the Hebrews, were involved in the mixture of peoples and cultures. The exact origin of the Hebrews is not

known with certainty, but the biblical tradition of their origin in a clan that migrated from Mesopotamia to Canaan (Palestine) early in the 2nd millennium BCE has analogues in what is known of the movements of other groups in that area and period. There are, moreover, obvious Mesopotamian motifs in biblical cosmogony and primeval history in the early part of the Bible, and Mesopotamian place-names are the obvious bases of some of the personal names of the clan's forebears (see also ABRAHAM). Canaanite influences are evident in the Hebrew alphabet, poetry, and certain mythological themes. Linguistic and other similarities with neighbouring Semitic peoples, such as the Amorites and Moabites, are also evident.

Exodus and conquest. According to biblical tradition, the clan migrated to Egypt because of a famine in the land of Canaan, were later enslaved and oppressed, and finally escaped from Egypt to the desert east of the Isthmus of Suez under a remarkable leader, Moses. The account—a proclamation, celebration, and commemoration of the event—is replete with legendary elements, but present-day scholars tend to believe that behind the legends there is a solid core of fact; namely, that Hebrew slaves who built the fortified cities of Pithom and Rameses somehow fled from Egypt, probably in the 13th century BCE, under a great leader (see also MOSES). A stele (inscribed stone pillar) of the pharaoh Merneptah of that time in which he claims to have destroyed Israel is the first known nonbiblical reference to the people by name. Whether the destruction was in the intervening desert or in Canaan (and whether a true or a false claim) is not clear. The tradition ascribes to Moses the basic features of Israel's faith: a single God, called YHWH, who cannot be represented iconically, bound in a covenant relationship with his special people Israel, to whom he has promised possession of (not, as with their forefathers, mere residence in) the land of Canaan. There is some dispute among scholars as to when such features as the Mosaic Covenant actually emerged and as to which of the traditional 12 tribes of Israel entered Canaan at the end of the period of wandering in the desert.

The work and faith of Moses

The biblical account of the conquest of Canaan is again, from the point of view of historical scholarship, full of legendary elements that express and commemorate the elation and wonder of the Israelites at these events. The conquest of Canaan—according to tradition, a united national undertaking led by Moses' successor, Joshua—was a rather drawn out and complicated matter. Archaeological evidence tends to refute some of the elements of the biblical account, confirm others, and leave some open. According to the tradition, after an initial unified assault that broke the main Canaanite resistance, the tribes engaged in individual mopping-up operations. Scholars believe that Hebrews who had remained resident in Canaan joined forces with the invading tribes, that the other Canaanite groups continued to exist, and that many of them later were assimilated by the Israelites.

The tribal league. The invading tribes who became masters of parts of Canaan, although effectively autonomous and lacking a central authority, considered themselves a league of 12 tribes, although the number 12 seems to have been more canonical or symbolical than historical. Some scholars, on the analogy of Greek leagues of six or 12 tribes or cities with a common sanctuary, speak of the Israelite league as an "amphictyony," the Greek term for such an association; but others hold that there is no evidence that the Israelites maintained a common shrine. Certain leaders arose, called judges, who might rule over several tribes, but this arrangement was usually of a local or regional character. However, the stories about such "judges" (who were frequently local champions or heroes, such as Gideon, Jephthah, and Samson), though encrusted with legend, are now thought to be substantially historical. The period from about 1200 to 1020 is called, after them, the period of the judges. It was during this period that Israelite assimilation of Canaanite cultural and religious ideas and practices began to be an acute problem and that other invaders and settlers became a threat to the security of Israel. One of the chief threats was from the Philistines, an Aegean people who settled (*c.* 12th century BCE) on the coast of what later came to be called, after them, Palestine. Organized in a league of five cities, or principalities, the Philistines, who possessed a monopoly of iron implements and weapons, pushed eastward into the Canaanite hinterland and subjugated Israelite tribes, such as the Judahites and Danites, that stood in their way, even capturing the sacred ark from the famous shrine of Shiloh when it was brought into battle against them. The Philistine threat was probably the decisive factor in the emergence of a permanent political (but at first primarily military) union of all Israel under a king—what historians call the united monarchy (or kingdom).

The period of the judges

The united monarchy. The monarchy was initiated during the career of Samuel, a prophet of great influence and authority who was also recognized as a judge and is depicted in varying biblical accounts as either favouring or not favouring the reign of a human king over Israel (see also SAMUEL). In any case, he anointed Saul, a courageous military leader of the tribe of Benjamin, as king (*c.* 1020 BCE). Saul won substantial victories over the Ammonites, Philistines, and Amalekites, leading the tribes in a "holy war," and for a time the Philistine advance was stopped; but Saul and his son Jonathan were killed in a disastrous battle with the Philistines in central Palestine (see also SAUL). His successor, David, a former aide (and also his son-in-law) who had fallen out of favour with him, at first took over (*c.* 1010) the rule of Judah in the south and then of all Israel (*c.* 1000). Through his military and administrative abilities and his political acumen, David established a centralized rule in Israel, cleared the territory of foreign invaders, and, in the absence of any aggressive foreign empire in the area, created his own petty empire over neighbouring city-states and peoples. He established his capital in Jerusalem, which until then had maintained its independence as a Canaanite city-state wedged between the territories of Saul's tribe Benjamin and David's tribe Judah, and moved the ark there from the small Israelite town in which it had been stored by the Philistines, establishing it in a tent shrine. This felicitous combination of holy ark, political reign, and central city was to be hailed and proclaimed by future ages. Under David's successor, his son Solomon (reigned *c.* 961–922), Israel became a thriving commercial power; numerous impressive buildings were erected, including the magnificent Temple (a concrete symbol of the religio-political unity of Israel); a large harem of foreign princesses was acquired, sealing relations with other states; the country was divided into 12 districts for administrative, supply, and taxation purposes. Foreign cults set up to serve the King's foreign wives and foreign traders led to charges of idolatry and apostasy by religious conservatives. In the latter years of his reign, Solomon's unpopular policies, such as oppressive forced labour, led to internal discontent and rebellion, while externally the vassal nations of Damascus (Aram) and Edom staged successful revolts against his rule. The central and northern tribes, called Israel in the restricted sense, were especially galled by the oppressive policies, and soon after Solomon's death Israel split off to become a separate kingdom. The united monarchy thus became the divided monarchy of Israel (the northern kingdom) and Judah (the southern kingdom).

The centralization of state and religion in Jerusalem

FROM THE PERIOD OF THE DIVIDED MONARCHY THROUGH THE RESTORATION

The divided monarchy: from Jeroboam I to the Assyrian conquest. Jeroboam I, the first king of the new state of Israel, made his capital first at Shechem, then at Tirzah. Recognizing the need for religious independence from Jerusalem, he set up official sanctuaries at Dan and Bethel, at the two ends of his realm, installing in them golden calves (or bulls), for which he is castigated in the anti-northern account in the First Book of the Kings. Israel engaged in conflicts with Judah and, sometimes jointly with Judah, against foreign powers. At first there was great dynastic instability in the northern kingdom, until the accession of Omri (reigned *c.* 884–*c.* 872),

one of its greatest kings, who founded a dynasty that lasted through the reign of his two grandsons (to 842). Under Omri an impressive building program was initiated at the capital, Moab was subjugated (an event confirmed in an extrabiblical source, the Moabite Stone), and amicable relations were established with Judah. The Phoenician kingdom of Tyre was made an ally through the marriage of his son Ahab to the Tyrian princess Jezebel. Ahab (reigned *c.* 874–853 BCE)—unless the episode recounted in I Kings, chapter 20 actually took place four reigns later—fought off an attempt by Damascus, heading a coalition of kings, to take over Israel. Near the end of his reign, Ahab joined with Damascus and other neighbouring states to fight off the incursions of the great Assyrian Empire in their area. Peaceful relations were cemented with Judah through the marriage of Ahab's daughter (or sister) Athaliah to Jehoram, the son of the king of Judah (not to be confused with Ahab's son, Jehoram of Israel). But the establishment of a pagan Baal temple for Jezebel and her attempt to spread her cult aroused great opposition on the part of the zealous Yahwists among the common people. There was also resentment at the despotic Oriental manner of rule that Ahab, incited by Jezebel, exercised. She and her cult were challenged by Elijah, a prophet whose fierce and righteous character and acts, as illumined by legend, are dramatically depicted in the First Book of the Kings. In the reign of Ahab's son Jehoram, Elijah's disciple Elisha inspired the slaughter of Jezebel and the whole royal family, as well as of all the worshippers of Baal, thus putting a stop to the Baalist threat. Jehu, Jehoram's general who led this massacre, became king and established a dynasty that lasted almost a century (*c.* 842–745), the longest in the history of Israel.

Jezebel and the Baal cult: Elijah and Elisha

Meanwhile, in Judah, the Baal cult introduced by Athaliah, the queen mother and effective ruler for a time, was suppressed after a revolt, led by the chief priests, in which Athaliah was killed and her grandson Joash (Jehoash) was made king. In the ensuing period, down to the final fall of the northern kingdom, Judah and Israel had varying relations of conflict and amity and were involved in the alternative expansion and loss of power in their relations with neighbouring states. Damascus was the main immediate enemy, which annexed much of Israel's territory, exercised suzerainty over the rest, and exacted a heavy tribute from Judah. Under Jeroboam II (783–741) in Israel, and Uzziah (Azariah; 783–742) in Judah, both of whom had long reigns at the same time, the two kingdoms cooperated to achieve a period of prosperity, tranquillity, and imperial sway unequalled since Solomon's reign. The threat of the rising Assyrian Empire under Tiglath-Pileser III soon reversed this situation. When a coalition of anti-Assyrian states, including Israel, marched against Judah to force its participation, the Judahite king Ahaz (*c.* 735–720) called on Assyria for protection; the result was the defeat of Israel, which suffered heavily in captives, money tribute, and lost provinces, while Judah became a vassal state of Assyria. In about 721, after an abortive revolt under King Hoshea, the rump state of Israel was annexed outright by Assyria and became an Assyrian province; its elite cadre, amounting to nearly 30,000 according to Assyrian figures, was deported to Mesopotamia and Media, and settlers were imported from other lands. Thus, the northern kingdom of Israel ceased to exist. Its decline and fall were a major theme in the prophecies of Amos, Hosea, Isaiah, and Micah.

The final period of the kingdom of Judah. Meanwhile, the southern kingdom of Judah was to have another century and a half of existence before a similar and even grimmer fate befell it. Hezekiah (reigned *c.* 715–*c.* 686), who instituted a religious reform to return worship to a pure Yahwist form, also displayed political independence, joining a coalition of Palestinian states against Assyria. But the coalition was soon defeated, and Judah —with Jerusalem besieged—bought off the Assyrians, led by Sennacherib, with tribute. In the reign of Manasseh (692–638) there was a revival of pagan rites, including astral cults in the very forecourts of the temple of YHWH, child sacrifice, and temple prostitution; hence, he is usually portrayed as the most wicked of the kings of Judah. If he had any tendencies toward independence from Assyrian domination, they apparently were suppressed by his being taken in chains to Babylon, where he was molded into proper vassal behaviour, although one edifying and probably unhistorical biblical account reports his repentance and attempt at religious reform after his return to Judah. The great religious reform took place in the reign of his grandson Josiah (640–609) during a period when the Assyrian empire was in decline and was precipitated by the discovery of the Book of the Law during the restoration of the Temple. It was proclaimed by the king to be the Law of the realm, and the people pledged obedience to it. In accordance with its admonitions, the pagan altars and idols in the Temple were removed, rural sanctuaries ("high places") all the way into Samaria were destroyed, and the Jerusalem Temple was made the sole official place of worship. (For an identification of the law book with the legal portion of Deuteronomy, see below *Old Testament literature: Deuteronomy*.) Josiah also made an attempt at political independence and expansion but was defeated and killed in a battle with the Egyptians, the new allies of the fading Assyrian Empire. During the reigns of his sons Jehoiakim (*c.* 609–598) and Zedekiah (597–586), Judah's independence was gradually extinguished by the might of the new dominant Babylonian Empire under Nebuchadrezzar. The end came in 586 with the Babylonian capture of Jerusalem and the destruction of the principal buildings, including the Temple and the fortifications. The first deportation of Judahites to Babylon, during the brief reign of Josiah's grandson Jehoiachin in 597, was followed by the great deportation of 586, which was to be a theme of lament and remembrance for millennia to come. (Numerous Jews also migrated to Egypt during this troubled time.) Exhortations and prophecies on the decline and fall of Judah are to be found in Zephaniah, Nahum, Habakkuk, and Jeremiah (who played a significant role in the events), while the conditions and meaning of the exile are proclaimed by Ezekiel and Deutero-Isaiah (chapters 40–55 of Isaiah).

Josiah's reform and the Book of the Law

The Babylonian Exile and the restoration. The Babylonian Exile (586–538) marks an epochal dividing point in Old Testament history, standing between what were subsequently to be designated the pre-exilic and post-exilic eras. The Judahite community in Babylonia was, on the whole, more Yahwist in religion than ever, following the Mosaic Law, emphasizing and redefining such distinctive elements as circumcision and the sabbath and stressing personal and congregational prayer—the beginnings of synagogal worship. It is possible that they also reached an understanding of historical events (like that taught by the great pre-exilic and exilic prophets)—as the chastening acts of a universal God acting in history through Nebuchadrezzar and other conquerors. To this period is also ascribed the beginning of the compilation of significant portions of the Old Testament and of the organizing view behind it. In any event, it was from this community that the leadership and the cadres for the resurrection of the Judahite nation and faith were to come when Cyrus the Great (labelled "the Lord's anointed" in Deutero-Isaiah) conquered Babylon and made it possible for them to return (538). A contingent of about 50,000 persons, including about 4,000 priests and 7,000 slaves, returned under Sheshbazzar, a prince of Judah.

The first great aim was the rebuilding of the Temple as the centre of worship and thus also of national existence; this was completed in 515 under the administration of Zerubbabel and became the place of uninterrupted sacrificial worship for the next 350 years. The next task was to rebuild the walls of Jerusalem, which was undertaken by Nehemiah, a Babylonian Jew and court butler who was appointed governor of Judah and arrived in 444. Nehemiah also began religious reforms, emphasizing tithing, observance of the sabbath, and the prohibition against intermarriage with "foreign" women. This reform was carried through systematically and zealously by Ezra,

Rebuilding of the Temple and religious reform under Nehemiah and Ezra

a priest and scribe who came from Babylon about 400 BCE, called the people together, and read them the "book of the law of Moses" to bring them back to the strict and proper observance maintained in Babylon: circumcision, sabbath observance, keeping the feasts, and, to seal it all, avoiding intermarriage. (In this presentation, modern critical scholarship is being followed, placing Nehemiah before Ezra instead of the traditional sequence, which reverses the positions.) Haggai, Zechariah, and Malachi are the prophets of this restoration period. Ezra and Nehemiah are its narrators.

It was in this period that enmity between the Jews, or Judaeans, as they came to be called, and the Samaritans, a term applied to the inhabitants of the former northern kingdom (Israel), was exacerbated. It has been surmised that this goes back to the old political rivalry between Israel and Judah or even further back to the conflict between the tribes of Joseph and Judah. Scholars ascribe the exacerbation of enmity in the restoration period variously to the Samaritans' being excluded from participating in the rebuilding of the Temple; to Nehemiah's rebuilding of the walls of Jerusalem (regarded as a threatening act by the Samaritan authorities); or to the proscriptions of intermarriage by Ezra. The animus of the Jews against the Samaritans is frequently expressed in the biblical books dealing with the restoration (expressions perhaps engendered by later events), but the attitude of the Samaritans and a good deal else about them is not evident. At some time they became a distinct religious community, with a temple of their own on Mt. Gerizim and a Scripture that was limited solely to the Pentateuch, excluding the Prophets and Writings.

Old Testament history proper ends with the events described in the books of Ezra and Nehemiah. The books of Chronicles give all the preceding history, from Adam to the Babylonian sack of Jerusalem and the exile. The last two verses of the Second Book of the Chronicles are repeated in the first two verses of Ezra: God inspires Cyrus to send the Jews back to Jerusalem to rebuild the Temple. The Persian period of Jewish history ended with the conquest of Alexander the Great in 323 BCE to begin the Hellenistic era, in which some of the biblical (including apocryphal or deuterocanonical) writings were created (for Hellenistic Judaism, see JUDAISM, HISTORY OF).

IV. Old Testament literature

THE TORAH (LAW, PENTATEUCH, OR FIVE BOOKS OF MOSES)

Composition and authorship. The Torah, or Pentateuch (Five Scrolls), traditionally the most revered portion of the Hebrew canon, comprises a series of narratives, interspersed with law codes, providing an account of events from the beginning of the world to the death of Moses. Modern critical scholarship tends to hold that there were originally four books (Genesis, Exodus, Leviticus, and Numbers) resulting from the division into manageable scrolls—a so-called Tetrateuch—to which later was added a fifth scroll, or book, Deuteronomy. A theory, once widely held, that the Book of Joshua was originally integral with the first five books to form a Hexateuch (Six Scrolls) is now generally regarded as dubious.

The traditional Jewish and Christian view has been that Moses was the author of the five books, that "of Moses" means "by Moses," citing in support passages in the Pentateuch itself that claim Mosaic authorship. Since these claims, however, are written in the third person, the question still arises as to the authorship of the passages; *e.g.*, in Deuteronomy, chapter 31, verse 9: "And Moses wrote this law, and gave it to the priests . . . and to all the elders of Israel." The last eight verses of Deuteronomy (and of the Pentateuch), describing Moses' death, were a problem even to the rabbis of the 2nd century CE, who held that "this law" in the verse quoted refers to the whole Torah preceding it. There are also other passages that seem to be written from the viewpoint of a much later period than the events they narrate.

The documentary hypothesis. Beyond these obvious discrepancies, modern literary analysis and criticism of the texts has pointed up significant differences in style, vocabulary, and content, apparently indicating a variety of original sources for the first four books, as well as an independent origin for Deuteronomy. According to this view, the Tetrateuch is a redaction primarily of three documents: the Yahwist, or J (after the German spelling of Yahweh); the Elohist, or E; and the Priestly code, or P. They refer, respectively, to passages in which the Hebrew personal name for God, YHWH (commonly transcribed "Yahweh"), is predominantly used, those in which the Hebrew generic term for God, Elohim, is predominantly used, and those (also Elohist) in which the priestly style or interest is predominant. According to this hypothesis, these documents—along with Deuteronomy (labelled D)—constituted the original sources of the Pentateuch. On the basis of internal evidence, it has been inferred that J and E are the oldest sources (perhaps going as far back as the 10th century BCE), probably in that order, and D and P the more recent ones (to about the 5th century BCE). Genesis, Exodus, and Numbers are considered compilations of J, E, and P, with Leviticus assigned to P and Deuteronomy to D.

J, E, P, and D documents

The Yahwist, or J, is the master of narrative in biblical literature, who sketches people by means of stories. He takes his materials wherever he finds them, and if some are crude he does not care, as long as they make a good story. The book of Genesis, for example, contains the story of Abraham's passing off his wife as his sister, so if the king took her as a concubine he would honour her supposed brother instead of having her husband killed, a story told by J without any moralistic homily. Not given to subtle theological speculations, J nearly always refers to the Deity as YHWH, by his specifically Israelite personal name (usually rendered "the Lord" in English translations), though he is not hidebound and also employs the term Elohim ("God"), especially when non-Hebrews are speaking or being addressed. He presents God as one who acts and speaks like human persons, a being with whom they have direct intercourse. The Yahwist, however, has one very definite theological (or theo-political) preoccupation: to establish Israel's divinely bestowed right to the land of Canaan.

More reflective and theological in the apologetic sense is the Elohist, or E. No fragment of E on the primeval history (presented in the first 11 chapters of Genesis) has been preserved, and it is probable that none ever existed but that the Elohist began his account with the patriarchs (presented in the remainder of Genesis, in which the J and E strands are combined). The first passage that can be assigned to E with reasonable certainty is chapter 20 of Genesis, which parallels the two J variants of the "She is my sister" story noted above. Unlike these, it tries to mitigate the offensiveness of the subterfuge: though the patriarch did endanger the honour of his wife to save his life, his statement was not untrue but merely (deliberately) misleading. The Elohist is also distinct from the Yahwist in generally avoiding the presentation of God as being like a human person and treating him instead as a more remote, less directly accessible being. Significantly, E avoids using the term YHWH throughout Genesis (with one apparent exception), and it is only after telling how God revealed his proper name to Moses, in chapter 3 of Exodus, that he refers to God as YHWH regularly, though not exclusively. This account (paralleled in the P strand in chapter 6 of Exodus) is apparently based on a historical recollection of Moses' paramount role in establishing the religion of YHWH among the Israelites (the former Hebrew slaves). Also noteworthy is E's choice of the term prophet for Abraham and his characterization of a prophet as one who is an effective intercessor with God on behalf of others. This is in line with his speculations on the unique character of Moses as the great intercessor as compared with other prophets (and also with Joshua as Moses' attendant).

The Elohistic tone and stress

It is inferred from certain internal evidence that E was produced in the northern kingdom (Israel) in the 8th century BCE and was later combined with J. Because it is not always possible or important to separate J from E, the two together are commonly referred to as JE.

The third major document of the Tetrateuch, the Priestly code, or P, is very different from the other two. Its narrative is frequently interrupted by detailed ritual instructions, by bodies of standing laws of a ritual character, and by dry and exhaustive genealogical lists of the generations. According to one theory, the main author of P seems to have worked in the 7th century and to have been the editor who combined the J and E narratives; for his own part, he is content to add some brief, drab records—with frequent dates—of births, marriages, and migrations. The P material is to be found not merely in Leviticus but throughout the Tetrateuch, including the early chapters of Genesis and one of the creation accounts and ranging from the primeval history (Adam to Noah) to the Mosaic era. Like the Elohist, P uses the term Elohim for God until the self-naming of God to Moses (Exodus, chapter 3, in the P strand) and shows a non-anthropomorphic transcendent stress.

The Deuteronomist, or D, has a distinctive hortatory style and vocabulary, calling for Israel's conformity with YHWH's covenant laws and stressing his election of Israel as his special people (for a detailed consideration of D, see below *Deuteronomy*). To the Deuteronomist or the Deuteronomic school is also attributed the authorship of the former prophets (Joshua, Judges, Samuel, and Kings), which scholars call the "Deuteronomic history."

Other Pentateuchal theories. This documentary theory of the composition of the Pentateuch has been challenged by eminent 20th-century scholars who have offered alternative or additional methods of analysis and interpretation. Form criticism, for example, has stressed particular literary forms and the historical setting out of which they arose: the sagas, laws, legends, and other forms and the particular tribal or cultic context that gives them meaning. Tradition criticism centres on the pre-literary sources; *i.e.*, on the oral traditions and the circles out of which they originated as accounting for the variety of the materials in the Pentateuch. Archaeological criticism has tended to substantiate the reliability of the typical historical details of even the oldest periods and to discount the theory that the Pentateuchal accounts are merely the reflection of a much later period. The new methods of criticism have served to direct attention to the life, experience, and religion out of which the Pentateuchal writings arose and to take a less static and literal view of the constituent documentary sources; yet most scholars still accept the documentary theory, in its basic lines, as the most adequate and comprehensive ordering of the variegated Pentateuchal materials. The following presentation rests mainly on an analysis and interpretation of the literary sources. (See also EXEGESIS AND HERMENEUTICS, BIBLICAL.)

In any case, the five books that have come down in various texts and versions have been seen as a unit in the religious communities that preserved them. Their basic content may be divided thus: (1) beginnings of the world and man—the primeval history; (2) patriarchal narratives—from Abraham to Joseph; (3) Egyptian slavery and the Exodus; (4) the revelation and Covenant at Sinai; (5) wanderings and guidance in the wilderness (divisible into two separate sub-blocks, before and after Sinai); (6) various legal materials—the Decalogue, Covenant Code, and passages of cultic and Deuteronomic laws—interspersed in the narrative, which take up the greater portion of the Pentateuch.

Genesis. This book is called *Bereshit* in the Hebrew original, after its first word (and the first word of the Bible), meaning "In the beginning." It tells of the beginnings of the world and man and of those acclaimed as ancestors of the Hebrew people—all under the shaping action and purpose of God. The book falls into two main parts: chapters 1–11, dealing with the primeval history, and chapters 12–50, dealing with the patriarchal narratives; the latter section is again divisible into the story of Abraham, Isaac, and Jacob (chapters 12–36) and the story of Joseph (chapters 37–50), which may be treated as a unit of its own.

The primeval history. The Bible begins with the creation of the universe. It tells the story with images borrowed from Babylonian mythology, transformed to express its own distinctive view of God and man. Out of primary chaos, darkness, void, depths, and waters God creates the heaven and the earth and all that dwell therein—a coherent order of things—by his will and word alone. He says, "Let there be . . ." and there is. Actually, there are two creation accounts: the first (1–2:4), ascribed to P, simply gives a terse day-by-day account including the culminating creation of man, in the divine "image and likeness," followed by the primordial sabbath on the seventh day. The other (2:4–25), ascribed to J, starts with an arid wasteland and the creation of man (Adam), described specifically as being formed by God out of dust and made into a living thing by God blowing the breath of life into him. He and the woman (Eve) created for him out of his rib are put into a paradisal garden (Eden), especially created for them to till and to tend and to sustain life. The two are forbidden only to eat of the tree of the knowledge of good and evil on pain of death (there is also a tree of life in the middle of the garden). The cosmic setting and concern of the P account is thus followed by the human setting and concern of the J account. Creation is followed by temptation, disobedience, and fall and all that follows from that for the history of mankind. At the instigation of the serpent, the shrewdest of the beasts, who holds out the possibility of attaining godlike knowledge, the woman eats of the fruit of the tree of knowledge and gives some to her husband to eat also. Their distinction from beasts and children manifests itself immediately by a sense of modesty about exposing their bodies, and loincloths become the first products of the higher knowledge. The primal human couple are punished by God for their disobedience by being driven out of the idyllic garden into the world of pain, toil, and death.

The eating of the forbidden fruit

The reason given by YHWH to the divine beings is: "Behold, the man has become like one of us, knowing good and evil; and now, lest he put forth his hand and take also of the tree of life, and eat, and live for ever." These words apparently point back to the polytheistic mythology (the existence of divine, magical powers; the gods' jealousy of mankind; the tree of eternal life; etc.) from which the Yahwist drew his images and symbols explaining man's suffering, frustration, and limitation. In the biblical framework and rendering (and subsequent interpretation), the archaic stories and images acquire a different meaning, suitable to the idea of a transcendent deity and an imperfect mankind.

With the exile from the garden, human history and culture begins. In the story of Adam's sons, Cain and Abel, man has already become a herdsman and farmer, and also a murderer: again probably a reflection of older mythical material and, again, one that puts an emphasis on human sin and estrangement from God. In the story of the Flood that follows there are evident borrowings from the Mesopotamian stories of a flood sent by the gods to destroy mankind, but in the biblical account it is emphasized that man's extreme wickedness is the cause and that Noah is saved along with his family by God's deliberate choice because he is a righteous man. (In the flood story in the Babylonian Gilgamesh epic, by contrast, there is no apparent moral reason why the gods resolved to destroy mankind, and the only reason why the hero of the Flood and his kin are saved is that he is favoured by one of the gods, who tricks the others, including the chief god.) After the Flood, God blesses Noah and bestows on man the earth and the things on it for sustenance and makes a covenant with Noah and all creatures that he will never again unleash a world-destroying flood. The permanent order of the world is assured, and God's blessing and covenant make their first explicit appearance in the Bible.

The Flood and the Tower of Babel

In the story of the Tower of Babel, the final story in the primeval history, a primal unity of mankind in which there is only one language is shattered when, in their pride, men decide to build a city and a tower that will reach up to the heavens. YHWH again takes steps to check dangerous collaboration: He says (to the celestial council), "Come, let us go down, and there confuse their

language, that they may not understand one another's speech," and scatters them over the earth. Again, the Yahwist has apparently used ancient mythological motifs to explain the diversity of mankind; the story may be regarded as simply a direct borrowing from the older traditions, without any monotheistic adaptation; in its textual setting, however, it may also be taken as another instance of the ruin of primal harmony by human willfulness and pride.

The fathers of the Hebrew people

The patriarchal narratives. The universal primal history of man in the first 11 chapters of Genesis is followed by an account of the fathers of the Hebrew people; *i.e.*, of the origins of a particular group. From a literary point of view, this portion may be divided into the sagas of Abraham, Isaac, and Jacob and the story of Joseph. Although these narratives are not historical in the ordinary sense, they have an evident historical setting and refer to various particulars that fit in with what is generally known of the time and area. They apparently rest on the traditions of particular families, clans, or tribes and were probably passed down orally before they took written form. Theologically, they are an account of a divine promise and Covenant and of man's faith and unfaith in response, with Abraham as the model man of faith (see also ABRAHAM).

The Elohist, as well as J and P, tells the remarkable story of how God singled out Abraham (Abram) to migrate from Mesopotamia and sojourn in Canaan, promised him that he would make him the ancestor of great nations and that his posterity would inherit the land of his sojournings, and singled out as the heirs to the latter promise first Isaac, Abraham's son by his chief wife Sarah, and then Jacob, the younger of Isaac's two sons; how Jacob acquired the additional name of Israel and how the wives, children, and children's children who, in Jacob-Israel's own lifetime, came to constitute a family of 70 souls, became the nucleus of the Israelite people; and how it came about that this ethnic group, prior to becoming, as promised, the masters of the land of their sojournings, first vacated it to sojourn for a time in Egypt. Apart from the low-keyed P strand, it is mostly splendid narrative, including the Elohist's account of the (aborted) sacrifice of Isaac by his father in response to God's command, a terse story packed with meaning, and the Joseph story about the son of Jacob who is sold into slavery by his brothers, rises to a high post in the Egyptian court, and ultimately helps his family to settle in Egypt. The twelve sons of Jacob-Israel are eponymous ancestors of Israelite tribes (ancestors after whom the tribes are named); the actions and fortunes of the eponymous ancestors, including certain blessings and other pronouncements of Jacob-Israel, account for the future positions and fortunes of the particular tribes. Though there is less history and more legend, much of the atmosphere of an older age is preserved, with the patriarchs represented as semi-nomadic, essentially peaceful and pastoral tent dwellers—alien residents—among the settled Canaanites and as observing customs otherwise only attested in Mesopotamia. Anachronistic features, however, insinuate themselves from time to time.

The God of the patriarchs

The God of the patriarchs is presented as Yahweh—explicitly by the Yahwist and implicitly by E and P—*i.e.*, as the same God who would later speak to Moses. God apparently was originally the personal, tutelary deity of each of the patriarchs, called by a variety of names and later unified into the one God of Abraham, Isaac, and Jacob. There are various cult legends in this portion of Genesis, etiological accounts of the origins of various cult sites and practices; though probably of Canaanite origin, these all indicate the places and customs held holy by the Israelites and perhaps also by their claimed Hebrew ancestors. There are direct appearances of God to some of the main figures in the narratives, intimate personal communication between men and God. God's particular blessing upon and Covenant with Abraham is the paradigmatic high point, to be referred back to continually in later biblical and post-biblical traditions.

Exodus. The title (in the Greek, Latin, and English versions) means "a going out," referring to the seminal event of the liberation of Israel from Egyptian bondage through the wondrous acts and power of God. The book celebrates and memorializes this great saving event in song and story and also the awesome revelation and covenant at Mt. Sinai. The contents of the book may be summarized thus: (1) Israel in Egypt, (2) the Exodus and wanderings, (3) the Covenant at Sinai, (4) the apostasy of the people and renewal of the Covenant, and (5) the instructions on building the Tabernacle and their execution.

The God of the Exodus

Redemption and revelation. Significant in the early chapters is God's special concern for the Hebrew slaves, his reference to them as "my people," and his revelation to Moses, the rebel courtier whom he has picked to be their leader, that he is YHWH, the God of their fathers, an abiding presence that will rescue them from their misery and bring them into Canaan, the land of promise. This assurance is repeated at the critical moments that follow (*e.g.*, "And I will take you for my people, and I will be your God"). In the series of frustrations, obstacles, and redeeming events that are narrated, God's special causal power and presence are represented as being at work. God hardens the Pharaoh's heart, sends plagues that afflict the Egyptians but spare the Hebrews, causes the waters to recede in the Sea of Reeds (or Papyrus Marsh) to permit passage to the fleeing Israelites and then to engulf the pursuing Egyptians ("the horse and his rider he has thrown into the sea"), and gives the people guidance in their wandering in the wilderness. The cryptic "name" that God gives to himself in his revelation to Moses (*'ehye 'asher 'ehye*), often translated "I am that I am" or "I will be what I will be," may also be rendered "I will cause to be that which I will cause to be." In either case, it is a play on, and an implied interpretation of, the name YHWH.

The constancy of God's directive power and concern is displayed notably in the period (40 years) of wilderness wandering (on the eastern and southern borders of Canaan), when Israel is tested and tempered not only by hardship but also by rebellious despair that looks back longingly to Egyptian bondage (see also below *Numbers*). God sends the people bread from heaven (manna) and quail for their sustenance (J and P strands) and, through Moses, brings forth hidden sources of water (JE strand). When the Amalekites (a nomadic desert tribe) attack, Moses, stationed on a nearby hill, controls the tide of battle by holding high the rod of God (a symbol of divine power), and when the enemy is routed he builds an altar called "The Lord is my banner" (E strand). Also inserted here is the account (E) of the visit of Moses' father-in-law, Jethro, a priest of another people (Midianite) who, impressed by YHWH's marvellous deliverance of Israel, blesses, extols, and sacrifices to him—under the name Elohim, but in the context the same God is clearly meant.

The Covenant at Mt. Sinai

God's power and presence manifest themselves impressively in the culminating account of the Covenant at Mt. Sinai (or Horeb). The people, forewarned by God through Moses, agree beforehand to carry out the terms of the Covenant that is to be revealed, because God has liberated them from Egypt and promises to make them his special holy people; they purify themselves for the ensuing Covenant ceremony, according to God's instructions. Yahweh appears in fire and smoke, attended by the blare of a ram's horn at the top of the mountain, where he reveals to Moses the terms of the Covenant, which Moses then passes on to the people below. Here follow in the text the Ten Commandments and the so-called Covenant Code (or Book of the Covenant) of lesser, specific ordinances, moral precepts, and cultic regulations, accompanied by a promise to help the people conquer their enemies if they will serve no other gods. After this comes the Covenant ceremony with burnt offerings and the sacrifice of oxen, with the blood of the animals thrown both on the altar and on the people to sacramentally seal the Covenant, followed by a sacral meal of Moses and the elders at the mountaintop, during which they see God. Many modern scholars hold that this is presented as the initial form of a Covenant renewal ceremony that was repeated either annually or every seven years in ancient Israel.

The stone tablets inscribed by God and Moses

There are certain problems and apparent discrepancies in this account that are explained by critical scholarship as deriving from the combination of different sources, mainly J and E, traditions, or emphases. In the opening portion (chapter 19) the people are gathered at the foot of the mountain so as to hear and meet God, and Moses himself brings down to them God's words. In a later portion (24:12–18, also 32:15–20), after the sacral meal, Moses goes up on the mountain to receive "the tables of stone, with the law and commandments," inscribed by God himself, and returns with two stone tablets written on both sides by the hand of God—which he breaks in anger at the people's worship of the molten calf that has developed in his absence. Later (chapter 34), at God's command, Moses cuts two new stone tablets, upon which after hearing God's various promises and exhortations, he writes "the words of the covenant, the ten commandments"; finally, he brings the new tablets down to the people and tells them what YHWH has commanded. There seem to be two parallel accounts of the same event, woven together by the skillful redactor into a continuing story. There also seem to be two distinct strands in the account of the sealing of the Covenant in the first 11 verses of chapter 24. According to one, the elders are to worship from afar, and only Moses is to come near YHWH; in the other strand, as noted the elders eat the sacred meal on the mountain top in the direct presence of God.

Legislation. The book of Exodus includes not only the narrative and celebration of God's redemptive action in the Exodus and wanderings and his revealing presence at Mt. Sinai but also a corpus of legislation, both civil and religious, that is ascribed to God and this revelation event. The Covenant Code, or Book of the Covenant, presented in chapters 20–23, immediately following the Decalogue (Ten Commandments), opens with a short passage on ritual ordinances, followed by social and civil law applying to specific situations (case law), including the treatment of slaves, capital crimes, compensation for personal injuries and property damage, moneylending and interest, precepts on the administration of justice, and further ritual ordinances. Scholars generally date this code in the later agricultural period of the settlement in Canaan, but some hold that it is analogous to more ancient Near Eastern law codes and may go back to Moses or to his time. In any case, it seems to be a compilation from various sources, inserted into and breaking the flow of the narrative.

Instructions on the Tabernacle. Also interspersed in the story (chapters 25–31) are God's detailed instructions to Moses for building and furnishing the Tabernacle, the clothing and ordination of priests, and other liturgical matters. According to this segment (evidently P in inspiration), an elaborate structure is to be set up in the desert, in the centre of the camp, taken apart, transported, and assembled again, like the simple "Tent of Meeting" outside the camp, where Moses received oracular revelations from God. Indeed, the two concepts seem to have fused and the Tabernacle is also called the Tent of Meeting. Its prime function is to serve as a sanctuary in which sacrifices and incense are offered on altars and bread presented on a table; it is also equipped with various other vessels and furnishings, including a wooden ark, or cabinet, to contain the two tablets of the Covenant—the famous ark of the Covenant. It is, moreover, to be the place of God's occasional dwelling and meeting with the people. Scholars believe that the elaborate details and materials described stem from a later, Canaanite, period but that the essential concept of a tent of meeting goes back to an earlier desert time. An account of the execution of the instructions for the building of the Tabernacle is presented in chapters 35–40 (following the apostasy, tablet breaking, and Covenant-renewal episodes), which duplicates to the letter the instructions in chapters 25–31. After the Tabernacle is completed and consecrated, it is occupied by the "glory," or presence, of YHWH, symbolized by a cloud resting upon it. It is on this note that the book of Exodus ends.

The portable "Tent of Meeting" and elaborate sanctuary

Leviticus. The cultic and priestly laws presented in Exodus are expanded to take up virtually the whole of Leviticus, the Latin Vulgate title for the third of the Five Books of Moses, which may be translated the Book (or Manual) of Priests. With one exception (chapters 8–10), the narrative portions are brief connective or introductory devices to give an ostensibly narrative framework for the detailed lists of precepts that provide the book's content. The source of Leviticus, both for the legal and narrative passages, is definitely identified as P; it is the only book in the so-called Tetrateuch to which a single source is attributed. Apparently the book consists of materials from various periods, some of them going back to the time of Moses, which were put together at a later date, possibly during or after the Babylonian Exile. Recent scholarship tends to emphasize the ancient origin of much of the material, as opposed to the previous tendency to ascribe a late, even post-exilic date. Despite its content and its dry, repetitive style, many interpreters caution against taking Leviticus as merely a dull, spiritless manual of priestly ritual, holding that it is strictly inseparable from the ethical emphasis and spiritual fervour of the religion of ancient Israel. It is in Leviticus that the so-called law of love, "You shall love your neighbour as yourself," first appears. The rituals set forth drily here probably presuppose an inward state in offering to God, as well as humanitarian and compassionate ethics.

The book may be divided thus: chapters 1–7, offerings and sacrifices; chapters 8–10, inauguration of priestly worship; chapters 11–16, purification laws; chapters 17–26, holiness code; chapter 27, commutation of vows and tithes.

Offerings, sacrifices, and priestly worship. The first verse attributes these regulations to YHWH, who speaks to Moses from the Tent of Meeting, beginning with the rules for offerings by the individual layman. These include burnt, cereal, peace, sin, and guilt offerings, all described in precise details. The prescription for priestly offerings is about the same, with some slight differences in the order of actions, and is presented much more briefly. In chapters 8–10 the narrative that was interrupted at the end of Exodus is resumed, and the ordination of Aaron and his sons by Moses, before the people assembled at the door of the Tent of Meeting is described, as are various animal sacrifices by Aaron and his sons under Moses' direction and the subsequent appearance of God's "glory" to the people. Aaron's two older sons are burned to death by fire issuing forth from God because they have offered "unholy fire." This story apparently emphasizes the importance of adherence to the precise cultic details, as does also the account (at the end of the chapter) of Moses' anger at Aaron's two remaining sons for not eating the sin offering. These stories were apparently used by the priestly authors to buttress the authority of the Aaronic priesthood.

The Aaronic priesthood

Purification laws. With chapter 11 begin the regulations on ritual cleanness and uncleanness, starting with animals and other living things fit and unfit to eat—the basis of the famous Jewish dietary laws. Then come the uncleanness and required purification of women after childbirth, skin diseases, healed lepers, infected houses, and genital discharges. Chapter 16, which belongs in the narrative flow immediately after chapter 10, describes the priestly actions on the Day of Atonement, the culmination of ritual cleansing in Israel. It is a chapter rich in details on Israelite ritual and bound up with the salient religious theme of atonement.

The Holiness Code. Next (chapters 17–26) comes what has been designated the "Holiness Code," or "Law of Holiness," which scholars regard as a separate, distinctive unit within the P material (designated H). It calls upon the people to be holy as God is holy by carrying out his laws, both ritual and moral, and by avoiding the polluting practices of neighbouring peoples; and it proceeds to lay down laws, interspersed with exhortations, to attain this special holiness. Although many scholars tend to date its compilation in the exilic period, some see evidence that it was compiled in pre-exilic times; in any case, the consensus is that the laws themselves come from a much earlier time.

These—a most miscellaneous collection—begin with in-

junctions on the proper (kosher) slaughtering of animals for meat; go on to a list of precepts against outlawed sexual relations (incest, homosexuality) and an injunction against defiling the (holy) land; proceed to a list of ethical injunctions, including the law of love and kindness to resident aliens, all interspersed with agronomic instructions and warnings against witchcraft; and then, after an injunction against sacrificing children, return to the listing of illicit sexual relations and the warning that the land will spew the people out if they do not obey the divine norms and laws. There follow special requirements for preserving the special holiness of priests and assuring that only unblemished animals will be used in sacrifices; instructions on the observance of the holy days—the sabbath, feasts, and festivals; commands on the proper making of oil for the holy lamp in the Tent of Meeting and of the sacred shewbread, to which are appended the penalties for blasphemy and other crimes; and finally, rules for observance of the sabbatical (seventh) and jubilee (50th) years, in which the land is to lie fallow, followed by rules on the redemption of land and the treatment of poor debtors and Hebrew slaves.

Punishment and atonement for iniquity

This miscellany, presented in chapters 17–25, is followed by a final exhortation, in chapter 26, promising the people that if they follow these laws and precepts, all will go well with them but warning that if they fail to do so all kinds of evil will befall them, including exile and the desolation of the Promised Land. Yet, if they confess their iniquity and atone for it, God will not destroy them utterly but will remember his Covenant with their forebears. Such a passage points to a later time but not necessarily to the exilic period, as some commentators have assumed. The chapter concludes: "These are the statutes and ordinances and laws which the Lord made between him and the people of Israel on Mt. Sinai by Moses," connecting these precepts with the primal revelation in Exodus.

Commutation of vows and tithes. In the final chapter of Leviticus (27), the P material is resumed with a presentation of the rules for the commutation of votive gifts and tithes. It provides for the release from vows (of offerings of persons, animals, or lands to God) through specified money payments. Some commentators understand the vow to offer persons to refer originally to human sacrifice, others as pledging their liturgical employment in the sanctuary. Special provisions are made for the poor to relieve them from the stipulated payments. Only grain and fruit tithes, not animal tithes, are redeemable. This chapter and the book of Leviticus end, like chapter 26, with the verse, "These are the commandments which the Lord commanded Moses for the people of Israel on Mount Sinai."

Numbers. In the Hebrew Bible this book is entitled Bemidbar (In the Wilderness) after one of its opening words, while in English versions it is called Numbers, a translation of the Greek Septuagint title Arithmoi. Each of the titles gives an indication of the content of the book: (1) the narrative of "40 Years" of wanderings in the wilderness, or desert, between Sinai and Canaan; and (2) the census of the people and other numerical and statistical matters, preceding and interspersing that account. It is a composite of various sources (J, E, and predominantly P) and traditions, which as a whole continue the story of God's special care and testing of his people in the events of the archaic period that formed them. Numbers continues the account of what many modern scholars call the "salvation history" of Israel, which apprehends and narrates events (or the image and impact of events) as involving divine action and direction.

The wilderness wanderings and the census of the people

The book may be divided into the following sections: (1) the conclusion of the Sinai sojourn (1:1–10:10), covering 20 days; (2) the wanderings in the desert of Paran (10:11–20:13), covering 38–40 years; and (3) the events in Edom and Moab (20:14–36:13), covering five months.

The conclusion of the Sinai sojourn. The book opens with a command from God to Moses, early in the second year after the Exodus, to take a census of the arms-bearing men over 20 in each of the clans of Israel. Moses and Aaron, aided by the clan chiefs, take the count, clan by clan, and reach a total of 603,550 men—according to critical scholars, an unbelievably large total for the time and conditions. The Levites, to whom is entrusted the care of the Tabernacle and its equipment, are exempted from this secular census and are counted in a later census, of males one month and over, along with a census of firstborn males from other tribes. The Lord had required that the latter be consecrated to him when he slew all the firstborn of the Egyptians but spared those of the Israelites; now the bulk of them were released by the Levites being taken in their stead to minister to the priests, while for the excess of firstborn over Levites "redemption" payments were collected. A further census of men 30–50 years old is taken among the Levite clans, so as to assign them their various duties, which are here stipulated. Also specified are the positions of the tribes (separated into four divisions of three tribes each) in the camp and on the march, with an assignment of specific portions of the Tabernacle and its equipment to be carried by the Levite clans. YHWH is to give the signal to break camp by lifting the cloud by day or the fire by night from above the Tabernacle and then to advance it in the direction the people are to march. YHWH's signal is to be followed by a blast by the priests (Aaron's sons) on two specially made silver trumpets.

YHWH's signal: a cloud by day and a fire by night

The above directions are set forth in chapters 1–4 and 9–10 (through verse 10). There are intervening chapters containing various materials: expelling leprous or other unclean persons from the camp, the ordeal for a woman suspected of adultery, regulations for Nazirites (those who take special ascetic vows), the offerings brought at the dedication of the Tabernacle, and the purification of the Levites preparatory to taking up their special sacred functions. The priestly emphasis of the materials in chapters 1–10 is evident, and it is also clear that there are various strands of priestly interpretation involved.

Wanderings in the desert of Paran. This section apparently combines various traditions of how the Israelites came into Palestine, and J, E (or JE), and P sources have been discerned in these chapters. The traditional "40 years" in the wilderness (38 or 39, according to critical calculations) were spent mostly in the wilderness of Paran, with a short stay in the oasis of Kadesh, according to P; while, according to J, they spent most of their time in Kadesh; and chapter 13, verse 26, puts Kadesh in the wilderness of Paran, thus encapsulating both traditions. The discrepancy may stem from two separate traditions of how the tribes entered Canaan: from the south or from the north through Transjordan.

The P narrative begins (chapter 10, verse 11) with the lifting of the cloud from the Tabernacle and the setting out of the Israelites for the Promised Land, with their holy Tabernacle and ark, in the order prescribed in chapter 2. According to the P account (verses 11–28), the cloud settles down over the wilderness of Paran, the signal to make camp; whereas in the JE account (verses 29–36) it is the ark of the Covenant that goes ahead to seek out a stopping place, and where it stops the Israelites rest, the cloud simply accompanying them overhead (perhaps to shield them from the blazing desert sun). Chapters 11–12 (JE) deal with the complaints of the people about their hardships and the rebellion of Miriam and Aaron against their brother Moses. When the people express their longing for the good food they had in Egypt and their disgust with the unvarying manna, God sends them a storm of quail, which remain uneaten because he also sends them a plague. This is a somewhat different account from that in Exodus, but the point is the same: the mighty, infinite power of God (chapter 11, verse 23). (Also inserted here is the story of God visiting his spirit on 70 selected elders so that they may share Moses' burdens.) When Miriam and Aaron question God's speaking only through Moses, God proclaims his unique relation with Moses, who alone receives direct revelations from God, not indirectly through dreams and visions, like the prophets.

Chapters 13–14 tell of the despatch of spies from Paran to reconnoiter Canaan and of the despair, rebellion, and unsuccessful foray of the people in response to the spies'

The spies' report and the people's rebellion

reports. Scholars discern two separate accounts of the spying incident artfully woven together. According to the JE account, the spies go only as far as Hebron in the south and return with a glowing report of a fertile land, which is, however, they warn, too strongly defended to be taken from that quarter: only one spy, Caleb, advocates attacking it. In the P account the spies reconnoiter the whole country and give a pessimistic report of it as a land that "devours its inhabitants," who are, moreover, giants compared to the Israelites. The people cry out in despair at this report and want to go back to Egypt, while Caleb and Joshua (added by P) plead with them to trust in God and go forward to take the land. God, disgusted with the people, condemns them to wander in the wilderness for 40 years and decrees that only their children, along with Caleb and Joshua, shall enter into the land of promise. Ruefully, the people now decide to attack and go forth, against Moses' warning, to a resounding defeat.

Chapter 15 is a P document or addition, setting forth various ritual regulations. Chapters 16–18 deal with the comparative rights and duties of priests and Levites. Chapter 16 is a composite document dealing with revolts against Moses and Aaron by certain Levites who question their special authority in a community where all are holy, as also by certain Reubenites who resent Moses' leadership. The dispute is settled when 250 revolting Levites attempt to offer incense (a priestly Aaronic function) and are consumed by fire sent by God, while the leaders of the revolt are swallowed up in the earth. Yet the stubborn people continue their complaint against Moses and Aaron, bringing forth the Lord's anger and a plague, from which they are saved by Aaron's (proper and effective) offering of incense. This latter incident occurs in chapter 17 in the Hebrew text and Jewish translations but concludes chapter 16 in some Christian versions. Chapter 17 in both arrangements, with its story of Aaron's rod, associates Levitical with Aaronic authority; Aaron's name is inscribed on the staff of Levi, which alone among the staffs of the chiefs of the tribes of Israel blossoms and bears fruit, thus authenticating Aaron's, and thereby the Levites', special claims. The relative functions and payments (tithes) of priests and Levites are prescribed in chapter 18. Chapter 19, inserted here, has to do with purification from uncleanness incurred through touching the dead, accomplished through washing in water mixed with the ashes of a red heifer.

Events in Edom and Moab. Chapter 20, verse 14, resumes the narrative of Israel's onward march, starting with their arrival in the wilderness of Zin and stay at Kadesh, marked by Miriam's death and God's exclusion of Moses and Aaron from entering the Promised Land because of their ascribed lack of confidence in God when Moses drew forth water from a rock in response to still more Israelite complaints, but did so in anger and impatience, striking the rock twice with his rod, instead of telling it to give forth water, as the Lord had instructed (the incident of the waters of Meribah). Refused permission by the King of Edom to pass through that land, over the much-used King's Highway, they proceed from Kadesh to Mt. Hor, where Aaron dies and is succeeded by his son Eleazar, and from which they proceed (chapter 21) to bypass Edom in an attempt to approach Canaan from the east. Arrived at the border of what was geographically part of Moab but politically the Amorite kingdom of Sihon, they are refused passage and proceed to defeat the Amorites and take possession of their land. This is from the JE strand of the composite narrative; the P strand does not recognize the existence of settled and politically organized populations between Kadesh and the plains of Moab.

Balaam, the pagan prophet of YHWH

At this point, in chapters 22–24, apparently a very mixed composite of various J and E strands, is presented the fascinating story (or collection of stories) of the non-Israelite seer, or prophet, Balaam, from the region of the Middle Euphrates. Alarmed at the Israelite host encamped at his border, the King of Moab commissions the seer Balaam to put a curse on them, but Balaam refuses, at the order of YHWH, who is also the God of Balaam. On three occasions at the King's request Balaam seeks an oracle from God against Israel, but each time, to the King's rage, he is told by the Lord that Israel is graced with the divine blessing and cannot be cursed. The seer, who is ordered back to his own country, without payment by the disgruntled King, offers a final, unsolicited oracle prophesying the destruction of Moab and other nations by Israel's might: "I will let you know what this people will do to your people in the latter days."

Chapter 25 (combining JE and P strands) provides a lurid interlude in which the Israelites go whoring after Moabite women and offer sacrifices and worship to their god, Baal of Peor. Phinehas, the son of Eleazar, is so incensed at the sight of an Israelite consorting with a Midianite woman that he kills them both, thus ending a plague that has broken out and earning God's special favour: a covenant of perpetual priesthood with him and his descendants (a forward reference to the Zadokite priesthood of post-exilic times). This account is connected by the last two verses with God's call for Israel to harass and smite the Midianites (see below). After the plague ends, in the account (P) in chapter 26, a second census of arms-bearing men and of the Levites is taken, and again a fantastically large total, 601,730, is given, perhaps referring to a much later time. It is noted at the end that all of the previous 603,730 had died in the wilderness, as prophesied, except for Caleb and Joshua, who have been especially picked out by God. This census, coming at the end of the 40-year period of wilderness wanderings, is for the purpose of allotting lands to the various tribes and families. Hence the logical positioning of the passage (P) in the first 11 verses of chapter 27 assuring that a family may inherit through a daughter when there is no son and through a brother when there are no children and through the closest relative when there are neither.

At this point (chapter 27, verse 12) comes the impressive and poignant passage (also P) in which Moses ascends the heights, at God's bidding, to look over the Promised Land, which he is not to enter, and calls on God to appoint a leader to succeed him. At God's command, Moses selects Joshua, and before the priest Eleazar and the whole community he lays his hands on him and commissions him to lead Israel. It is noteworthy that Joshua is invested only with some of Moses' authority and is to learn God's will through Eleazar and the sacred lot (Urim), not directly, as did Moses.

Sacrificial rules for holy days

Again, the narrative is interrupted by three chapters (P) dealing with various religious regulations. Chapters 28–29 stipulate the sacrifices to be made by the whole community daily, on the sabbath, at the new moon, and on these holidays: the Feast of Unleavened Bread (Passover), the Feast of Weeks (Shavuot), The Feast of Trumpets, *i.e.*, New Year (Rosh Hashana), the Day of Atonement (Yom Kippur), and the Feast of Tabernacles (Sukkot). The last two verses of chapter 29 specify that these public offerings are in addition to individual offerings, such as those specified in chapter 15. Critical scholars hold that these elaborate regulations stem from a much later (post-exilic) period, though they may go back to very ancient practices. Some see them as a liturgical commentary on chapter 23 of Leviticus, which presents the cycle of feasts and festivals (see above *Leviticus*). Chapter 30 gives women special exemption from keeping vows (presumably of offerings or abstinence) when countermanded by a father or husband; only widows or divorcees are bound, like men, unconditionally to keep their vows.

Chapter 31, likewise from P, deals with the annihilation of the Midianites following God's command at the end of chapter 25. The Israelites, a thousand from each tribe, go forth to battle led by the priest Eleazar, who carries the sacred vessels and the trumpets. They kill every man and seize all the movable property but spare the women and children. Moses, however, orders every male child and all nonvirgin women killed. There follow instructions for purification for the stain caused by killing a person or touching a dead body and for the distribution of the booty, which includes sheep, cattle, asses, and 32,000 virgins. The rules are that half of the spoils go to the fighting men, half to the rest of the people; in addition, the Lord's share is allotted thus: one five-hundredth of the

fighting men's portion goes to the priest, and one-fiftieth of the people's portion goes to the Levites. Scholars are inclined to treat this chapter as a piece of fiction intended really to set forth the rules for purification and dividing the spoils through an invented story. The seer-diviner Balaam is here (verse 16) blamed for the whoring and apostasy incidents in chapter 25; but texts providing his connection with these events are lacking.

Special settlement of Reuben and Gad east of the Jordan

Chapter 32, dealing with the settlement east of the Jordan, concludes the narrative portion of Numbers and thus of the Tetrateuch (a story that is continued in chapter 34 of Deuteronomy and in the Book of Joshua). This very composite account (JEP) tells how the tribes of Reuben and Gad, after an initial angry remonstrance from Moses, are granted permission to settle in the rich pasturelands east of the Jordan on the assurance that after they erect sheepfolds and fortified towns for their flocks and families, they will provide the shock troops spearheading the advance of the Israelites into Canaan, and will not return to their homes until their brethren hold the land. Thereupon Moses allots the various conquered kingdoms and towns east of Jordan to the Gadites and Reubenites. The various Gadite, Reubenite, and Manassite towns are listed.

The rest of the book of Numbers (P in its final form) consists of an itemized summary of the route from Egypt to the plains of Moab outside Canaan (chapter 33) and various additional materials (chapters 34–36). Verses 50–56 of chapter 33 present the divine command to dispossess the people of Canaan, destroy their idols and cultic places, and apportion the land to each clan by lot. In chapter 34 the Lord specifies the boundaries of the whole land of Canaan that is to be Israel's inheritance and names the tribal leaders who, along with Eleazar and Joshua, are to oversee the division of the land by lot. In chapter 35, the Lord orders 48 towns with extensive pasturelands to be set aside for the Levites; six of these are to be cities of refuge for manslayers whose guilt of intentional murder has not yet been determined and who are provided sanctuary from the traditional blood vengeance. Although these settlements do not constitute an independent tribal territory but are scattered through the territories of the other tribes, the contradiction with chapter 18, verse 24, of Leviticus, commanding that the Levites are to have no share of the land but are to subsist solely on tithes, is obvious and raises critical questions. Finally, chapter 36 concludes the book of Numbers with a supplement to the law of inheritance through daughters laid down in chapter 27, enjoining daughters from marrying outside the tribe, so that the tribe will hold its portion of the land, which was given from God, in perpetuity. As before, the general injunction is laid down in a story dealing with a particular case (the daughter of Zelophehad).

Deuteronomy. *Special nature and problems.* The English title of this work, meaning "second law," is derived from a faulty Greek translation of chapter 17, verse 18, referring to "a copy of this law": the implication being that the book is a second law or an expanded version of the original law for the new generation of Israelites about to enter Canaan. Hebrew texts take the opening words of the book as title, Ele ha-Devarim (These Are The Words), or simply Devarim (Words). As noted in *Composition and authorship*, above, the book is in a class by itself in the Pentateuch, so much so that modern scholars tend to consider it apart from the other four books, and some see it in style, content, and concerns more closely related to the succeeding books of Joshua, Judges, Samuel, and Kings, constituting a "Deuteronomic history." In spite of its homogeneous style and tone—it is assigned for the most part to a single source, D—the content indicates to critical scholars very composite traditions, ages, and situations behind the finished form. This book has elicited a library of scholarship going back to the early 19th century, not only because of the complicated critical and historical problems calling for solution but also because of its spiritual and theological message, which gives it a special place among Old Testament writings.

In form, the book is ostensibly a discourse by Moses "to all Israel" in the final month in Moab before they go over the Jordan into Canaan. Actually it comprises three separate discourses, a set of laws, two poems, and various other matters, all ascribed to Moses directly—here it is Moses who sets forth the laws, not God through him. These materials are centred on the presentation of the rules of life and worship for the coming stay in the Promised Land, along with exhortations and explanations pointing to YHWH, the marvellous liberator from Egypt and guide in the wilderness, as the divine source and reason for the commands. The traditional view was that, with the possible exception of the account of Moses' death, the whole book was written by Moses, based on the phrase "And Moses wrote this song" in chapter 31, verse 22.

Deuteronomy identified with Josiah's "book of the law"

Some early Church Fathers identified the book with "the book of the law" (II Kings, Chapter 22, verse 8), found in the 18th year of King Josiah's reign (*c.* 621 BCE), and made the basis of his great religious reform the following year. Wilhelm M.L. de Wette, a German biblical scholar, in 1805 established the predominant modern view that Deuteronomy (or its nucleus, or main portion) was found in Josiah's time and was a distinctive book, separate from the Tetrateuch. He also held that it was composed shortly before its discovery; other, more recent, scholars would put it as much as a century earlier and connect it with earlier reforms, while some associate it with the writings and teachings of the 8th-century-BCE prophet Hosea and with the E source. Furthermore, the references to localities near Shechem as cultic places, taken with certain passages in Joshua, indicate a northern provenance for the book and not the southern source connected with a cultic centre at Jerusalem, as had been previously supposed from the associated material in II Kings. Some scholars see the form and occasion of Deuteronomy as a Covenant renewal ceremony in which the whole law is read, as in Joshua, chapter 8, verses 30–35, and thus view it as a liturgical document, as well as a lawbook. In any case, the tendency is to see various layers of materials and lines of transmission, perhaps going back to quite early preliterary sources, before its final formation in the 8th or 7th century BCE.

The book may be divided as follows: (1) introductory discourse to the whole book (chapter 1 to chapter 4, verse 43); (2) introductory discourse to the lawbook (chapter 4, verse 44, through chapter 11); (3) the lawbook (chapters 12–28); (4) concluding exhortation and traditions about the last days and death of Moses (chapters 29–34).

First introductory discourse of Moses. The first introductory discourse, spoken by Moses, traces the journey of the Israelites from Mt. Horeb to Moab, with some noticeable differences in detail from the account in Exodus and Numbers and an emphasis on Moses being banned from entrance into the Promised Land because the Lord was angry at the Israelites. To this historical retrospect is appended an exhortation to the people to obey God's laws and norms, recalling the imageless God of the revelation and Covenant at Horeb as a warning against making images and serving man-made gods. The uniqueness and soleness of the God of the Exodus and Covenant, his power and presence in his marvellous acts of redemption and revelation, and his gracious selection of Israel are proclaimed in rhetorical questions; moreover, it is emphasized that the God of Israel ("YHWH your God") "is God in heaven above and on the earth beneath; there is no other." The injunctions against idolatry appear to come from later experience and religious crisis in Canaan. The fact that other nations have their own gods and objects of worship is recognized elsewhere in Deuteronomy.

Second introductory discourse. The second discourse, also ascribed to Moses, again refers to the Covenant at Horeb and sets forth the Ten Commandments, which the people are admonished to obey rigorously, emphasizing the mediating function of Moses at Horeb between the awesome divine presence and the awestruck people. Israel is further admonished to obey the law through wholehearted love of God, expressed in what became the cen-

The Shema: "Hear, O Israel"

tral liturgical expression of Israel's faith, beginning, "Hear, O Israel: The Lord is our God, the Lord Alone. You must love the Lord your God with all your heart and with all your soul and with all your might." If they obey God's laws, avoid other gods, and do what is right and good, they will possess the land promised by God—him who rescued them from Egypt and has brought them thus far. They are to avoid marriage and all other intercourse with the peoples of the land, utterly destroying them and their idolatrous altars and cultic places, for they are a special, holy people chosen by God out of all the peoples because of his love, not because of their greatness or power. This marvellous love will continue to be exercised, and the people will be blessed with all good things—prosperity, fertility, health, and success in battle—if they obey God's ordinances. They are urged to remember the 40-year period of wilderness wandering, in which they were tested (disciplined) by God through hardship and hunger (to find out whether or not they would keep his commands) and saved by him: man does not live by bread alone but, rather, by whatever God provides (*e.g.*, manna from heaven). Another time of testing will come when they live in the rich, fertile land of Canaan and eat their fill and perhaps forget the Lord and his laws, ascribing their wealth to their own power and might and even venturing into idolatrous worship of the gods of the land. If they do so they shall perish, just as the idolatrous nations of the land shall.

A long list of the apostasies of Israel is presented in chapter 9 to demonstrate the point that Israel is going in to possess the land of Canaan not through any virtue of their own but because of God's promise to the patriarchs. This is followed in chapter 10 by a moving declaration of what God requires of Israel—fear (reverence), walking in his ways, love, wholehearted service, and keeping his commandments—and an extolling of the wondrous, unique, powerful God who liberated them from Egypt. Chapter 11 extols the richness of the land of Canaan and describes how it will bloom for them if they are observant of God's commandments and promises that they will hold the territory from the wilderness to Lebanon and from the Euphrates to the western sea (Mediterranean). It closes with the choice set before them by Moses of "a blessing and a curse"—the former if they obey the commandments, the latter if they do not. This choice is posed to them immediately before the presentation of the laws and norms beginning in chapter 12.

The lawbook. The laws are the central core and purport of the book of Deuteronomy. They are couched in a hortatory, sermonic style that has led to their being categorized as preached law. Emphatic statements of what must or must not be done are connected with exhortations to fulfill these injunctions, pointing to the motivations and spirit in which they should be carried out. There is a wide variety of laws here—ritual, criminal, social—but they are all set within this preaching context and aimed at the service of God. This is no dry legal code but, rather, a book written in fluent and moving prose. Scholars have seen duplications and parallels between the laws presented here and those in the Covenant Code in chapters 21–23 of Exodus; but to this a common source may be ascribed, and Deuteronomy may be considered a work in its own right and not a mere expansion of the Covenant Code.

The basic injunction: a central sanctuary

The lawbook comprises chapters 12–26, supplemented by chapters 27–28. After an initial order to destroy the pagan cultic places and idols, the lawbook goes to its basic injunction: to set up a single central sanctuary in Canaan, where all Israel is to make their offerings, as distinct from the present unregulated practice, "every man doing whatever is right in his own eyes." The spot is designated only "the place which the Lord your God will choose," which some interpreters, following King Josiah, have understood to be Jerusalem and which others understand to be Shechem. (The blessing and curse passage immediately preceding in chapter 11 specifies Mts. Gerizim and Ebal, on either side of Shechem, as the places of blessing and curse, respectively; and an even more elaborate ritual is prescribed for the same locality in chapter 27.) Instructions are given for the proper killing of animals for food, previously connected with the sacrificial cult, and the people are admonished when they settle in Canaan not to inquire about how other nations serve their gods, possibly to follow their abominable practices. Inserted at this point is the striking exhortation, "Everything that I command you you shall be careful to do; you shall not add to it or take from it."

Chapter 13 warns the people to beware of the temptations to apostasy arising from the urging or example of prophet-diviners, kinfolk or friends, or a whole town; they are to kill the tempters and destroy the towns. Chapter 14 is devoted mainly to a list of living things that may or may not be eaten, the "clean" and "unclean," similar to the list in Leviticus, chapter 11; and to laws for tithes and first fruits to be brought annually to the central sanctuary and triennially to the Levites in the towns, who are specified as having no "portion" of their own (two years to the centre, the third year to the town Levites). Chapter 15 deals mainly with the releases to be granted every seventh year to debtors of their debts and Hebrew slaves of their bondage; lenders are exhorted and commanded not to refuse loans to the poor in the sabbatical year of release, and God's redemption of Israel from Egypt is given as the reason for freeing one's Hebrew slaves in the sabbatical release. The first section of chapter 16, verses 1–17, gives the rules for celebrating the three main festivals of the religious year: Unleavened Bread, Weeks, and Booths, which are to be observed at the central sanctuary (hence later called the three pilgrim festivals).

Beginning with verse 18 of chapter 16 there is a discussion of the appointment and character of judges, and of judicial procedures and punishments for apostasy, homicide, and other crimes; similarly, beginning with verse 14 of chapter 17 there are rules on the selection of a king and for his conduct, and the injunction that he read from "a copy of this law," so that he may be edified and chastened. The first portion of chapter 18 deals with the office and support of priests, referred to here as "the Levitical priests . . . all the tribe of Levi," not distinguishing the Aaronic priests from the lesser Levites. This is followed—after a passage inveighing against abominable cultic and divinatory practices of the nations of the land—by a promise that God will raise up prophets among the people and instructions on how to tell true from false prophets. Thus the offices of judge, king, priest, and prophet are considered in chapters 16–18.

Chapter 19 deals again with crime and punishment. It distinguishes between unintentional manslaughter and murder, setting up cities of refuge for the manslayer and ordering the murderer to be killed by the blood avengers. It also lays down the rules for witnesses and the punishment for perjury. It closes with the famous lex talionis: "Life for life, eye for eye, tooth for tooth, hand for hand, foot for foot," which in context may spell out what is to happen to the false witness and even could be interpreted as a moderating, rather than an inhumane, precept (no more than an eye for an eye, etc.). Chapter 20 gives the rules for holy war, listing the situations that exempt men from military service (*e.g.*, a newly married man) and distinguishing the treatment of non-Canaanite and Canaanite cities; the latter are to be utterly destroyed, yet it is forbidden to destroy fruit-bearing trees. There are also rules on holy war in 21:10–14; 23:9–14; 24:5; and 25:17–19. Chapters 20–25 contain a great variety of laws; the just treatment of women captives, sexual offenses, exclusions from the religious community, public hygiene in campgrounds, and many other things.

The lex talionis

The last of the laws are set forth in chapter 26, dealing with the first fruits offering and tithes. At the annual offering (or soon after entering Canaan), in the central sanctuary, the worshipper is to recite a piece beginning, "A wandering Aramaean was my father," affirming his link with the patriarchs and extolling God's wondrous deeds on behalf of Israel. And every third year he is to set aside his tithe "to the Levite, the sojourner, the fatherless, and the widow" and make an affirmation "before the Lord" that he has complied and avoided any ritual stain.

The final passage in chapter 26 proclaims that "this

day" God has proclaimed his law, Israel has affirmed its commitment to God and his law, and God has affirmed his choice of Israel as his special, holy people, to be set up high above all the nations. This is the hortatory conclusion to chapters 12–26 and to the "second law," or Covenant, contained therein.

Covenant ratification and renewal ceremonies

The emphasis on the laws given on "this day" is continued in the supplementary chapters 27–28, which deal with Covenant ratification and renewal ceremonies, apparently a reference to an original ceremony in Moab, one in Canaan on the first day in the land, and subsequent, possibly annual, renewal ceremonies. Blessings and curses are to be pronounced from Mts. Gerizim and Ebal for respectively fulfilling or disobeying the Covenant: all good things or all bad things will befall the people, as they keep or fail to keep the Covenant. Some of the curse consequences in chapter 28, referring to siege, subjugation, and exile, are believed by some scholars to reflect late pre-exilic or exilic situations. The curse consequences fill up the bulk of these chapters and are recounted in powerful, moving language, ending with a threat to return the people to Egypt.

Concluding exhortation and traditions about the last days of Moses. Chapters 29–31 comprise the third and last address of Moses to the people of Israel. They are preceded by an introductory verse referring to "these words" as a covenant made in Moab, in addition to the one made at Horeb (Sinai). After reminding them of all that God has done for them, Moses calls on the whole people to enter into the sworn Covenant made this day that they may be his people and he may be their God, warning the secret apostate of the calamities that will befall him. Yet the possibility of a return to God and the land is held out to those who will suffer exile and persecution as punishment for their apostasy, again presumably a reflection of the exilic situation (chapter 30 verses 1–10 seems clearly to be an interpolation inspired by the actual experience of exile). This law, it is emphasized, is no recondite, remote thing up in the sky but is, rather, very close to men, "in your mouth and in your heart"; what is revealed is made plain, it is not the secret things of God. Moses sets before them the classic Deuteronomic choice: "life and good" over "death and evil." The people are given that choice and told the consequences of loving the Lord and keeping the Covenant or of going the other way.

The last words and acts of Moses

The final chapters are concerned with the last words and acts of Moses: directing Joshua to lead Israel after his death, writing down "this law," calling for a sabbatical renewal ceremony of it on the Feast of Booths, ordering that it be put beside the ark of the Covenant, and uttering two poems. The first, "The Song of Moses" (chapter 32), praises the faithfulness and power of the Lord, decries the faithlessness and wickedness of Israel, and predicts the consequent divine punishment; it adds, however, that in the end the Lord will relent and will vindicate his people. The second poem, "The Blessing of Moses" (chapter 33), blesses each of the tribes of Israel, one by one, and the blessings are associated with God's love, the law commanded by Moses, and the kingship of God over his people. There are indications in both poems of a considerably later date (after Joshua's time, perhaps in the period of the Judges); Moses is spoken of in the third person in "The Blessing" poem.

The narrative of Deuteronomy, and thus of the Pentateuch, ends with Moses' ascent to the top of Mt. Pisgah, his being shown the Promised Land by God, and his death there in the land of Moab, buried by God in an unknown grave. It is emphasized in the closing words that Moses was a unique prophet "whom the Lord knew face to face" and through whom the Lord wrought unique "signs and wonders" and "great and terrible deeds." Thus end the Five Books of Moses. (S.C.)

THE NEVI'IM (THE PROPHETS)

The canon of the Prophets. The Hebrew canon of the section of the Old Testament known as the Nevi'im, or the Prophets, is divided into two sections: the Former Prophets and the Latter Prophets. The Former Prophets contains four historical books—Joshua, Judges, Samuel, and Kings; the Latter Prophets includes four prophetic works—the books of Isaiah, Jeremiah, Ezekiel, and the Twelve (Minor) Prophets. The Twelve Prophets, formerly written on a single scroll, include the books of Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, and Malachi. Thus, in the Hebrew canon of the Prophets there are, in effect, eight books.

Variations in the canon of the Prophets

The Christian canon of the Prophets does not include the Former Prophets section in its division of the Prophets; instead, it calls the books in this section Historical Books. In addition to Isaiah, Jeremiah, and Ezekiel, the Christian canon of the Prophets includes two works from the division of the Hebrew canon known as the Ketuvim (the Writings): the Lamentations of Jeremiah and the Book of Daniel. The Twelve (Minor) Prophets are separated into individual books. The number of works in the Christian canon, however, varies. The Protestant canon contains all the books of the Latter Prophets and the two books from the Ketuvim, thus listing 17 works among the prophetic writings. The Roman Catholic canon accepts one other book as a canonical prophetic work, namely, Baruch (including the Letter of Jeremiah); the number of prophetic writings in the Roman Catholic canon is, therefore, 18. The Greek Orthodox Synod of Jerusalem in 1672 did not accept Baruch as canonical.

As far as the Former Prophets is concerned, the Protestant canon, following the Septuagint, separates Samuel and Kings into two sections each: I and II Samuel, and I and II Kings. The Roman Catholic and Orthodox churches in the past divided these two works into I, II, III, and IV Kings, but most Roman Catholic translations now follow the listing as it is in the Septuagint.

Hebrew prophecy. Hebrew prophecy was rooted in the prophetic activities of various individuals and groups from the nations and peoples of the ancient Near East. Though prophecy among ancient Egyptians, Mesopotamians, and Canaanites—as well as among the peoples of the Aegean civilization—generally was connected with "foretelling" (or predicting) the future, the Hebrew view of prophecy centred on "forthtelling" (or proclaiming), though it included predictive aspects. Thus, in Hebrew prophecy the phrase "Thus says the Lord" is repeated constantly to emphasize the "forthtelling" motif. The Hebrew prophets were very conscious of the absolute holiness (separateness) of God and the purpose of God for his chosen people, Israel. Because of this consciousness, they developed an acute awareness of sin and its effects on man and society and, in consequence of such an awareness, a radical ethical outlook that applied to both the individual and the community.

Emphasis on "forthtelling"

The Hebrew term for prophet (*navi'*) is probably related etymologically to the Akkadian verb *nabū*, meaning "to call" or "to name." The Hebrew prophet may thus be viewed as a "caller," or spokesman, for God. Other designations for prophet in the Old Testament are *ro'e*, or "seer," and *ḥoze*, or "visionary," the two latter terms indicating that the predictive element was operative in Hebrew prophecy. The distinctive element of Hebrew prophecy, however, was the relationship of the prophet to God, the Lord of the Covenant, and to Israel, the covenant people. He spoke for the sovereign Lord to remind, cajole, castigate, reprove, comfort, and give hope to the people of the covenant, constantly reminding them that they were chosen to witness to the nations of the love, mercy, and goodness of God.

Some of the Hebrew prophets, from the 11th to the 8th century BCE, belonged to bands or guilds of ecstatic prophets. Such prophets were spokesmen for God whose uncontrollable actions and words caused them to be feared and, sometimes, held in contempt. In II Kings, chapter 9, verse 11, a prophet—who came to Jehu, the 9th-century-BCE army commander who became king of Israel, in order to anoint him—was called a "madman" (*meshugga'*). Other Hebrew prophets were more independent, such as Nathan and Elijah, though they continued to maintain the quality of being uncontrollable—at least as far as the political authorities were concerned.

Both of these early nonwriting prophets spoke out against the oppression of the weak by the strong, a theme that came to be expressed constantly in Judaism throughout the centuries. The activities of such early prophets, including also Micaiah and Elisha in the 9th century BCE, are described in the Former Prophets.

In the 8th century BCE, the writing prophets—*i.e.*, the Latter Prophets—began their activities. Though all the books that bear their names probably have been edited by schools of a prophet or by individuals or groups that were influenced by their ideas, the editors or disciples of the prophets preserved as well as was possible the words, activities, and idiosyncratic themes of the prophetic personalities. Some of the Latter Prophets may have been connected with the priestly class, such as Isaiah, Jeremiah, and Ezekiel; most of the Latter Prophets, however, were independent of priestly connections. All of the Latter Prophets stood out in contrast to the court prophets who, in the tradition of court prophets of most ancient Near Eastern peoples, seldom contradicted what they believed was expected of them by their sovereigns or the people (see also PROPHECY).

Joshua. The Book of Joshua takes its name from the man who succeeded Moses as the leader of the Hebrew tribes—Joshua, the son of Nun, a member of the tribe of Ephraim. In post-biblical times Joshua himself was credited with being the author of the book, though internal evidence gives no such indication. According to the views of the German biblical scholar Martin Noth, which have been accepted by many contemporary biblical critics, the Book of Joshua was the second of a series of five books (Deuteronomy, Joshua, Judges, Samuel, and Kings) written by a Judaean oriented historian after the fall of Jerusalem in 586 BCE. This writer (called the Deuteronomist and designated D) constructed the history of Israel from the death of Moses to the beginning of the Babylonian Exile (586–538 BCE). The Deuteronomist, according to this view, used sources, both oral and written, from various periods to produce the history of Israel in these five books. The Book of Joshua probably contains elements from the J and E documents, as well as local and tribal traditions, all of which were modified by additions and editing until the book assumed its present form. The main theme of the Deuteronomist historian was that under the guidance of and in obedience to Yahweh, Israel would persevere and conquer its many enemies.

The work of the Deuteronomic historian

This theme is especially and dramatically presented in Joshua. Under the guidance of Yahweh, the people of Israel entered and conquered Canaan in fulfillment of the promise of God to Abraham and his descendants in Genesis, chapter 12. Joshua is interpreted as a second Moses—*e.g.*, he sent out spies, led the people in crossing the Jordan River on dry land as Moses had crossed the Sea of Reeds, and ordered the males to be circumcised with flint knives as Zipporah, Moses' wife, had earlier circumcised the son of Moses (and probably Moses himself). He was obedient to the will of Yahweh, and because of this obedience he was able to lead the Israelite tribes in their battles against the Canaanites. As long as they were faithful to their covenant promise, the land would be theirs as a trust.

The book may be divided into three parts: the story of the conquest of Canaan (chapters 1–12); the division of the land among the tribes of Israel (chapters 13–22); and Joshua's farewell address, the renewal of the Covenant, and Joshua's death (chapters 23–24).

The conquest of Canaan. As told by the Deuteronomist, the conquest of Canaan by Joshua and the Israelite tribes was swift and decisive. No conquest of central Canaan (in the region of Shechem), however, is mentioned in the book; and some scholars interpret this to mean that the central hill country was already occupied either by ancestors of the later Israelite tribes prior to the time of Moses or by portions of Hebrew tribes that had not gone to Egypt. Because these people made peace with the tribes under Joshua, a conquest of the area apparently was not necessary. Archaeological evidence supports portions of Joshua in describing some of the cities (*e.g.*, Iachish, Debir, and Hazor) as destroyed or conquered in the late 13th century BCE, the approximate time of the circumstances documented in Joshua. Some of the cities so reported, however, apparently were devastated at some time prior to or later than the 13th century. Jericho, for example, was razed at the end of the Middle Bronze Age (*c.* 1550 BCE) and most likely had not been rebuilt as a strongly fortified town by the time of Joshua, though the site may well have been inhabited during this period. The city of Ai was destroyed about 600 years before; but it may have been a garrison site for the city of Bethel, which was destroyed later by the "house of Joseph." Though many of the cities of Canaan were conquered by the Israelites under Joshua, historical and archaeological evidence indicates that the process of conquering the land was lengthy and not completed until David conquered the Jebusite stronghold of Jerusalem in the early 10th century BCE. At any rate, the 13th century was an ideal time for a conquest of the area because of the international turmoil involving the great powers of the time: Egypt and Babylonia. A political vacuum existed in the area, permitting small powers to strengthen or to expand their holdings.

Archaeological evidence supporting the biblical narrative

The introductory section of Joshua (chapters 1 and 2), in dealing with the Deuteronomist's view of the ideal man of faith—one who is full of courage and faithful to the law that was given to Moses—relates the story of spies sent to Jericho, where they were sheltered by Rahab, a harlot, whose house was spared by the Israelites when they later destroyed the city. In the Gospel According to Matthew, in the New Testament, Rahab is listed as the grandmother of Jesse, the father of David (the architect of the Israelite empire), which may be the reason why this story was included in Joshua. Also in the New Testament, in the Letter to the Hebrews, Rahab is depicted as an example of a person of faith. After the return of the spies, who reported that the people of Canaan were "fainthearted" in the face of the Israelite threat, Joshua launched the invasion of Canaan; the Israelite tribes crossed the Jordan River and encamped at Gilgal, where the males were circumcised after a pile of stones had been erected to commemorate the crossing of the river. They then attacked Jericho and, after the priests marched around it for seven days, utterly destroyed it in a *herem; i.e.*, a holy war in which everything is devoted to destruction. Prior to the Israelites' further conquests it was discovered that Achan, a member of the tribe of Judah, had broken the *herem* by not devoting everything taken from Jericho to Yahweh. Because he had thus sinned in keeping some of the booty, Achan, his family, and all of his household goods were destroyed and a mound of stones was heaped upon them. The Israelite tribes next conquered Ai, made agreements with the people of the region of Gibeon, and then campaigned against cities to the south, capturing several of them, such as Lachish and Debir, but not Jerusalem or the cities of Philistia on the seacoast. Joshua moved north, first conquering the city of Hazor—a city of political importance—and then defeating a large number (31) of the kings of Canaan, though the conquests of their cities did not necessarily follow.

The significance of the *herem*

Division of the land and renewal of the Covenant. The division of the land among the tribes is recounted in chapters 13–22. Two sources were apparently used by the Deuteronomist in dealing with the division of the land: a boundary list from the pre-monarchical period (*i.e.*, before the late 11th century BCE) and a list of cities occupied by several tribes from the 10th to the 7th century BCE. The tribes who occupied territories were: Reuben, Gad, Manasseh, Caleb, Judah, the Joseph tribes (Ephraim and Manasseh), Benjamin, Simeon, Zebulun, Issachar, Asher, Naphtali, and Dan. Certain cities (*e.g.*, Hebron, Shechem, and Ramoth) were designated Levitical cities. Though the Levites probably did not control the cities politically, as the priestly class they were of cultic significance—and therefore feared and respected—in cities that were the sites of sanctuaries.

As Moses had before him, Joshua gave a farewell address (chapter 23) to his people, admonishing them to

The Covenant renewal ceremony

be loyal to the Lord of the Covenant; and in the closing chapter (24), the Israelites reaffirmed their loyalty to Yahweh at Shechem: first having heard the story of God's salvatory deeds in the past, they were asked to swear allegiance to Yahweh and to repudiate all other gods, after which they participated in the Covenant renewal ceremony. After the people were dismissed, Joshua died and was buried in the hill country of Ephraim; the embalmed body of Joseph that had been carried with the Hebrews when they left Egypt more than a generation earlier was buried on purchased land; and Eleazar, the priestly successor to Aaron (Moses' brother), was buried at Gibeah.

Besides the obvious emphases on the conquest of Canaan and the division of the land, the Deuteronomist gave special attention to the ceremony of Covenant reaffirmation. By means of a regularly repeated Covenant renewal the Israelites were able to eschew Canaanite religious beliefs and practices that had been absorbed or added to the religion of the Lord of the Covenant, especially the fertility motifs that were quite attractive to the Hebrew tribes as they settled down to pursue agriculture, after more than a generation of the nomadic way of life.

Judges. The Book of Judges, the third of the series of five books that reflect the theological viewpoint of the Deuteronomic historian, covers the history of the Israelite tribes from the death of Joshua to the rise of the monarchy, a period comprising nearly 200 years (*c.* 1200–*c.* 1020 BCE). Though the internal chronology of Judges points to a period of about 400 years, the editor may have arbitrarily used the formula of 40 years for a generation of rule by a judge; and he may have compiled the list in the form of a series of successive leaders who actually may have led only a particular tribe or a group of tribes during the same generation as another judge. In other words, the reign of two or more judges may well have overlapped.

The Deuteronomic "theology of history." The Deuteronomic "theology of history" shows through very clearly in Judges: unless the people of the Covenant remain faithful and obedient to Yahweh, they will suffer the due consequences of disobedience, whether it be an overtly willful act or an unthinking negligence in keeping the Covenant promise. The Deuteronomist worked out a formula for his theology of history that was based in a very dramatic way on the historical events of the period: (1) obedience to Yahweh brings peace and well-being; (2) a period of well-being often involves a slackening of resolve to keep the commandments of Yahweh or outright disobedience; (3) disobedience leads to a weakness of the faith that had bound the community together and thus leaves the community open to repression and attacks from external enemies; and (4) external repression forces the community to reassess its position and ask the cause of the calamities, thus leading to repentance and eventual strength to resist all enemies.

Canaanite culture and religion. The Israelite tribes during the period of the guidance and leadership of Moses and Joshua mainly had to contend with nomadic tribes; in their contacts with such groups they absorbed some of the attitudes and motifs of the nomadic way of life, such as independence, a love of freedom to move about, and fear of or disdain for the way of life of settled, agricultural, and urban peoples.

Canaanite cultural accomplishments and religious beliefs and practices

The Canaanites, with whom the Israelites came into contact during the conquest by Josha and the period of the Judges, were a sophisticated agricultural and urban people. The name Canaan means Land of Purple (a purple dye was extracted from a murex shellfish found near the shores of Palestine). The Canaanites, a people who absorbed and assimilated the features of many cultures of the ancient Near East for at least 500 years before the Israelites entered their area of control, were the people who, as far as is known, invented the form of writing that became the alphabet, which, through the Greeks and Romans, was passed on to many cultures influenced by their successors—namely, the nations and peoples of Western civilization.

The religion of the Canaanites was an agricultural religion, with pronounced fertility motifs. Their main gods were called the Baalim (Lords) and their consorts, the Baalot (Ladies), or Asherah (singular), usually known by the personal plural name Ashtoret. The god of the city of Shechem, which city the Israelites had absorbed peacefully under Joshua, was called Baal-berith (Lord of the Covenant) or El-berith (God of the Covenant). Shechem became the first cultic centre of the religious tribal confederacy (called an amphictyony by the Greeks) of the Israelites during the period of the judges. When Shechem was excavated in the early 1960s, the temple of Baal-berith was partially reconstructed; the sacred pillar (generally a phallic symbol or, often, a representation of the *ashera*, the female fertility symbol) was placed in its original position before the entrance of the temple.

The Baalim and the Baalot, gods and goddesses of the Earth, were believed to be the revitalizers of the forces of nature upon which agriculture depended. The revitalization process involved a sacred marriage (*hieros gamos*), replete with sexual symbolic and actual activities between men, representing the Baalim, and the sacred temple prostitutes (*qedeshoṭ*), representing the Baalot. Cultic ceremonies involving sexual acts between male members of the agricultural communities and sacred prostitutes dedicated to the Baalim were focussed on the Canaanite concept of sympathetic magic. As the Baalim (through the actions of selected men) both symbolically and actually impregnated the sacred prostitutes in order to reproduce in kind, so also, it was believed, the Baalim (as gods of the weather and the Earth) would send the rains (often identified with semen) to the Earth so that it might yield abundant harvests of grains and fruits. Canaanite myths incorporating such fertility myths are represented in the mythological texts of the ancient city of Ugarit (modern Ras Shamra) in northern Syria; though the high god El and his consort are important as the first pair of the pantheon, Baal and his sexually passionate sister-consort are significant in the creation of the world and the renewal of nature.

Attraction of the Israelites toward Canaanite religion

The religion of the Canaanite agriculturalists proved to be a strong attraction to the less sophisticated and nomadic-oriented Israelite tribes. Many Israelites succumbed to the allurements of the fertility-laden rituals and practices of the Canaanite religion, partly because it was new and different from the Yahwistic religion and, possibly, because of a tendency of a rigorous faith and ethic to weaken under the influence of sexual attractions. As the Canaanites and the Israelites began to live in closer contact with each other, the faith of Israel tended to absorb some of the concepts and practices of the Canaanite religion. Some Israelites began to name their children after the Baalim; even one of the judges, Gideon, was also known by the name Jerubbaal ("let Baal contend").

As the syncretistic tendencies became further entrenched in the Israelite faith, the people began to lose the concept of their exclusiveness and their mission to be a witness to the nations, thus becoming weakened in resolve internally and liable to the oppression of other peoples (see also SYRIAN AND PALESTINIAN RELIGIONS).

The role of the judges. Under these conditions, the successors to Joshua—the judges—arose. The Hebrew term *shofeṭ*, which is translated into English as "judge," is closer in meaning to "ruler," a kind of military leader or deliverer from potential or actual defeat. In a passage from the so-called Ras Shamra tablets (discovered in 1929) the concept of the judge as a ruler is well illustrated:

> Our king is Triumphant Baal,
> Our judge, above whom there is no one!

The magistrates of the Phoenician-Canaanite city of Carthage, which competed with Rome for supremacy of the Mediterranean world in the 3rd century BCE, were called *suffetes*, thus pointing toward the political authority of the judges.

The office of judgeship in the tribal confederacy of the Israelites, which was centred at a covenant shrine, was not hereditary. The judges arose as Yahweh saw fit, in

The significance of the office of the judge

order to lead an erring and repentant people to a restoration of a right relationship with him and to victory over their enemies. The quality that enabled a person selected by Yahweh to be a judge was charisma, a spiritual power that enabled the judge to influence, lead, and control the people caught between the allurements of the sophisticated Canaanite culture and the memory of the nomadic way of life with its rugged freedom and disdain for "civilization." Though many such leaders are mentioned, the Book of Judges focusses attention upon only a few that are singled out as especially significant: Deborah and Barak, Gideon, Abimelech, Jephthah, and Samson. In spite of the Israelites' repeated apostasy, such leaders, under the guidance and spiritual powers granted to them by Yahweh, were able to lead their tribes in successfully defeating or driving back their opponents.

The Book of Joshua may be divided into four parts: (1) the conquests of several tribes (chapter 1); (2) a general background for the subsequent events according to the interpretation of the Deuteronomic historian—"And the people of Israel did what was evil in the sight of the Lord and served the Baals"—(chapter 2 through chapter 3, verse 6); (3) the exploits of the judges of Israel (chapter 3, verse 7, through chapter 16); and an appendix (chapters 17 through 21).

Judges, chapter 1, shows that the conquest of Canaan, in contradistinction to the view presented in Joshua, was incomplete, inconclusive, and lengthy. Though conquests of some of the tribes (Judah, Simeon, Caleb, and the "house of Joseph") are noted, the main emphasis is on the cities and areas that the tribes had *not* conquered—*e.g.*, "And Ephraim did not drive out the Canaanites who dwelt in Gezer, but the Canaanites dwelt in Gezer among them" (chapter 1, verse 29).

The second section gives the Deuteronomic interpretation of the consequences of such a policy:

> they forsook the Lord, the God of their fathers, who had brought them out of the land of Egypt; they went after other gods, from among the gods of the peoples who were round about them; and they provoked the Lord to anger. They forsook the Lord, and served the Baals and the Ashtaroth. (chapter 2, verses 12–13.)

The purpose of allowing the Canaanites to continue to exist

In chapter 3, an explanation is given as to why the Canaanites had not been annihilated and were allowed to remain with the Israelites: they enabled the Israelites to be tested in the techniques of warfare; the Philistines, for example, had a monopoly on the smelting of iron in the area—and the iron used in their weapons was far superior to the bronze used by the Israelites for their swords, shields, and armaments—until the secret had been wrested from them by the first king of Israel, Saul, in the latter part of the 11th century BCE. The Canaanites also served to test the faith of the Israelites in the one, true God, Yahweh.

The role of certain lesser judges. The third section relates the exploits of the various judges. Othniel, a member of the tribe of Caleb, delivered the erring Israelites from eight years of oppression by Cushan-rishathaim, king of Mesopotamia. The king, however, was most likely an area ruler, rather than a king of the Mesopotamian Empire. Another judge, Ehud, a left-handed Benjamite, delivered Israel from the oppression of the Moabites. Ehud, a fat man who had hidden a sword under his garments on his right side so that when a search of his person was made it would be overlooked, brought tribute to Eglon, the Moabite king. Upon Ehud's claiming to have a secret message for the king, Eglon dismissed the other people carrying tribute. Ehud then said to the King, "I have a message from God to you," assassinated him, locked the doors to the chamber, and escaped. Rallying the Israelites around him, Ehud led an attack upon the Moabites that was decisive in favour of the Israelites. Shamgar, the third judge, is merely noted as a deliverer who killed 600 Philistines.

The roles of Deborah, Gideon, and Jephthah. The first notably important judge of the tribal confederacy was Deborah, who was primarily a seer, poetess, and interpreter of dreams but still a person endowed with the kind of charisma that identified her as a judge sent from Yahweh. The story of the victory of the Israelites under the charismatic leadership of Deborah and the military leadership of Barak, her commander, is related in prose (chapter 4) and repeated in poetry (chapter 5, which is known as the "Song of Deborah"). The Canaanites, under the leadership of Jabin, king of a reestablished Hazor, and his general Sisera, had oppressed an apostate Israel. Deborah sent word to all the tribes to unite against the Canaanites, but only about half the tribes responded. The Canaanites had asserted control over the Valley of Jezreel, which was an important commercial thoroughfare and was commanded by the city of Megiddo. In this valley dominated by the hill of Megiddo (Armageddon)—a site of many later crucial military battles and which later became the symbolic name for the final battle between the forces of good and the forces of evil in apocalyptic literature—the Israelites met the Canaanites near the river Kishon in open battle. A cloudburst occurred, causing the river to flood, thus limiting the manoeuvrability of the Canaanite chariots. The Canaanite general Sisera, seeing defeat for his forces, fled, seeking refuge in the tent of a Kenite woman, Jael. A supporter of the cause of Israel, Jael gave Sisera a drink of milk (fermented?) and he fell asleep "from weariness." Jael pounded a tent peg through his temple, thus ending decisively the threat of the Canaanites of Hazor. The victory song of Deborah in chapter 5 is one of the oldest literary sections of the Old Testament. It is a hymn that incorporates the literary forms of a confession of faith, a praise of Yahweh's theophany (manifestation), an epic, a curse, a blessing, and a hymn of victory.

The significance of Deborah and Gideon

Another important judge, perhaps the most important other than Samuel, was Gideon, whose exploits are related in chapters 6–8. The oppressors of Israel during the time of Gideon were the camel-borne raiders from Midian, roving bands that pillaged the farms and unfortified villages for seven years. A prophet appeared among the Israelites and denounced them for their apostasy, after which, according to the account, an angel of Yahweh visited and then commissioned Gideon, a member of the tribe of Manasseh, to lead the Israelites against the enemies from the Transjordan. After sacrificing to Yahweh, building an altar to the Lord (which he named Yahweh Shalom, or "Yahweh is peace"), and destroying an altar of Baal and an *ashera* (most likely a wooden pole symbolizing the goddess) beside it, he sent out messengers to gather together the tribes in order to meet an armed force of the Midianites and Amalekites that had crossed the Jordan River and were encamped in the Valley of Jezreel. He went to a threshing floor (a common place to seek divinatory advice) and sought a sign from Yahweh—dew on a fleece of wool placed overnight on the threshing floor, with the rest of the area remaining dry. After receiving the positive divinatory sign, Gideon assembled a large force, reduced it to 300 men, and infiltrated the outposts of the Midianite camp with his servant—overhearing a Midianite telling another of his dream about a barley cake rolling into the camp of the Midianites and striking a tent so that it fell down and was flattened (which Gideon interpreted as a sign of victory for the forces under him). He encircled the camp of the Midianites about midnight. On signal, the men broke jars, shouted, waved torches, blew rams' horns, and attacked the encampment. The Midianites, in the confusion, were routed and harassed in their flight. In their pursuit of the fleeing Midianites, Gideon and his forces were refused aid by the cities of Succoth and Penuel, which was a violation of the tribal confederacy agreements. The Midianites, however, were again the objects of a surprise attack and their two kings (Zebah and Zalmunna) were captured and later executed by Gideon because they had killed his brother. The leaders of Succoth were punished and the men of Penuel were killed in retaliation for their refusal to aid the forces of Gideon.

The request for a hereditary monarchy

After the victory, the people, recognizing their need for centralized leadership of the confederacy, petitioned to Gideon that he establish a hereditary monarchy, with himself as the first king. Gideon refused, however, on the basis that "the Lord will rule over you."

After Gideon died, the people returned to worshipping the gods of the Canaanites, especially Baal-berith. Abimelech, one of the 70 sons of the wives and concubines of Gideon, went to Shechem to solicit support for his attempt to establish a monarchy. After receiving financial support from those who controlled the treasury of the shrine of Baal-berith, he hired a band of assassins—who killed all of his brothers except Jotham, the youngest of Gideon's sons. Abimelech was declared king by the Shechemites. The surviving Jotham told a parable about trees that sought a king—after all the larger trees refused the kingship, the bramblebush, which was highly inflammable, accepted the offer. The point of the parable was that as the bramblebush is highly inflammable, so also would the reign of Abimelech be the source of fires of rebellion and revolution. Revolution did occur, and after being wounded at Thebez by a millstone dropped by a woman from a tower, Abimelech asked his armour bearer to kill him. The attempt of Abimelech and the Shechemites to establish a monarchy thus proved to be abortive and premature.

After a brief account of the rule of two judges, Tola of the tribe of Issachar and Jair from Gilead, the Deuteronomist describes the apostasy of the Israelites and the consequent oppression of the tribes by the Philistines from the seacoast and the Ammonites from the Transjordan. The Israelites looked for a leader and found one in the person of Jephthah, the son of a harlot, who had been rejected by the sons of his father and who had gathered about him a band of men who made their living by raiding others. Jephthah made several attempts to negotiate with the Ammonites and Moabites; when the Ammonites did not cooperate, Jephthah moved against them. Seized by the Spirit of the Lord—*i.e.*, ecstatically inspired—he began his campaign with a vow to sacrifice the first person he saw upon his return home as a burnt offering to Yahweh. He was victorious over the Ammonites, but the first person he saw on return home was his only child, a daughter. Upon learning of her destined fate, she requested a two-month period to be with her friends to bewail her virginity and approaching death. The story is reminiscent of the fertility myths of the ancient Near East. After she was sacrificed, Jephthah subdued a contingent of the Ephraimites in the Transjordan in order to bring peace to the area. A password was used to separate the Ephraimites from the men under Jephthah: "shibboleth." Because the Ephraimites could not pronounce the word correctly, in that their dialect was different from the others, they were thus identified and killed.

The judgeship of Jephthah

In chapter 12, three judges are given cursory treatment: Izban of Bethlehem, Elon the Zebulunite, and Abdon the Ephraimite.

The exploits of Samson

The role of Samson. The exploits of the great Israelite strong-man judge, Samson (a member of the tribe of Dan), are related in chapters 13–16. Dedicated from birth by his mother to Yahweh, Samson became a member of the Nazirites, an anti-Canaanite reform movement. As a Nazirite, he was required never to cut his hair, drink wine, or eat ritually unclean food. He married a Philistine woman whom he then left when she helped her fellow Philistines avoid payment to Samson in a riddle contest by giving them the answer. Returning later to find her given to another man, he burned the grainfields of the Philistines. They sought revenge by killing Samson's wife and her father. The exploits of Samson against the Philistines from then on are numerous. After he met the temptress Delilah, who wrested from him the secret of his great strength (*i.e.*, his long uncut hair because of his vow), Samson was captured by the Philistines after his hair had been cut short. After imprisonment, blinding, and humiliation, Samson finally avenged his loss of self-respect by pulling down the main pillars of the temple of the Philistine god Dagon, after which the temple was destroyed, along with numerous Philistines. Though Samson was more a folk hero than a judge, he was probably included in the list of judges because his ventures against the Philistines slowed their movements inland against the Israelite towns and villages. The Philistines were a group of "sea peoples" united in a confederacy of five city-states: Gaza, Ashkelon, Ashdod, Gath, and Ekron. To the area they gave their name, which has endured to the 20th century: Palestine.

The final section of the Book of Judges is an appendix divided into two parts: (1) the story of Micah, the repentant Ephraimite, a Levite priest who deserted him to be priest of the tribe of Dan, and the establishment of a shrine at the conquered city of Laish (renamed Dan) with the cult object taken from the house of Micah; and (2) the story of the Benjamites who were defeated in a holy war after they had killed a concubine of a Levite. The book ends with a critique of the period: "In those days there was no king in Israel; every man did what was right in his own eyes" (chapter 21, verse 25).

Samuel. The book of Samuel covers the period from Samuel, the last of the judges, through the reigns of the first two kings of Israel, Saul and David (except for David's death). The division of Samuel and its succeeding book, Kings (Melakhim), into four separate books first appeared in the Septuagint, the Greek translation of the Old Testament from the 3rd to 2nd centuries BCE.

The two main sources of Samuel

Theological and political biases. Containing two primary sources, the book of Samuel is the result of the editorial skill of the Deuteronomic historians of the post-exilic period. The early source, which is pro-monarchical and may have been written by a single author, is found in I Samuel, chapter 9, verse 1, through chapter 10, verse 16, as well as chapter 11 and most of II Samuel. The chapters just noted were probably written by a chronicler during the reign of Solomon; possible authors of these chapters were Abiathar, a priest of the line of Eli (who was Samuel's predecessor at the shrine of Shiloh), or Ahimaaz, a son of Zadok (who originally may have been a priest of the Jebusite city of Jerusalem that David made his capital). The chapters in I Samuel are sometimes called the "Saul" source because it is in them that Saul's charismatic leadership is legitimized in the form of kingship. The chapters of II Samuel, also displaying a pro-monarchical bias—as far as content is concerned—are the "book of David." In the early source, Samuel, a seer, prophetic figure, and priest of the shrine at Shiloh, is viewed mainly as the religious leader who anointed Saul to be king. The later source, which displays a somewhat anti-monarchical bias and shows the marks of disillusionment on the part of the Deuteronomic historians of the post-exilic period is found in I Samuel, chapter 7, verse 3, to chapter 8, verse 22, chapter 10, verses 17–27, and chapter 12. Sometimes called the Samuel source, the later source interprets the role of Samuel differently; he is viewed as the last and most important judge of the whole nation, whose influence extended to the shrines at Bethel, Gilgal, and Mizpah. The two sources illustrate the two opposing tendencies that lasted for centuries after the conquest of Canaan.

The international and area situation during the 11th–10th centuries BCE

During the period of Samuel, Saul, and David (the 11th–10th century BCE), the Israelites were still threatened by various local enemies. The great nations—Egypt, Assyria, and the Hittite Empire—were either involved in domestic crises or concerned with areas other than Palestine in their expansionist policies. Of the various peoples pressing to break up the Israelite confederacy, the Philistines (the "sea peoples") of the Mediterranean coast proved to be the most dangerous. Expanding eastward with their iron-weapon equipped armies, the Philistines threatened the commercial routes running north and south through Israelite territory. If they captured and controlled such areas as the Valley of Jezreel, they would eventually strangle the economic life of the Israelite confederacy.

To meet this threat, the tribal confederacy had four options open to it. First, the tribes could continue as before, loosely held together by charismatic leaders, who served only as temporary leaders. Second, they could create a hereditary hierocracy (rule by priests), which the priest of the shrine at Shiloh, Eli, apparently attempted to inaugurate. A third possible course of action was to establish a hereditary judgeship, which was the aspiration of the judge Samuel. But in either of these two possibili-

ties, the sons of Eli and Samuel were not of the same stature as their fathers; and the apparent hopes of their fathers could not be realized. The fourth alternative was a hereditary monarchy. The book of Samuel is an account of the eventual success of those who supported the monarchical position, along with the Deuteronomic interpretation that pointed out the weaknesses of the monarchy whenever it departed from the concept of Israel as a covenant people and became merely one kingdom among other similar kingdoms.

The book of Samuel may be divided into four sections: (1) the stories of Samuel, the fall of the family of Eli, and the rise of Saul (I Samuel, chapters 1–15); (2) the accounts of the fall of the family of Saul and the rise of David (I Samuel, chapter 16, to II Samuel, chapter 5); (3) the chronicles of David's monarchy (II Samuel, chapter 6, to chapter 20, verse 22); and (4) an appendix of miscellaneous materials containing a copy of Psalm 18, the "last words of David," which is a psalm of praise, a list of heroes and their exploits, an account of David's census, and other miscellaneous materials.

The early life and "call" of Samuel

The role of Samuel. The first section (chapters 1–15) begins with the story of Samuel's birth, after his mother Hannah (one of the two wives of the Ephraimite Elkanah) had prayed at the shrine at Shiloh, the centre of the tribal confederacy, for a son. She vowed that if she bore a son, he would be dedicated to Yahweh for lifetime service as a Nazirite, as indicated by the words "and no razor shall touch his head."

Three years after she had borne a son, whom she named Samuel—which is interpreted "asked of God," a phrase that fits the meaning of Saul's name but may actually mean "El has heard"—Hannah took the boy to the shrine at Shiloh. Hannah's song of exultation (chapter 12, verses 1–10) probably became the basis of the form and content of the Magnificat, the song that Mary, the mother of Jesus, sang in Luke, chapter 1, verses 46–55, in the New Testament. Eli, the priest at Shiloh (who had heard Hannah's vow), trained the boy to serve Yahweh at the shrine, which Samuel's mother and father visited annually. The sons of Eli, Hophni and Phinehas, are depicted as corrupt, misusing their positions as servants of the shrine to take offerings the people gave to Yahweh for their own gratification, in contrast to Samuel, who "continued to grow in stature and favour with the Lord and with men." Because the sons of Eli failed to heed the admonition of their father, the house of Eli was condemned by a "man of God," who told Eli that his family was to lose its position of trust and power. This condemnation, an interruption of the later source, is the Deuteronomic historian's answer as to why Abiathar, a priest of the family of Eli at the time of David, was excluded from the priesthood at Jerusalem, which became the central shrine of the monarchy.

While a youth (about 12 years old), Samuel experienced a revelation from Yahweh in the shrine at night. First going to Eli three times after hearing his name called, Samuel responded to Yahweh at Eli's suggestion. What was revealed to him was the fall of the house of Eli, a message that Samuel hesitatingly related to Eli. After this religious experience, Samuel's reputation as a prophet of Yahweh increased.

The fall of Shiloh and the house of Eli

In chapter 4 is an account of the fall of Shiloh and the loss of the ark of the Covenant to the Philistines. Leaving the ark, the symbol of Yahweh's presence, at Shiloh, the Israelites go out to battle against the Philistines near the Mediterranean coast but are defeated. The Israelites return to Shiloh for the ark; but even though they carry it back to the battleground, they are again defeated at great cost—the sons of Eli are killed, and the ark is captured by the enemy. When Eli, old and blind, hears the news of the disaster, he falls over backward in the chair on which he is sitting, breaks his neck, and dies. The wife of his son Phinehas gives birth to a son at this time; and, upon hearing of what had happened to Israel and her family, names the boy Ichabod, meaning "where is the glory?"—because, as she says, "The glory has departed from Israel."

Though the Philistines had captured the ark, they eventually discovered that it did not bring them good fortune. Their god Dagon, an agricultural fertility deity probably meaning "grain," fell to the ground whenever the ark was placed in close proximity to it; and, even more calamitous to them, the Philistines suffered from "tumours," probably the bubonic plague, wherever they carried the ark. After experiencing such disasters for seven months, the Philistines returned the ark to Beth-shemesh in Israelite territory, along with a guilt offering of five golden tumours and five golden mice carried in a cart drawn by two cows. Because many Israelite men in Beth-shemesh also died—"because they looked into the ark of the Lord"—the ark was taken to Kiriath-jearim (the "forest of martyrs" in modern Israel), where it was placed in the house of Abinadab, whose son Eleazar was consecrated to care for it. The ark was not returned to Shiloh, probably because that shrine centre had been destroyed, along with other Israelite towns, by the Philistines.

In chapter 7, verse 3, to chapter 12, verse 25, the Deuteronomic historian depicts the way in which Samuel assumed leadership as judge and Covenant mediator of Israel. The Philistines continued to oppress Israel, though under Samuel's leadership the Israelites were able to reconquer territory lost to their western enemies. When Samuel grew old, his sons were trained to take his place; but they—like the sons of Eli—were corrupt ("they took bribes and perverted justice"), so that the Israelites demanded another form of government—a monarchy. Samuel attempted to dissuade them, pointing out that if they had a highly centralized form of government (*i.e.*, a monarchy), they would have to give up much of their freedom and would be heavily taxed in goods and services. Samuel obeyed both the elders of the people, who demanded a king, and Yahweh, who said, "make them a king."

The request for a monarchy

The rise and fall of Saul. The man selected to become the first monarchical ruler of Israel was Saul, son of Kish, a wealthy Benjamite landowner. Because Kish had lost some donkeys, Saul was sent in search of them. Unsuccessful in his search, he went to the seer-prophet Samuel at Ramah. In the early source, from which this narrative comes, he did not know Samuel's name. The day before Saul went to Ramah, Samuel the seer (*ro'e*), who was depicted by the Deuteronomic historian as a prophet (*navi'*), received notice from Yahweh that Saul was the man chosen to reign over Israel. At the sacrificial meal, Saul, a tall young man, was given the seat of honour, and the next day Samuel anointed him prince (*nagid*) of Israel in a secret ceremony. Before returning home, Saul joined a band of roving ecstatic prophets and prophesied under the influence of the spirit of Yahweh. In chapter 10, verse 17–27, generally accepted as part of the later source, the Deuteronomic historian's views are depicted—Saul was chosen by lot at Mizpah. The early source picks up the story of Saul in chapter 11, which illustrates Saul's military leadership abilities and describes his acclamation as king at Gilgal. Samuel's farewell address, a Deuteronomic reworking of the later source, recapitulates the history of the Israelite tribes from the time of the patriarch Jacob through the period of the judges and forcefully presents the conservative view that the request for a monarchy will bring about adversity to Israel.

The reign of Saul

The early reign of Saul and his confrontations with Samuel until the last judge's death is the subject of chapters 13–15. Saul's early acts as king centred about battles with the Philistines. Because his son Jonathan had defeated one of their garrisons at Geba, the Philistines mustered an army to counterattack near Beth-aven (probably another name for Bethel). Saul issued a request for volunteers, who gathered together for battle but awaited the performance of the sacrifice before the battle by Samuel. Because Samuel did not come for seven days, Saul, acting on his own, presided at the sacrifice. Immediately after the burnt offering had been completed, Samuel appeared (perhaps waiting for such an opportunity to reassert his leading position) and castigated Saul for overstepping the boundaries of his princely prerogatives—even though Saul had been more than patient.

Samuel warned him that this type of act (which Saul, in the early source, and later David and Solomon also often performed) would cost Saul his kingdom. In spite of Samuel's apparent animosity, Saul continued to defend the interests of the newly formed kingdom.

The tragedy of Saul was that he was a transitional figure who had to bear the burden of being the man who was of an old order and at the same time of a new way of life among a people composed of disparate elements and leading figures. Both Samuel, the last judge of Israel, and David, the future builder of the small Israelite empire, opposed him. Saul was more a judge—a charismatic leader—than a monarch. Unlike most kings of his time and area, he levied no taxes, depended on a volunteer army, and had no harem. He did not construct a court bureaucracy but relied rather on the trust of the people in his charismatic leadership and thus did not alter the political boundaries or structure of the tribal confederacy.

The dispute between Samuel and Saul

The issue between Saul and Samuel came to a head in the events described in chapter 15 (a section from the later source). Samuel requested Saul to avenge the attacks by the Amalekites on the Israelite tribes during their wanderings in the wilderness after the Exodus from Egypt about 200 years earlier. Saul defeated the Amalekites in a holy war but did not devote everything to destruction as was required by the ban (*herem*). Because Saul had not killed Agag, the Amalekite king, and had saved sheep and cattle for a sacrifice, Samuel informed Saul that he had disobeyed Yahweh and was thus rejected by God, for "to obey is better than to sacrifice." Samuel then asked that Agag be brought to him, and he hacked the Amalekite king to pieces. After that, Saul and Samuel saw each other no more.

The rise and significance of David. The next section contains the account of Saul's fall from power and David's rise to the position of king over all Israel. Samuel, still a charismatic and political power of great consequence, received from Yahweh the message that he was to go to Bethlehem to anoint a new ruler. Because he feared reprisal from Saul, Samuel went to Bethlehem (whose elders had the same fears) under the pretense of presiding at a sacrifice. There he anointed David, son of Jesse, to be future king. David then went to the court of Saul to be the king's armour bearer and court singer.

In a battle with the Philistines David is reported to have killed the 10-foot-tall Philistine champion Goliath of Gath. In II Samuel, chapter 21, verse 19, however, Goliath is killed in a later period by one of David's warriors, Elhanan. According to some biblical scholars, the name of Goliath may have been inserted for an unnamed Philistine warrior killed by David apparently while he was armour bearer to Saul and was unrecognized by Saul, thus indicating the reworking of more than one source by the Deuteronomic historian.

The rise of David and the death of Saul

Chapters 18 through 26 depict the rise of David in the court of Saul, his friendship with Jonathan, the beginning of Saul's jealousy of David, the young David's winning of Saul's daughter Michal in marriage for killing a large number of Philistines, Saul's attempt on David's life, David's escape and formation of an outlaw band in the Judaean hills, his acceptance by the priests of the house of Eli at Nob (all of whom were killed by Saul except Abiathar, who became David's priest), Samuel's death, and other incidents.

Because he feared for his life, David, along with 600 of his men, fled to the Philistine city of Gath, where he became a supposed leader of one of their military contingents against the Israelites. The last four chapters of I Samuel depict the final futile effort of Saul to retain control of his throne and thwart the Philistines: Saul attempted to receive advice from the spirit of the dead Samuel through the necromancer (sometimes called the witch or medium) of Endor, even though he had earlier banned such practices in his realm. Through her mediumship, Samuel foretold the death of Saul and his sons by the Philistines. The armies of the Philistines poured into the Valley of Jezreel. Some of the Philistine leaders distrusted David, who was sent back to his garrison town of Ziklag, which the Amalekites had overrun and in which they had taken many prisoners. Thus, David did not witness the defeat of the Israelites under Saul, who was mortally wounded by the Philistines and whose sons were killed. In an act of heroism so that he, the king of Israel, would not be captured, Saul committed suicide by falling on his own sword. Thus ended the career of the tragic hero who tried to serve Yahweh and Israel but was caught between the old, conservative ways (led by Samuel) and the new, liberal views (championed by David).

The early reign of David

The Second Book of Samuel, as noted earlier, relates the exploits of David and the events of his monarchy. After mourning the death of Saul and executing an Amalekite who claimed to have killed the former king, David began to consolidate his position as the successor to Saul. He was anointed king of Judah at Hebron while Ishbosheth ("man of shame," originally Ishbaal, or "man of Baal"), Saul's son, reigned in the rest of Israel under the guidance of Abner, Saul's general. After seven years, the army of Israel, under Abner, and the army of Judah, under Joab, David's general and nephew, met at Gibeon —each chose 12 champions to fight each other, and all were killed. After the minor battle, a major engagement ensued, with the forces of Judah emerging victorious. A long war of attrition developed between the house of Saul and the house of David. Abner attempted to deliver Israel to David but was killed by Joab to avenge his brother Asahel's death at Abner's hand in the first engagement between the two reigning houses. With Abner dead, Ishbosheth's position became exceedingly insecure, and he was beheaded by two of his own captains, whom David, in turn, executed for murdering the last ruler of the house of Saul.

Because of the course of events, the Israelites asked David to become king over all of Israel, and David made a covenant with the elders of northern Israel. He next engaged in a war with the Jebusite (Canaanite) stronghold of Jerusalem, which he captured. He selected this city as his new capital because it was a neutral site and neither the northerners nor the southerners would be adverse to the selection. From the very beginning of his reign, David showed the political astuteness and acumen that made for him a reputation that has continued for 3,000 years. He built at his new capital a palace, fortified the defenses, and established a harem. The Philistines, concerned about the man whom they had considered a former vassal, decided to move against David, which proved to be their undoing. David effectively contained them in a small area of the Mediterranean coast.

The expansion of the Davidic Empire. The third section of Samuel (II Samuel, chapter 6 through chapter 20, verse 22) contains the account of the reign of David from Jerusalem, ruling over a minor empire that stretched from Egypt in the south to Lebanon in the north and from the Mediterranean Sea in the west to the Arabian Desert in the east. He thus controlled the crossroads of the great empires of the ancient Near East. His second act of political astuteness was to bring the ark of the Covenant to Jerusalem; but because of pressures from conservative elements who wanted to retain the tent that housed the ark (which had symbolic value from the days of the Exodus), David was not able to build a temple. Because the ark was now in Jerusalem, however, the city became both the political and the religious cult centre of his kingdom. In chapter 8 is a summary account of David's extension of his kingdom by military means and of the military, administrative, and priestly leaders of Israel.

II Samuel, chapters 9 through 20, verse 22—together with I Kings, chapters 1 and 2, the so-called Succession History, or the Family History of David, which, according to many scholars, forms the oldest section of historiography in Scripture—contains accounts of the domestic problems of David's reign. Though he showed generosity to Mephibosheth, the sole surviving son of the house of Saul, he showed his weakness for the charms of Bathsheba, the wife of Uriah, one of his generals. After ensuring Uriah's death by sending him into the front lines in

a battle with the Ammonites, David married Bathsheba, who had become pregnant by the King. When the prophet Nathan came to David and told him of a rich man's unjust actions toward a poor man, David's response was one of anger and a demand for justice, whereupon Nathan said, "You are the man," and that Yahweh would exact retribution by not allowing the child to live. David then repented. He later went to Bathsheba and she conceived and bore another child, Solomon, who was to be the future king of Israel.

Domestic problems of the house of David

Though David was viewed as a master in the art of governing a nation, he was depicted as an unsuccessful father of his family. One son, Amnon (half-brother to Absalom and his sister Tamar), raped Tamar, for which act Absalom later exacted revenge by having Amnon assassinated at a feast. Absalom then fled to Geshur, stayed there three years, was taken back to Jerusalem by Joab, and two years later was reconciled to his father. Absalom's ambition to succeed his father as king caused him to initiate a revolt so that David had to flee from Jerusalem. Absalom was crowned king at Hebron, went to the concubines of David's harem in the palace, and decided to raise a massive army to defeat David. If he had then heeded the advice of Ahithophel, one of David's former counsellors, and attacked David's forces while they were disorganized, he probably would have been successful in retaining the throne. The forces of David under Joab, however, defeated Absalom's army "in the forest of Ephraim." While in flight on a mule, Absalom caught his head in an oak tree, and, when Joab heard of his predicament, he killed the hanging son of David. When David heard of the death of his rebellious son, he uttered one of the most poignant laments in literature: "O my son Absalom, my son, my son Absalom! Would I had died instead of you, O Absalom, my son, my son!" David then returned to Jerusalem and settled some of the quarrels that had erupted in his absence. A revolt led by the conservative Benjaminite Sheba, under the old rallying cry "every man to his tents, O Israel," was thwarted by Joab, who had to kill David's newly appointed commander Amasa to accomplish this end.

The appendix (chapter 20; verse 23, through chapter 24) has been noted earlier in this section (see SAMUEL; SAUL; DAVID).

Kings. The fourth book of the Former Prophets (I and II Kings in the Septuagint) continues the history of the nation Israel from the death of David, the reign of Solomon and the divided monarchy through the collapse of both Israel (the northern kingdom) and Judah (the southern kingdom). Whereas Samuel was composed primarily of the early and the later sources with some editing on the part of the Deuteronomic historians, the Deuteronomic editors of Kings, in addition to these two sources, used other sources—such as the book of the acts of Solomon, the Book of the Chronicles of the Kings of Israel, the Book of the Chronicles of the Kings of Judah, temple archives, and traditions centring on certain major kings and prophets. The Deuteronomic historians wrote from the vantage points of the reign of King Josiah of Judah, who died in 609 BCE and was the ruler who accepted the Deuteronomic reform that began in 621 BCE, and of the Babylonian Exile, which traditionally lasted 70 years, though it began in 597 BCE, the temple was destroyed in 586, some exiles returned in 538, and the temple was restored in 516. The Deuteronomic view that national apostasy was the cause of the covenant people's predicament pervades this work.

The sources of the Book of Kings

The history of the period (10th–early 6th century BCE) is covered in the article JUDAISM, HISTORY OF, under the section *Biblical Judaism*, and therefore this article will concentrate only on the reigns of important monarchs and their relationships to the rising power of the prophetic movement in Israel.

The Book of Kings may be divided into four sections: (1) The last years of David and Solomon's succession to the throne (I Kings, chapter 1, to chapter 2, verse 11), (2) the reign of Solomon (I Kings, chapter 2, verse 12, to chapter 11, verse 43), (3) the beginning of the divided monarchy to the fall of Israel (I Kings, chapter 12, to II Kings, chapter 17), and (4) the last years of Judah II Kings, chapters 18–25).

The succession of Solomon to the throne. I Kings (chapters 1 and 2) continues the story of David and the struggle for the succession of his throne. The sides were drawn between Adonijah, David's eldest living son, and Solomon, the son of David and Bathsheba. Supporting Adonijah were the "old guard"—the general Joab and the priest Abiathar—and supporting Solomon were the priest Zadok, the prophet Nathan, and the captain of David's bodyguard, Benaiah. With David close to death, Adonijah prepared to seize control of the kingdom; Nathan, however, requested Bathsheba to go to David and persuade David to proclaim Solomon the next monarch. Following the advice of Nathan, David then appointed Solomon the heir to his throne; and Zadok the priest and Nathan the prophet anointed the son of Bathsheba king in Gihon (see also DAVID).

The appointment of Solomon as the future king of Israel

After David died, however, Adonijah attempted to regain some semblance of prestige by asking Solomon to give him Abishag, a young Shunammite woman who had been given to David in his old age, as his wife. To this request Solomon answered by ordering Adonijah's execution, which Benaiah carried out. Solomon also ordered the execution of the old general Joab for having killed Abner and Amasa years earlier as a loyal supporter of David, an execution again carried out by Benaiah, who also executed Shimei, a man who had cursed David a long time earlier. Prior to these executions, which David—before he had died—had requested of Solomon, the new king banished the priest Abiathar of the house of Eli to Anathoth, an act that confirmed the position of Zadok as the principal priest of Jerusalem.

The reign of Solomon. David had reigned from about 1000–962 BCE, a period in which he consolidated a federation of tribes that had been united under the charismatic leadership of Saul, who had reigned for about two decades before David began to construct his minor empire. Solomon, who inherited a strong monarchy, reigned for 40 years. His reputation as a monarch centred about his great wisdom (chapter 3), his reorganization of the administrative bureaucracy (chapter 4), and his building of the magnificent Temple (chapters 3–8). Though two sons of the prophet Nathan served Solomon, one as a court official and another as a priest, the prophetic movement apparently was little encouraged by the united monarchy's third king. Solomon is perhaps one of the most overrated figures in the Old Testament, in spite of his achievements in wisdom, construction, and commerce; he is recorded as having 1,000 wives and concubines—some of them merely guarantees of commercial treaties, to be sure—and as building a fleet of ships for a nearly landlocked Israel. To accommodate his desire for a seaport, he built the port of Ezion-geber at the head of the Gulf of Aqaba of the Red Sea. A son of the harem, Solomon had had little contact with the people of his realm, and he used many of them in labour battalions in his vast building programs to the economic disadvantage of Israel. By fostering social discontent in such ventures, Solomon prepared the way for the disintegration of the united kingdom and the resurgence of the prophetic movement that reflected the indigenous covenant concept peculiar to Israel.

Whereas David secured Israel's borders and property by military means, Solomon sought to extend Israel's influence through commercial treaties. To secure diplomatic and commercial treaties Solomon contracted marriage with various princesses—who brought with them their native deities. This defection from the Covenant obligations to Yahweh is viewed by the Deuteronomic historian as a continuance of Israel's constant flirting with apostasy, which had occurred under the judges, and the beginning of a long process of internal religious and political disintegration under the monarchical system. Solomon's oppressive taxation and commercial expansion brought about retaliation and rebellion.

The divided monarchy. After Solomon died (922 BCE), he was succeeded by Rehoboam, who proved to be unfit for the task of reigning. Prior to Solomon's death, Jero-

The rise of the divided kingdom: Israel and Judah

boam the Ephraimite, a young overseer of the forced labour battalions of the "house of Joseph" in the north, had encountered Ahijah, a prophet from the old shrine of the confederacy at Shiloh and Ahijah had torn a new garment into 12 pieces, prophesying that 10 pieces (tribes) would be given to Rehoboam and only two pieces (tribal political units) would be retained by the house of David. The dismemberment of the united monarchy was to be brought about by Yahweh because Solomon had "not walked in my ways, doing what is right in my sight and keeping my statutes and my ordinances, as David his father did." Though Solomon had worshipped the Sidonian goddess Ashtoreth, the Moabite god Chemosh, and the Ammonite god Milcom, his reign over Israel continued. Jeroboam's initial rebellion proved to be abortive, and he sought political asylum in Egypt under the protection of the pharaoh Sheshonk I (Shishak).

Rehoboam, having been crowned king of the united monarchy in Jerusalem, went north to Shechem, a shrine centre of the 10 northern tribes of the old confederacy, to have his position ratified by the northern units of the kingdom. Using this gathering as an opportune time to present their grievances against Solomon's oppressive domestic policies, the northerners, under the leadership of the returned political fugitive Jeroboam, asked the king from Jerusalem to lighten their load. Requesting three days to take their grievances under advisement, Rehoboam sought counsel from his advisers. The older counsellors advised moderation, the younger, retaliation. Assenting to the latter, Rehoboam returned to the people with an answer that was to lead to the disintegration of the united monarchy that had lasted for only about a century under three kings: "My father made your yoke heavy, but I will add to your yoke; my father chastised you with whips, but I will chastise you with scorpions." The response of the northerners was the ancient battle cry, "To your tents, O Israel." Rehoboam, ruling from the cities, sent Adoram, the leader of the forced labour battalions, to Israel (the name to be used henceforth for the northern area); but he was stoned to death. The uncrowned king of the north, unable to quell the rebellion, returned to Jerusalem in rapid flight. Heeding the advice of the prophet Shemaiah, Rehoboam allowed the situation to remain that of a stalemate, thus inaugurating the period of the divided monarchy that lasted in Israel in the north from 922–721 BCE and in Judah in the south until 586 BCE.

Turmoil in the monarchy of the northern kingdom

Though the Davidic monarchy continued in Judah until the fall of Jerusalem in 586 BCE, the monarchial situation in Israel was one of constant turmoil and confusion, except for the periods of a few dynasties. Jeroboam I of Israel (reigned 922–901 BCE) attempted to bring about religious and political reforms. Establishing his capital at Shechem, he set aside two pilgrimage sites (Dan in the north and Bethel in the south) as shrine centres. Though the Deuteronomic historian—with an anti-north prejudice—interpreted Jeroboam's use of golden bulls in the high place sanctuaries as a sin against Yahweh, Jeroboam's actions may have merely been an incorporation of religious symbols similar to the cherubim (winged animals) that guarded the empty throne of Yahweh in the temple of Solomon in Jerusalem. Jeroboam would not have been so politically and religiously naïve as to introduce polytheistic practices among the conservative-minded tribes of northern Israel. Thus, the golden bulls may have been meant to serve as pedestals for the invisible Yahweh just as the ark (throne) may have been the seat of the invisible Yahweh in the Holy of Holies (inner sanctuary) of the Temple in Jerusalem. Gods (such as the storm god Hadad) of other Syrian and Palestinian religions also were represented as standing on the backs of bulls.

Jeroboam remained true to Yahwistic religion, however, in that the God of the Israelites was not represented iconographically. The first king of the northern kingdom also inaugurated other religious reforms or reinstituted ancient practices that were interpreted as decadent by the Deuteronomic historian of the southern kingdom of Judah. He instituted a harvest thanksgiving festival on the 15th day of the eighth month, a change in the religious calendar that would preclude the journey of many northern Israelites to a similar festival in Jerusalem; he reformed the priesthood by installing non-Levites (the traditional shrine functionaries) to serve Yahweh at the shrines, an action that had been carried out in Jerusalem by David but without the opprobrium inferred by the Deuteronomic historian on a similar action by Jeroboam.

The dynasties of the northern kingdom were short-lived. Jeroboam was succeeded by his son Nadab, who reigned for two years before he was overthrown by Baasha, who decimated the house of Jeroboam. Reigning for 24 years, Baasha (who "did what was evil in the sight of the Lord" like all of the northern kings, according to the interpretation of the Deuteronomists) had to concern himself not only with charismatic leaders who were traditionally powerful in the north but also with the rising power of anti-monarchical prophets, such as Jehu—who prophesied the end of the house of Baasha (chapter 16). Elah, Baasha's son, ruled only two years before he was assassinated while in a drunken state by Zimri, a chariot commander, who exterminated all of the members of the house of Baasha. Reigning for the brief period of seven days, Zimri was besieged in the citadel at Tirzah by Omri, commander of the army. Zimri burned to death in the king's house. Much of this political turmoil and confusion in the north occurred during the reign of Asa, king of Judah from *c.* 913–873 BCE, who inaugurated religious reforms, such as banning male cult prostitutes and the worship of the Canaanite goddess Asherah that had been sponsored by his mother, Maachah, the queen regent.

The dynasty of Omri and the career of Elijah

The significance of Elijah. With the dynasty of Omri (*c.* 876–842), the prophetic movement begins to assume a position of tremendous importance in Israel and Judah. Omri (reigned *c.* 876–869) re-established Israel's economic and military significance among the Syrian and Palestinian minor kingdoms, so much so that years after his death the Assyrians referred to the northern kingdom as "the land of Omri." He is mentioned in the Moabite Stone of King Mesha (9th century BCE) as a king who "humbled Moab many years." To strengthen an alliance with the Phoenicians, Omri contracted a marriage between Jezebel, princess of Sidon, and his son Ahab. The marriage proved to be fateful for Israel and was a catalyst that brought the prophetic movement into a course of action and a form that became Israel's contribution to Near Eastern prophecy.

The reign of Omri's son Ahab coincided with the activities of the prophet Elijah, as recorded in I Kings, chapter 16, verse 29, to chapter 22, verse 40. Ahab, under the influence of his queen Jezebel, allowed her to foster the worship of the fertility god Baal in Samaria—the capital that Omri had built—and in all Israel, even though he himself remained a worshipper of Yahweh. A temple was built for Baal in Samaria; Jericho was rebuilt (even though the ban against its existence still remained) by Hiel of Bethel, who sacrificed two of his own sons and placed them in the foundation and the gates of the walls of the city. During these apostate activities the great prophet Elijah the Tishbite appeared. A man of erratic behaviour, wearing a garment of hair with a leather belt around his waist, using uncouth language, and preferring the wilderness areas to the towns, Elijah bore many of the outward signs of social rebels. At odds with the court authorities, he began his prophetic career just prior to a retreat in the wilderness during a drought, which he had announced to Ahab, thus pointing out that Yahweh, rather than Baal, is the Lord of nature. In the desert he performed two miracles: he ensured a widow and her son of continuous food for her act of generosity to him and cured her son, apparently dead, who had stopped breathing, by stretching himself on top of the boy three times. Elijah then went to the court of Ahab at Samaria, after having met one of the leading prophets (Obadiah) who had escaped Jezebel's attempt to destroy the leaders of the cult of Yahweh, and stood before Ahab, accusing the king of being the

"troubler of Israel" for having followed the cult of Baal. Elijah hurled a challenge to the Baalists, supported by Jezebel, to meet him in a contest on Mt. Carmel.

The Mt. Carmel contest

The contest between Elijah and the 450 prophets of Baal was dramatic. Elijah first taunted the spectators, "How long will you go limping with two different opinions? If the Lord is God, follow him; but if Baal, then follow him." Elijah then laid the ground rules: two bulls were to be sacrificed, one each on an altar, on which firewood was to be laid, but no one was to light the fire—only the God "who answers by fire." The prophets of Baal had the first opportunity, and they prayed to Baal loudly for a full half day, until noon. During this time, Elijah, in coarse language, taunted them. Eliminating the euphemisms in most English versions of the Bible, Elijah mocked the Baalists by saying that Baal might not be responding because he was out urinating ("gone aside"), on a trip, or sleeping. The Baalists then attempted to use sympathetic magic. By cutting themselves they hoped that as their life blood flowed on the ground Baal would send rain, the life blood of the Earth.

When the Baalists had failed, Elijah rebuilt an old altar of Yahweh, poured water on the wood three times (perhaps a remnant of an ancient rainmaking ceremony?), and prayed to Yahweh to answer his servant; "the fire of the Lord fell, and consumed the burnt offering, and the wood, and the stones and the dust, and licked up the water that was in the trench." Though some authorities explain the action by suggesting that Elijah poured naphtha on the wood, this does not explain the ignition of the wood at that particular time and that particular place even if by a bolt of lightning. The Deuteronomic historian emphasized the miracle wrought by Yahweh. The people, upon witnessing the miracle, cried out, "Yahweh, he is God," and proceeded to annihilate the prophets of Baal.

Elijah told Ahab to complete the festivities while he went to the top of Mt. Carmel to perform another rainmaking ceremony. When the rains came in a cloudburst, Ahab was riding in his chariot in the Valley of Jezreel. Elijah, in fear of retaliation from Jezebel, fled to the southern wilderness. At Mt. Horeb (Sinai) after a storm, wind, and an earthquake, Yahweh spoke to Elijah through silence and then revealed that he should anoint Hazael to be king of Syria, Jehu to be king of Israel, and Elisha to be his successor as prophet. I Kings, chapter 20, records a war between Ben-hadad, king of Syria, and Ahab. Though Ahab was victorious, he did not kill Ben-hadad according to the provisions of the *herem* (ban); and a prophet then informed Ahab that he would suffer for his inaction.

The affair of Naboth's vineyard

Upon Ahab's return to Samaria Jezebel attempted to coerce the king into confiscating the vineyards of Naboth of Jezreel, which was a Canaanite centre. Naboth asserted that as an Israelite the land was not his own but was a trust from Yahweh and that he could not sell it. Taken to court on trumped-up charges of blasphemy, Naboth was convicted and stoned to death. Ahab, following Jezebel's advice, then went to Naboth's vineyard and took possession of it. Upon hearing of Ahab's unjust act as king, Elijah proclaimed to him, "In the place where dogs licked up the blood of Naboth shall dogs lick your own blood." The prophet also announced, "The dogs shall eat Jezebel within the bounds of Jezreel."

In I Kings, chapter 22, another prophet, Micaiah, prophesied to Ahab and to King Jehoshaphat of Judah who were preparing for battle against the Syrians that in a vision he saw "all Israel scattered upon the mountains, as sheep that have no shepherd." Micaiah was put in prison to test the validity of his vision. It turned out to be true—Ahab, even though he disguised himself, was mortally wounded by an arrow shot by a Syrian archer. In 850 he was succeeded by his son Ahaziah, who reigned for only two years.

The Second Book of Kings continues the history of the monarchies of Israel and Judah and of the prophetic movement. Ahaziah fell from an upper chamber of his palace in Samaria and sought help from Baalzebub, the god of Ekron. Elijah met the messengers to castigate them for not seeking aid from Yahweh, the God of Israel, and told a third delegation that had been sent out to return to tell Ahaziah that because of his apostasy he would die. After the death of Ahaziah, Elijah conferred his mantle, the symbol of his prophetic authority, on Elisha, and "Elijah went up by a whirlwind into heaven."

The significance of Elisha. The stories of Elijah and his successor, Elisha, are of a different literary genre from the historical accounts of the political developments of the 9th century. The historical accounts are based on the viewpoints and biases of the monarchy, nobility, and military leaders. The stories of Elijah and Elisha are legendary, popular accounts, probably having arisen among the common people. They demonstrate the predilection of the common people to accent what appears to them as the miraculous and the supernatural, much as has been the case among many Roman Catholics and Eastern Christians in stories of their saints. Elijah was depicted, in several instances, as a second Moses—*e.g.*, he fled to the wilderness to escape the retaliation of a ruler, and he encountered a theophany (manifestation of a deity) of Yahweh on Mt. Horeb. As Moses appointed Joshua as his successor, so also Elijah passed on his prophetic mantle to Elisha. Elisha is depicted in typical folk story embellishments and legendary motifs. The original beginning and ending of the Elijah story apparently was lost, but the Deuteronomic historian incorporated the popular accounts of Elijah and Elisha into the court history that gives scholars significant insights into the religious movements of the 9th century. During the reigns of King Jehoshaphat of Judah (*c.* 873–849 BCE) and King Jehoram (Joram) of Israel (*c.* 849–842), Elisha began his prophetic career. Elisha was unlike his mentor Elijah in many ways: he did not use uncouth language, he did not shun towns, he wore more fashionable clothing, and he used music to bring about the prophetic spirit—much as Saul had done earlier. A cycle of miracle stories arose around Elisha; he was said to have made bitter water sweet, revived the son of a Shunammite woman from death by breathing into his mouth and lying on top of him, helped a woman to avoid giving up her two sons to a creditor who would make them slaves, informed the Syrian captain Naaman how to be cured from his skin disease, and many other similar actions. In addition to being a miracle worker, Elisha was a political power. He prophesied the defeat of the Moabites as a result of a huge rainfall and advised Joram how to defeat Ben-hadad, king of Syria. By performing this last act Elisha instigated a revolt in Syria; Hazael murdered the sick and dying Ben-hadad.

The prophetic career of Elisha and the end of the Omri dynasty

Elisha sent "one of the sons of the prophets" to anoint Jehu, an army commander, to be the future king of Israel. Rushing in his chariot to Jezreel, Jehu exterminated Jehoram, the last king of the Omri dynasty, his nephew Ahaziah (king of Judah), who was visiting him, and the queen mother Jezebel, who "had painted her eyes, and adorned her head" before she was thrown out of the window and so mangled by the trampling of horses that "they found no more of her than the skull and the feet and the palms of her hands." Jezebel's end had come about in a manner similar to the way in which Elijah had prophesied.

The revolution of Jehu was not only politically inspired. A driving force behind him was the arch conservative Rechabite faction, led by Jehonadab. Despising the Canaanites and their agricultural way of life, the Rechabites—descendants of the ancient Kenites of Midian where Moses had experienced the theophany of the burning bush—lived in tents, refused to drink wine, and attempted to retain as many of the accoutrements of the "good old life" of ancient nomadism as possible. With excessive revolutionary zeal they helped Jehu to annihilate the worshippers of Baal, who were tricked into coming to their temple and there murdered. To further emphasize their revolutionary intent, the followers of Jehu, in addition to the holocaust, made the site of the temple of Baal a latrine.

Because the king of Judah (Ahaziah) had been killed in the revolution—along with the remaining northern mem-

bers of the house of Omri—the southern kingdom was ruled over by the queen mother, Athaliah, the daughter of Ahab and Jezebel. In her zeal to propagate the faith of her mother, Athaliah seized the opportunity to destroy the line of David that tended to be loyal to Yahweh. Liquidating all the male heirs to the throne of David—except the infant Joash (Jehoash) who received asylum in "the house of the Lord"—Athaliah ruled for six years. With support from the priests led by Jehoiada, the army and "the people of the land" revolted, killing Athaliah and her high priest of Baal, Mattan, and destroying the temple of Baal.

The reign of Jeroboam II and the demise of the northern kingdom

In the north, Jehu was succeeded by his son Jehoahaz (reigned *c.* 815–*c.* 801), who, in turn, was followed by his son Joash, or Jehoash. During the latter king's reign, the prophet Elisha died. Though the Deuteronomic historian says little about Israel's next king, Jeroboam II, he was a major monarch, re-establishing the northern kingdom's ancient boundaries and fostering a period of economic prosperity. During the reign of Jeroboam II (*c.* 786–*c.* 746 BCE), a time of both economic advances and social injustice, Amos, the great prophet of social justice, arose. During Jeroboam's last years another great prophet, Hosea, whose message centred on Covenant love, arose to call an apostate people back to their Covenant responsibilities.

The fall of Israel. After the death of Jeroboam II, however, Israel faced a period of continuous disaster; and no prophetic figure was able to arrest the steady internal decay. From 746–721, when Samaria finally fell to the Assyrians, there were six kings, the last being Hoshea, a conspirator who had assassinated the previous king. The Assyrian king Sargon II deported the leading citizens of Samaria to Persia and imported colonists from other lands to fill their places.

The fall of Judah. The southern kingdom of Judah, under the Davidic monarchy, was able to last about 135 years longer, often only as a weak vassal state. Hezekiah (reigned *c.* 715–*c.* 687), with the advice of the prophet Isaiah, managed to avoid conflict with or outlast a siege of the Assyrians. Hezekiah was succeeded by his son Manasseh, an apostate king who stilled any prophetic outcries, reintroduced Canaanite religious practices, and even offered his son as a human sacrificial victim. Soothsaying, augury, sorcery, and necromancy were also reintroduced. The Deuteronomic historian also notes that many innocent persons were killed during his reign. Manasseh was succeeded by his son Amon, who was assassinated in a palace revolution after a reign of only two years. His son Josiah, who succeeded him, reigned from 640–609 BCE, when he was killed in a battle with the pharaoh Necho II of Egypt. During his reign, one of the most significant events in the history of the Israelite people occurred—the Deuteronomic reform of 621 BCE. Occasioned by the discovery of a book of the Law in the Temple during its rebuilding and supported not only by Hilkiah, a high priest, and Huldah, a prophetess, but also by the young prophet Jeremiah, the Deuteronomic Code—or Covenant—as it has been called, became the basis for a far-reaching reform of the social and religious life of Judah. Though the reform was short-lived, because of the pressure of international turmoil, it left an indelible impression on the religious consciousness of the people of the Covenant, Israel, whether they were from the north or the south.

Reforms in Judah and the Babylonian Exile

From 609–586 Judah felt the coming oppression of Babylon under King Nebuchadrezzar. After the death of Josiah, four kings ruled in Jerusalem, the last being Zedekiah, who failed to heed the advice of the prophet Jeremiah—who had attempted to persuade the king not to trust the Egyptians in a rebellion against Babylon because there would be only one loser, the House of David. Jehoiakim, the predecessor of the puppet king Zedekiah, had been carried off into exile to Babylon in 598; but about 560 he was released from prison, thus leaving a hope that the Davidic line had not become extinct. Despite this small element of hope, the year 586 BCE marked the beginning of a tragic period for the people of Judah—the Babylonian Exile. During this period of rethinking the Covenant faith, the prophet Ezekiel preached, both in Jerusalem and Babylon, offering the people hope for a restoration of the symbols and cultic acts of their covenant religion.

Isaiah. The Book of Isaiah, comprising 66 chapters, is one of the most profound theological and literarily expressive works in the Bible. Compiled over a period of about two centuries (the latter half of the 8th to the latter half of the 6th centuries BCE), the Book of Isaiah is generally divided by scholars into two (sometimes three) major sections, which are called First Isaiah (chapters 1–39), Deutero-Isaiah (chapters 40–55 or 40–66), and—if the second section is subdivided—Trito-Isaiah (chapters 56–66).

The division of Isaiah into two or three distinct sections

The prophecies of First Isaiah. First Isaiah contains the words and prophecies of Isaiah, a most important 8th-century BCE prophet of Judah, written either by himself or his contemporary followers in Jerusalem (from *c.* 740–700 BCE), along with some later additions, such as chapters 24–27 and 33–39. The first of these two additions was probably written by a later disciple or disciples of Isaiah about 500 BCE; the second addition is divided into two sections—chapters 33–35, written during or after the exile to Babylon in 586 BCE, and chapters 36–39, which drew from the source used by the Deuteronomic historian in II Kings, chapters 18–19. The second major section of Isaiah, which may be designated Second Isaiah even though it has been divided because of chronology into Deutero-Isaiah and Trito-Isaiah, was written by members of the "school" of Isaiah in Babylon: chapters 40–55 were written prior to and after the conquest of Babylon in 539 by the Persian king Cyrus II the Great, and chapters 56–66 were composed after the return from the Babylonian Exile in 538. The canonical Book of Isaiah, after editorial redaction, probably assumed its present form during the 4th century BCE. Because of its messianic (salvatory figure) themes, Isaiah became extremely significant among the early Christians who wrote the New Testament and the sectarians at Qumrān near the Dead Sea, who awaited the imminent messianic age, a time that would inaugurate the period of the Last Judgment and the Kingdom of God.

Isaiah, a prophet, priest, and statesman, lived during the last years of the northern kingdom and during the reigns of four kings of Judah: Uzziah (Azariah), Jotham, Ahaz, and Hezekiah. He was also a contemporary of the prophets of social justice: Amos, Hosea, and Micah. Influenced by their prophetic outcries against the horrors of social injustice, Isaiah added themes that were peculiar to his prophetic mission. To kings, political and economic leaders, and to the people of the land, he issued a message that harked back nearly five centuries to the period of the judges: the holiness of Yahweh, the coming Messiah of Yahweh, the judgment of Yahweh, and the necessity of placing one's own and the nation's trust in Yahweh rather than in the might of ephemeral movements and nations. From about 742 BCE, when he first experienced his call to become a prophet, to about 687, Isaiah influenced the course of Judah's history by his oracles of destruction, judgment, and hope as well as his messages containing both threats and promises.

Themes peculiar to Isaiah and his call to become a prophet

Intimately acquainted with worship on Mt. Zion because of his priest-prophet position, with the Temple and its rich imagery and ritualistic practices, and with a deep understanding of the meaning of kingship in Judah theologically and politically, Isaiah was able to interpret and advise both leaders and the common people of the Covenant promises of Yahweh, the Lord of Hosts. Because they were imbued with the following beliefs—God dwelt on Mt. Zion, in the Temple in the city of Jerusalem, and in the person of the King—the messianic phrase "God is with us" (Immanuel) Isaiah used was not a pallid abstraction of a theological concept but a concrete living reality that found its expression in the Temple theology and message of the great prophet.

In chapters 1–6 are recorded the oracles of Isaiah's early ministry. His call, a visionary experience in the temple in Jerusalem, is described in some of the most influential symbolic language in Old Testament literature. In the

year of King Uzziah's death (742 BCE), Isaiah had a vision of the Lord enthroned in a celestial temple, surrounded by the seraphim—hybrid human-animal-bird figures who attended the deity in his sanctuary. Probably experiencing this majestic imagery that was enhanced by the actual setting and the ceremonial and ritualistic objects of the Jerusalem Temple, Isaiah was mystically transported from the earthly temple to the heavenly temple, from the microcosm to the macrocosm, from sacred space in profane time to sacred space in sacred time.

Yahweh, in the mystical, ecstatic experience of Isaiah, is too sublime to be described in other than the imagery of the winged seraphim, which hide his glory and call to each other:

"Holy, holy, holy is the Lord of hosts;
The whole earth is full of his glory."

With smoke rising from the burning incense, Isaiah was consumed by his feelings of unworthiness ("Woe is me! for I am lost"); but one of the seraphim touched Isaiah's lips with a burning coal from the altar and the prophet heard the words, "Your guilt is taken away, and your sin forgiven." Isaiah then heard the voice of Yahweh ask the heavenly council, "Whom shall I send, and who will go for us?" The prophet, caught up as a participant in the mystical dialogue, responded, "Here am I! Send me." The message to be delivered to the Covenant people from the heavenly council, he is informed, is one that will be unheeded.

The early oracles of Isaiah

The oracles of Isaiah to the people of Jerusalem from about 740 to 732 BCE castigate the nation of Judah for its many sins. The religious, social, and economic sins of Judah roll from the prophet's utterances in staccato-like sequence: (1) "Bring no more vain offerings; incense is an abomination to me. New moon and sabbath and the calling of assemblies—I cannot endure iniquity and solemn assembly," against religious superficiality; (2) "cease to do evil, learn to do good; seek justice, correct oppression; defend the fatherless, plead for the widow," against social injustice; and (3) "Come now, let us reason together, says the Lord: though your sins are like scarlet, they shall be as white as snow," a call for obedience to the Covenant. The prophet also cried out for peace: "and they shall beat their swords into plowshares, and their spears into pruning hooks; nation shall not lift up sword against nation, neither shall they learn war anymore." The sins of Judah, however, are numerous: the rich oppress the poor, the nation squanders its economic resources on military spending, idolatry runs rampant in the land, everyone tries to cheat his fellowman, women flaunt their sexual charms in the streets, and there are many who cannot wait for a strong drink in the morning to get them through the day. One of Isaiah's castigations warns: "Woe to those who are heroes at drinking wine, and valiant men in mixing strong drink, who acquit the guilty for a bribe, and deprive the innocent of his right!"

Isaiah's prophetic attacks on Judah's foreign policy

During the Syro-Ephraimitic war (734–732 BCE), Isaiah began to challenge the policies of King Ahaz of Judah. Syria and Israel had joined forces against Judah. Isaiah's advice to the young King of Judah was to place his trust in Yahweh. Apparently Isaiah believed that Assyria would take care of the northern threat. Ahaz, in timidity, did not want to request a sign from Yahweh. In exasperation Isaiah told the King that Yahweh would give him a sign anyway: "Behold, a young woman shall conceive and bear a son, and shall call his name Immanuel." Thus, by the time this child is able to know how to choose good and refuse evil, the two minor kings of the north who were threatening Judah will be made ineffective by the Assyrians. The name Immanuel, "God is with us," would be meaningful in this situation because God on Mt. Zion and represented in the person of the king would be faithful to his Covenant people. Ahaz, however, placed his trust in an alliance with Assyria under the great conqueror Tiglath-pileser III. In order to give hope to the people, who were beginning to experience the Assyrian encroachments on Judaean lands in 738 BCE, Isaiah uttered an oracle to "the people who walked in darkness": "For to us a child is born, to us a son is given; and the government will be upon his shoulder, and his name will be called Wonderful Counselor, Mighty God, Everlasting Father, Prince of Peace." Isaiah trusted that Yahweh would bring about a kingdom of peace under a Davidic ruler.

From 732 to 731 BCE, the year the northern kingdom fell, Isaiah continued to prophesy in Judah but probably not in any vociferous manner until the Assyrians conquered Samaria. The king of the Assyrians is described as the rod of God's anger, but Assyria also will experience the judgment of God for its atrocities in time of war. During one of the periods of Assyrian expansion towards Judah, Isaiah uttered his famous Davidic messianic (salvatory figure) oracle in which he prophesies the coming of a "shoot from the stump of Jesse," upon which the Spirit of the Lord will rest and who will establish the "peaceable kingdom" in which "the wolf shall dwell with the lamb." A hymn of praise concludes this first section of First Isaiah.

Chapters 13–23 include a list of oracles against various nations—Babylon, Assyria, Philistia, Moab, Syria, Egypt, and other oppressors of Judah. These probably came from the time when Hezekiah began his reign (*c.* 715). In 705 BCE, Sargon of Assyria died, however, and Hezekiah, a generally astute and reform-minded king, began to be caught up in the power struggle between Babylon, Egypt, and Assyria. Isaiah urged Hezekiah to remain neutral during the revolutionary turmoil. Though Sennacherib of Assyria moved south to crush the rebellion of the Palestinian vassal states, Isaiah—contrary to his previous advocacy of neutrality—urged his king to resist the Assyrians because the Lord, rather than the so-called Egyptian allies, who "are men, and not God," will protect Jerusalem. He then prophesied a coming age of justice and of the Spirit who will bring about a renewed creation.

Second Isaiah (chapters 40–66), which comes from the school of Isaiah's disciples, can be divided into two periods: Chapters 40–55, generally called Deutero-Isaiah, were written about 538 BCE after the experience of the Exile; and chapters 56–66, sometimes called Trito-Isaiah (or III Isaiah), were written after the return of the exiles to Jerusalem after 538 BCE.

Deutero-Isaiah and the Servant Songs

The prophecies of Deutero-Isaiah. Second Isaiah contains the very expressive so-called Servant Songs—chapter 42, verses 1–4; chapter 49, verses 1–6; chapter 50, verses 4–9; chapter 52, verse 13, and chapter 53, verse 12. Writing from Babylon, the author begins with a message of comfort and hope and faith in Yahweh. The people are to leave Babylon and return to Jerusalem, which has paid "double for all her sins." As creator and Lord of history, God will redeem Israel, his chosen servant. Through the Servant of the Lord all the nations will be blessed: "I have put my Spirit upon him, he will bring forth justice to the nations." The Suffering Servant, whether the nation Israel or an individual agent of Yahweh, will help to bring about the deliverance of the nation. Though Second Isaiah may have been referring to a hoped-for rise of a prophetic figure, many scholars now hold that the Suffering Servant is Israel in a collective sense. Christians have interpreted the Servant Songs, especially the fourth, as a prophecy referring to Jesus of Nazareth—"He was despised and rejected by men; a man of sorrows and acquainted with grief . . . ," but this interpretation is theologically oriented and thus open to question, according to many scholars.

The oracles of Trito-Isaiah. Chapters 56–66 are a collection of oracles from the restoration period (after 538 BCE). Emphasis is placed upon cultic acts, attacks against idolatry, and a right motivation in the worship of Yahweh. Repentance and social justice are themes that have been retained from the earlier Isaiah traditions, and the ever-present element of hope in the creative goodness of Yahweh that pervaded II Isaiah remains a dominant theme in the last chapters of the Book of Isaiah (see also ISAIAH).

Jeremiah. The prophet Jeremiah began to prophesy about 626 BCE during the reign of the Judaean king Josiah. From the town of Anathoth and probably from the priestly family of Eli, this prophet, who may have been instrumental in the Deuteronomic reform, dictated his

The call of Jeremiah and his difficulties with political and revolutionary leaders

oracles to his secretary Baruch. Only a youth in his late teens when he experienced the call by Yahweh to be a "prophet to the nations," Jeremiah was a hesitant reforming prophet, experiencing deep spiritual struggles regarding his adequacy as a prophetic leader from the very beginning of his call and throughout his prophetic ministry. After the death of Josiah in 609 BCE, however, he became an outspoken prophet against the national policy of Judah, a policy that he knew would lead to the disaster that came to be called the Babylonian Exile. Because of his prophecies, which were unpopular with the military and the revolutionists against the Babylonians, Jeremiah was kidnapped by conspirators after 586 and taken to Egypt, where he disappeared from sight.

The Book of Jeremiah is a collection of oracles, biographical accounts, and narratives that are not arranged in any consistent chronological or thematic order. One 20th-century German biblical scholar, Wilhelm Rudolph, has attempted to arrange the chapters of the book according to certain chronological details. He has divided the work into five sections: (1) prophecies against Judah and Jerusalem, chapters 1–25, during the reigns of kings Josiah (640–609) and Jehoiakim (609–598), and the period after Jehoiakim (597–586); (2) prophecies against foreign nations, chapter 25 and 66–61; (3) prophecies of hope for Israel, chapters 26–35 (probably after the death of Josiah in 609); (4) narratives of Jeremiah's sufferings, chapters 36–45 (from a post-586 period), and (5) an appendix, chapter 52. Jeremiah's own prophetic oracles are found particularly in chapters 1–36 and 46–52. Baruch's writings about Jeremiah are found primarily in chapters 37–45, 26–29, and 33–36.

During the reign of Josiah, after his call, Jeremiah preached to the people of Jerusalem and warned them against the sin of apostasy. Recalling the prophecies of the 8th-century Israelite prophet Hosea, Jeremiah reproached the Judaeans for playing harlot with other gods and urged them to repent. He prophesied that enemies from the north would be the instruments of Yahweh's judgment on the apostate land and Jerusalem would suffer the fate of a rejected prostitute. The idolatry and immorality of the Judaeans would inevitably lead to their destruction. Because of the impending threat from the north, Jeremiah warned the people to flee from the wrath that was to come.

At the beginning of Jehoiakim's reign, Jeremiah preached in the temple that because of Judah's apostasy "death shall be preferred to life by all the remnant that remains of this evil family in all the places where I have driven them, says the Lord of hosts." Because he spoke words that were unpopular, his own townsmen of Anathoth plotted against his life. To symbolize the fate of Judah, Jeremiah adopted some rather bizarre techniques. He buried a waist cloth and wore it when it was spoiled to illustrate the fate of Jerusalem, which had worshipped other gods than Yahweh.

Jeremiah's depression and allegorical messages

Throughout his career Jeremiah had moments of deep depression, times when he lamented that he had become a prophet. Because of the uncertainty of the times, Jeremiah did not marry.

A master of symbolic actions and the use of symbolic devices, Jeremiah used a potter's wheel to show that Yahweh was shaping an evil future for Judah; and he bought a flask, after which he broke it on the ground to illustrate again the fate of Judah. Because of such words and actions, Jeremiah often found himself in trouble. Pashur, a priest, had Jeremiah beaten and placed in stocks. When released, Jeremiah told Pashur he would go into captivity and die. Despite the plots against him, Jeremiah continued to rely on the grace of Yahweh. He was brought to trial for prophesying the destruction of Jerusalem, but his defense attorneys—"certain of the elders"—pointed out that King Hezekiah had not punished the prophet Micah of Moresheth in the 8th century for similar statements.

Continuing to prophesy against the moral and religious corruption of Jerusalem during the reign of Zedekiah (597–586), Jeremiah became even more unpopular for his advocacy to surrender to Babylon.

In spite of his apparent failure to win over the people to his cause, Jeremiah inaugurated a reform that had lasting effects. He helped to bring about a change in religion from the view that primarily accepted corporate responsibility to one that held that religion is more individualistic in terms of responsibility. His words in chapter 31, verse 33, are a summation of his reform: "But this is the covenant which I will make with the house of Israel after those days, says the Lord: I will put my law within them, and I will write it upon their hearts; and I will be their God, and they shall be my people" (see also JEREMIAH).

Ezekiel. The Book of Ezekiel, written by the prophet-priest Ezekiel, who lived both in Jerusalem prior to the Babylonian Exile (586 BCE) and in Babylon after the Exile, and also by an editor (or editors), who belongs to a "school" of the prophet similar to that of the prophet Isaiah, has captured the attention of readers for centuries because of its vivid imagery and symbolism. The book has also attracted the attention of biblical scholars who have noticed that although Ezekiel appears to be a singularly homogeneous composition displaying a unity unusual for such a large prophetic work, it also displays, upon careful analysis, the problem of repetitions, certain inconsistencies and contradictions, and questions raised by terminological differences. Though the book itself indicates that the prophecies of Ezekiel occurred from about 593–571 BCE, some scholars—who are in a minority—have argued that the book was written during widely divergent periods, such as in the 7th century and even as late as the 2nd century BCE. Most scholars, however, accept that the main body of the book came from the 6th century BCE, with the inclusion of some later glosses by redactors who remained loyal to the theological traditions of their master-teacher.

Stylistic and literary problems in Ezekiel

Containing several literary genres, such as oracles, mythological themes, allegory, proverbs, historical narratives, folk tales, threats and promises, and lamentations, the Book of Ezekiel is divided into three main sections: (1) prophecies against Judah and Jerusalem (chapters 1–24); (2) prophecies against foreign countries (chapters 25–32); and (3) prophecies about Israel's future.

Ezekiel—the man and his message. The man who wrote this book—at least the main body of the work—was undoubtedly one of the leaders of Jerusalem because he was among the first group of exiles to go into captivity—those who were forced to leave their homeland about 597 BCE in a deportation to Babylon on the orders of the conquering king Nebuchadrezzar. Belonging to the priestly class, perhaps of the line of Zadok, Ezekiel was a spiritual leader of his fellow exiles at Tel-abib, which was located near the river Chebar, a canal that was part of the Euphrates River irrigation system. According to his own account, Ezekiel, the priest without a temple, received the call to become a prophet during a vision "In the thirtieth year, in the fourth month, on the fifth day"—perhaps July 31, 593 BCE, if the dating is based on the lunar calendar, though the exact meaning of "thirtieth year" remains obscure. A married man who was often consulted by elders among the exiles, Ezekiel carried out his priestly and prophetic career during two distinct periods: (1) from 593–586 BCE, a date that was doubly depressing for the prophet because it was the period when his wife died and his native city was destroyed; and (2) from 586–571 BCE, the date of his last oracle (chapter 29, verse 17).

The personality of the prophet shows through his oracles, visions, and narrations. Frustrated because the people would not heed his messages from Yahweh, Ezekiel often exhibited erratic behaviour. This need not mean that he was psychologically abnormal. Like many great spiritual leaders, he displayed qualities and actions that did not fall within the range of moderation, and to perform an ex post facto psychological postmortem examination on any great historical figure in the face of a paucity of necessary details may be an interesting game but is hardly scientifically respectable or accurate. To be sure, Ezekiel did engage in erratic behaviour: he ate a scroll on one occasion, lost his power of speech for a

period of time, and lay down on the ground "playing war" to emphasize a point, an action that would certainly draw attention to him, which was his purpose. In spite of these peculiarities, Ezekiel was a master preacher who drew large crowds and a good administrator of his religious community of exiles. He held out hope for a temple in a new age in order to inspire a people in captivity. He also initiated a form of imagery and literature that was to have profound effects on both Judaism and Christianity all the way to the 20th century: apocalypticism (the view that God would intervene in history to save the believing remnant and that this intervention would be accompanied by dramatic, cataclysmic events).

Ezekiel's mystical call and symbolic acts

Prophetic themes and actions. The first section of the book (chapters 1–24) contains prophecies against Judah and Jerusalem. Ezekiel's call is recorded in chapter 1 to chapter 3, verse 15. It came in a vision of four heavenly cherubim, who appeared in a wind from the north, a cloud, and flashing fire (lightning?)—traditional symbolic elements of a theophany (manifestation of a god) in ancient Near Eastern religions. These winged hybrid throne bearers—with the faces of a man, a lion, an ox, and an eagle (which became iconographic symbols of the four Gospel writers of the New Testament)—bore the throne chariot of Yahweh. The cherubim, symbolizing intelligence, strength, and—especially—mobility, had beside them four gleaming wheels, or "a wheel within a wheel" (*i.e.*, set at right angles to each other), which further emphasized the omnimobility of the throne chariot. This vision harks back to Isaiah's mystical experience (Isaiah chapter 6) in which that prophet envisioned the throne of the ark, which symbolized the omnipresence of the invisible Yahweh. High above the cherubim was a firmament, or crystal platform, above which was the throne of Yahweh, who—in a "likeness as if it were of a human form"—spoke to Ezekiel. The Spirit of Yahweh entered him, and he was commissioned to preach to the people of Israel a message of doom to an apostate people. The significance of this vision is that it occurred not to a priest in the holy Temple at Jerusalem but to an exiled prophet-priest in a foreign land. The God of Israel was the God of the nations. The impact of his visionary experience so overwhelmed Ezekiel that he simply sat at Tel-abib for seven days.

Commissioned by Yahweh to be "a watchman for the house of Israel," Ezekiel performed a series of symbolic acts to illustrate the impending fate of the city from which he had been banished: he placed a brick on the ground to symbolize Jerusalem's future siege, lay down on the ground, bound himself to indicate capture, ate food first cooked on fuel composed of human feces and then animal excrement, and then cut his hair and beard. Though these acts were performed in Babylon, news of them was most likely communicated to the people of Jerusalem. Just as Jeremiah had tried to repress the false hopes that the residents of Jerusalem harboured concerning the downfall of Babylon, which had been predicted by the popular nationalistic prophet Hananiah (Jeremiah, chapter 28, verses 5–17), Ezekiel attempted to quash the ill-founded aspirations of the exiles for an immediate return to Jerusalem.

In chapters 6 and 7 Ezekiel prophesies that Jerusalem's "altars shall become desolate," its people will be "scattered through the countries," and "because the land is full of bloody crimes and the city full of violence," Yahweh "will put an end to their proud might and their holy places shall be profane." In chapter 8 he attacked the people of Jerusalem for their idolatry, as manifest by the women sitting before the entrance to the north gate of the Temple of Yahweh weeping in cultic despair for the Mesopotamian fertility deity Tammuz's "annual death."

After prophesying the fall of Jerusalem in chapters 9–11 because "the guilt of the house of Israel and Judah is exceedingly great," Ezekiel performed other symbolic acts such as packing baggage for an emergency exile, digging a hole in his house to illustrate the fact that some will try to escape, and eating and drinking with trembling actions to show the future fear that the Jerusalemites will experience; he also attacked prophets who gave the people false hopes. "Woe to the foolish prophets who follow their own spirit, and have seen nothing. Your prophets have been like foxes among ruins, O Israel." He tried to underline his message of urgency by relating the problem of apostasy to similar situations in Israel's past history.

The death of Ezekiel's wife and the fall of Jerusalem

About the time that Nebuchadrezzar besieged Jerusalem, Ezekiel's wife became ill. Though Ezekiel could mourn her impending death "but not aloud" (*i.e.*, only by himself so that the people would notice his unusual reaction and thus receive the full impact of his prophetic message), he was not to mourn her death publicly. When he did not eat the "bread of mourners" the people asked him for an explanation. He told them, and it was a shattering exposure: Jerusalem would be destroyed "and your sons and daughters whom you left behind shall fall by the sword"; when this happens—in spite of their pining and groaning—they will know the meaning of Ezekiel's actions.

In order to show that Yahweh was the Lord of the whole creation and of all nations, Ezekiel issued prophecies of impending disasters that would be experienced by many neighbouring Near Eastern countries. Nations that exulted in Judah's defeat—*i.e.*, Ammon, Moab, Edom, Philistia, and Phoenicia—would all suffer the same fate, as well as Egypt, the formerly great empire that had manoeuvred Judah into its disastrous foreign policy of opposing Babylon.

Ezekiel's oracular imagery

Oracles of hope. In the third section, chapters 33–48, Ezekiel proclaimed, in oracles that have become imprinted in theological discourse and folk songs, the hope that lies in the faith that God cares for his people and will restore them to a state of wholeness. As the good shepherd, God will feed his flock and will "seek the lost," "bring back the strayed," "bind up the crippled," and "strengthen the weak." He will also "set up over them one shepherd, my servant David, and he shall feed them." This Davidic ruler will be a *nasi* (prince), the term used for a leader of the tribal confederacy before the inauguration of the monarchy. In chapter 37, Ezekiel had a now-famous vision of the valley of dry bones, which refers not to resurrection from the dead but rather to the restoration of a scattered Covenant people into a single unity. To further emphasize the restoration of the scattered people of Yahweh, Ezekiel uttered the oracle of the two sticks joined together into one, which prophesied the re-unification of Israel and Judah as one nation. Chapters 38 and 39 contain a cryptic apocalyptic oracle about the invasion of an unidentified Gog of Magog. Who this Gog is has long been a matter of speculation; whoever he is, his chief characteristic is that he is the demonic person who leads the forces of evil in the final battle against the people of God. Gog and Magog have thus earned a position in apocalyptic literature over the centuries. Chapters 40–48 are a closing section in which Ezekiel has a vision of a restored Temple in Jerusalem with its form of worship re-established and a restored Israel, with each of the ancient tribes receiving appropriate allotments. Ezekiel's prophecies while in exile in Babylon were to have a significant influence on the religion of Judaism as it emerged from a time of reassessment of its religious beliefs and cultic acts during the Babylonian Exile (586–538 BCE) (see also EZEKIEL).

The Twelve. *Hosea.* The Book of Hosea, the first of the canonical Twelve (Minor) Prophets, was written by Hosea (whose name means "salvation," or "deliverance"), a prophet who lived during the last years of the age of Jeroboam II in Israel and the period of decline and ruin that followed the brief period of economic prosperity. The Assyrians were threatening the land of Israel and the people of the Covenant acted as though they were oblivious to the stipulations of their peculiar relation to Yahweh. The Book of Hosea is a collection of oracles composed and arranged by Hosea and his disciples. Like his contemporary Amos, the great prophet of social justice, Hosea was a prophet of doom; but he held out a hope to the people that the Day of Yahweh contained not just retribution but also the possibility of renewal.

His message against Israel's "spirit of harlotry" was dramatically and symbolically acted out in his personal life.

The Book of Hosea is divided into two sections: (1) Hosea's marriage and its symbolic meaning (chapters 1–3); and (2) judgments against an apostate Israel and hope of forgiveness and restoration (chapters 4–14).

The symbolism of Hosea's marriage and the emphasis on Covenant love and knowledge of God

In the first section, Hosea is commanded by Yahweh to marry a prostitute by the name of Gomer as a symbol of Israel's playing the part of a whore searching for gods other than the one true God. He is to have children by her. Three children are born in this marriage. The first, a son, is named Jezreel, to symbolize that the house of Jehu will suffer for the bloody atrocities committed in the Valley of Jezreel by the founder of the dynasty when he annihilated the house of Omri. The second, a daughter, is named Lo Ruḥama (Not pitied), to indicate that Yahweh was no longer to be patient with Israel, the northern kingdom. The third child, a son, is named Lo 'Ammi (Not my people), signifying that Yahweh was no longer to be the God of a people who had refused to keep the Covenant. In chapter 2, Hosea voiced what probably was a divorce formula—"she is not my wife, and I am not her husband"—to indicate that he had divorced his faithless wife Gomer, who kept "going after other lovers." The deeper symbolism is that Israel had abandoned Yahweh for the cult of Baal, celebrating the "feast days of Baal." Just as Yahweh will renew his Covenant with Israel, however, Hosea buys a woman for a wife—probably Gomer. The woman may have been a sacred prostitute in a Baal shrine, a concubine, or perhaps even a slave. He confines her for a period of time so that she will not engage in any attempt to search for other paramours and thus commit further adulteries.

The second section, chapters 4–14, does not refer to the marriage motif; but the imagery and symbolism of marriage constantly recur. The Israelites, in "a spirit of harlotry," have gone astray and have left their God." Their infidelity emphasized their lack of trustworthiness and real knowledge of love, a love that could not be camouflaged by superficial worship ceremonies. Thus, Hosea emphasized two very significant theological terms: *ḥesed*, or "Covenant love," and "knowledge of God." In attacking the superficiality of much of Israel's worship, Yahweh, through Hosea, proclaimed: "For I desire steadfast (Covenant) love and not sacrifice, the knowledge of God, rather than burnt offerings." Because they have broken Yahweh's Covenant and transgressed his law, however, the Lord's anger "burns against them." For "they sow the wind and they shall reap the whirlwind." Israel will be punished for its rebellion and iniquities, but Hosea's message holds out the hope that the holiness of Yahweh's love—including both judgment and mercy—will effect a triumphant return of Israel to her true husband, Yahweh.

Joel. The Book of Joel, the second of the Twelve (Minor) Prophets, is a short work of only three chapters. The dates of Joel (whose name means "Yahweh is God") are difficult to ascertain. Some scholars believe that the work comes from the Persian period (539–331 BCE); others hold that it was written soon after the fall of Jerusalem in 586 BCE. His references to a locust plague may refer to an actual calamity that occurred; the prophet used the situation to call the people to repentance and lamentation, perhaps in connection with the festival of the New Year, the "Day of Yahweh." " 'Yet even now,' says the Lord, 'return to me with all your heart, with fasting, with weeping, and with mourning; and rend your hearts and not your garments.' " Some scholars, however, believe that the plague of locusts refers to the armies of a foreign power (Babylonia?). In the remaining section of the book (chapter 2, verse 30 to chapter 3, verse 21), Joel, in apocalyptic imagery, predicts the judgment of the nations—especially Philistia and Phoenicia—and the restoration of Judah and Jerusalem.

Amos. The Book of Amos, the third of the Twelve (Minor) Prophets, has been one of the most significant and influential books of the Bible from the time it was written (8th century BCE) down to the 20th century. Comprising only nine chapters of oracles, it was composed during the age of Jeroboam II, king of Israel from 786 to 746 BCE. His reign was marked by great economic prosperity, but the rich were getting richer and the poor poorer. Social injustice ran rampant in the land. The economically weak could find no redress in the courts and no one to champion their cause—until the coming of Amos, a shepherd from Tekoa in Judah who also said that he was "a dresser of sycamore trees." Amos, thus, was no professional prophet nor a member of a prophetic guild.

The call for social justice by Amos

The book is divided into three sections: (1) oracles against foreign nations and Israel (chapters 1–2); (2) oracles of indictment against Israel for her sins and injustices (chapters 3–6); and (3) visions and words of judgment (chapters 7–9). Amos was the first of the writing prophets, but his work may be composed of oracles issued both by himself and disciples who followed his theological views.

His prophetic oracles begin with a resounding phrase: "The Lord roars from Zion." He then goes on to indict various nations—Syria, Philistia, Tyre, Ammon, and Moab—for the crimes and atrocities they have committed in times of peace: "Because they sell the righteous for silver, and the needy for a pair of shoes—they . . . trample the head of the poor into the dust of the earth, and turn aside the way of the afflicted" (chapter 2, verses 6–7).

The second section (chapters 3–6) contains some of the most vehement and cogent invectives against the social injustices perpetrated in Israel. Though the Israelites have prided themselves on being the elect of God, they have misinterpreted this election as privilege instead of responsibility. In chapter 4, Amos, in language that was sure to raise the ire of the privileged classes, attacked unnecessary indulgence and luxury. To the wealthy women of Samaria he said: "Hear this word, you cows of Bashan, who are in the mountain of Samaria, who oppress the poor, who crush the needy, who say to their husbands, 'Bring, that we may drink!' " (chapter 4, verse 1). After a series of warnings of punishment, Amos proclaimed the coming of the day of Yahweh, which is "darkness, and not light." His attacks against superficial pretenses to worship have become proverbial: "I hate, I despise your feasts, and I take no delight in your solemn assemblies" (chapter 5, verse 21). Another verse from Amos has become a rallying cry for those searching for social justice: "But let justice roll down like waters, and righteousness like an ever-flowing stream" (chapter 5, verse 24).

The third section (chapters 7–9) contains visions of locusts as a sign of punishment, a summer drought as a sign of God's wrath, and a plumb line as a sign to test the faithfulness of Israel. The priest of the shrine at Bethel, Amaziah, resented Amos' incursion on his territory and told him to go back to his home in the south. In reply to Amaziah, Amos prophesied the bitter end of Amaziah's family. Another vision in chapter 8, that of a basket of ripe fruit, pointed to the fact that Israel's end was near. A fifth vision, depicting the collapse of the Temple in Samaria, symbolized the collapse of even the religious life of the northern kingdom. He ended his work with a prophecy that the Davidic monarchy would be restored.

Obadiah. The Book of Obadiah, the fourth book of the Twelve (Minor) Prophets, contains only 21 verses. Nothing is known about the prophet as a person or about his times. It may have been written before the Exile, though many scholars believe that it was composed either some time after 586 BCE or in the mid-5th century, when the Jews returned to the area around Jerusalem. The prophet concentrates on the judgment of God against Edom and other nations, with the final verses referring to the restoration of the Jews in their native land.

Jonah. The Book of Jonah, containing the well-known story of Jonah in the stomach of a fish for three days, is actually a narrative about a reluctant prophet. This fifth book of the Twelve (Minor) Prophets contains no oracles and is thus unique among prophetic books. In

II Kings, chapter 14, verses 25–27, there is a reference to a prophet Jonah who lived during the early part of the reign of Jeroboam II (8th century BCE).

Jonah and the universal reign of Yahweh

The story, however, probably comes from a time after the fall of Jerusalem in 586 BCE. Probably living during the Exile, the author used the memory of the hated Assyrians to proclaim the mission of Israel—to teach all nations about the mercy and forgiveness of God. In the short book of four chapters, Jonah, Amittai's son, is commissioned by Yahweh to go to Nineveh, the capital of Assyria, to preach repentance. Attempting to avoid the command of Yahweh, Jonah boarded a ship, which soon was caught up in a storm. The frightened sailors drew lots to discover who was the cause of their unfortunate and calamitous condition. Jonah drew the unlucky lot and was thrown overboard, after which he was swallowed by a fish and stayed in that uncomfortable place for three days and nights. After he cried to the Lord to let him out, the fish vomited Jonah out onto dry land. Jonah, though still reluctant, went to Nineveh to preach repentance. His efforts were successful, which did not please him—because of his hatred for the Assyrians. In the end, however, Jonah realized that God was a universal God, and not the sole property of Israel.

Probably written sometime between 500 and 350 BCE (or, perhaps 250 BCE), the message of Jonah protested the exclusiveness of a post-exilic Judaism, with its policy of a pure blood race of Jews that the reformers Ezra and Nehemiah had implemented in the 5th century.

Micah. The Book of Micah, the sixth book of the Twelve (Minor) Prophets, was written by the prophet Micah in the 8th century BCE. Composed of seven chapters, the book is similar in many ways to the Book of Amos. Micah attacked the corruption of those in high places and social injustice, and the book is divided into two sections: (1) judgments against Judah and Jerusalem (chapters 1–3); and (2) promises of restoration for Judah and judgments against other nations (chapters 4–7).

Micah's attacks on corruption and his summation of the prophetic message

In the first section, Micah of Moresheth utters oracles against the corrupt religious and political leaders of Israel and Judah. He also attacks the prophets who attempted to give the people false hopes: "Thus says the Lord concerning the prophets who lead my people astray, who cry 'Peace' when they have something to eat, but declare war against him who puts nothing into their mouths . . . the seers shall be disgraced, and the diviners put to shame" (chapter 3, verses 5–7). In the second section, Israel's future is predicted as being glorious, and it is told that out of Bethlehem will come a ruler of the line of David who will bring peace to the earth. Though he issues an indictment against Judah for its idolatries, Micah proclaims what is necessary to renew the Covenant relationship between God and Israel; "and what does the Lord require of you but to do justice, and to love kindness, and to walk humbly with your God?" (chapter 6, verse 8). In this verse, Micah has given a brief summation of the messages of Amos, Hosea, and Isaiah.

Nahum. The Book of Nahum, seventh of the Twelve (Minor) Prophets, contains three chapters directed against the mighty nation of Assyria. Probably written between 626–612 BCE (the date of the destruction of Nineveh, the Assyrian capital), the book celebrates in oracles, hymns, and laments the fact that Yahweh has saved Judah from potential devastation by the Assyrians.

He begins with the words "The Lord is a jealous God and avenging . . . is slow to anger and of great might, and the Lord will by no means clear the guilty" (chapter 1, verses 2–3). From that beginning he predicts the overthrow of Assyria and the devastating manner in which Nineveh will be destroyed.

Habakkuk. The Book of Habakkuk, the eighth book of the Twelve (Minor) Prophets, was written by a prophet difficult to identify. He may have been a professional prophet of the Temple from the 7th century BCE (probably between 605–597 BCE). Containing three chapters, Habakkuk combines lamentation and oracle. In the first chapter, he cries out for Yahweh to help his people: "O Lord, how long shall I cry for help, and thou wilt not hear?" (chapter 1, verse 2). Though Yahweh will send mighty nations (*e.g.*, the neo-Babylonians will be the executors of his judgment), Habakkuk wonders who will then stop these instruments of God's justice, who use great force. The answer comes in a brief, almost cryptic verse, "but the righteous shall live by his faith." The rest of chapter 2 pronounces a series of woes against those who commit social injustices and engage in debauchery. The last chapter is a hymn anticipating the deliverance to be wrought by Yahweh.

"The righteous shall live by faith"

Zephaniah. The Book of Zephaniah, the ninth book of the Twelve (Minor) Prophets, is written in three chapters. Composed by the prophet Zephaniah in the latter part of the 7th century BCE, the book is an attack against corruption of worship in Judah, probably before the great Deuteronomic reform took place. He attacked the religious syncretism that had become established, especially the worship of Baal and astral deities, and predicted the coming catastrophe of the "Day of the Lord." He denounced both foreign nations and Judah, but issued a promise of the restoration of Israel: "Sing aloud, O daughter of Zion; shout, O Israel! Rejoice and exult with all your heart, O daughter of Jerusalem" (chapter 3, verse 14). The reason for exultation is that Yahweh will deliver his people.

Haggai. The Book of Haggai, the 10th book of the Twelve (Minor) Prophets, is a brief work of only two chapters. Written about 520 BCE by the prophet Haggai, the book contains four oracles. The first oracle calls for Zerubbabel, the governor of Judaea, and Joshua, the high priest, to rebuild the Temple (chapter 1, verses 1–11). A drought and poor harvests, according to Haggai, had been caused because the returnees from the Exile had neglected or failed to rebuild the Temple. The second oracle, addressed to the political and religious leaders and the people, sought to encourage them in their rebuilding efforts (chapter 2, verses 1–9). Apparently they were disappointed that the new Temple was not as splendid as the former one, so Haggai reassured them: "My Spirit abides among you, fear not." The third oracle was issued against the people for not acting in a holy manner (chapter 2, verses 10–19), and the fourth proclaimed that Zerubbabel would be established as the Davidic ruler (chapter 2, verses 20–23). His promise, however, remained unfulfilled.

Zechariah. The Book of Zechariah, the 11th book of the Twelve (Minor) Prophets, dates from the same period as that of Haggai—about 520 BCE. Though the book contains 14 chapters, only the first eight are oracles of the prophet; the remaining six probably came from a school of his disciples and contain various elaborations of Zechariah's eschatological themes.

The visions of Zechariah and messianic overtones

Though little is known about Zechariah's life, he probably was one of the exiles who returned to Jerusalem from Babylon. After an initial call to repentance (chapter 1, verses 1–6), Zechariah had a series of eight visions (chapter 1, verse 7 to chapter 6, verse 15). The first is of four horsemen who have patrolled the Earth to make sure that it is at rest. The second vision is of four horns (*i.e.*, nations that have conquered Israel and Judah), which will be destroyed. The third vision is of a man with a measuring line, but Jerusalem will be beyond measurement. The fourth vision shows Joshua the high priest in the heavenly court being prosecuted by Satan (the celestial adversary) and the high priest's eventual acquittal and return to his high position. The fifth vision is of a golden lampstand and an olive tree to emphasize the important positions of Joshua and Zerubbabel, which these two figures symbolize. The sixth and seventh visions—of a flying scroll and a woman of wickedness—symbolize the removal of Judah's previous sins. The eighth vision of four chariots probably refers to the anticipated messianic reign of Zerubbabel, a hope that was thwarted. Chapters 7 and 8 concern fasting and the restoration of Jerusalem.

The remaining chapters—9–14—are additions that contain messianic overtones. Chapter 9, verses 9–10, with its reference to a king riding on the foal of an ass and to a vast kingdom of peace, was used by New Testament Gospel writers in reference to Jesus' entrance into

Jerusalem prior to his crucifixion. The book closes on the note of the suffering Good Shepherd, the final battle between Jerusalem and the nations and eventual victory under God, and the universal reign of Yahweh, "king over all the earth."

Malachi. The Book of Malachi, the last of the Twelve (Minor) Prophets, was written by an anonymous writer called Malachi, or "my messenger." Perhaps written from about 500–450 BCE, the book is concerned with spiritual degradation, religious perversions, social injustices, and unfaithfulness to the Covenant. Priests are condemned for failing to instruct the people on their Covenant responsibilities, idolatry is attacked, and men are castigated for deliberately forgetting their marriage vows when their wives become older. In chapter 3, the message is that Yahweh will send a messenger of the Covenant to prepare for, and announce, the day of judgment. If the people turn from their evil ways, God will bless them, and those who "feared the Lord" will be spared. The book ends with a call to remember the Covenant and with a promise to send Elijah, the 9th-century prophet who ascended into heaven in a whirlwind on a chariot, "before the great and terrible day of the Lord comes." (L.F.)

THE KETUVIM

Contents of the Ketuvim

The Ketuvim (the Writings or the Hagiographa), the third division of the Hebrew Bible, comprises a miscellaneous collection of sacred writings that were not classified in either the Torah or the Prophets. The collection is not a unified whole: it includes liturgical poetry (Psalms and Lamentations of Jeremiah), secular love poetry (Song of Solomon), wisdom literature (Proverbs, Book of Job, and Ecclesiastes), historical works (I and II Chronicles, Book of Ezra, and Book of Nehemiah), apocalyptic, or vision, literature (Book of Daniel), a short story (Book of Ruth), and a romantic tale (Book of Esther); it ranges in content from the most entirely profane book in the Bible (Song of Solomon) to perhaps the most deeply theological (Job); it varies in mood from a pessimistic view of life (Job and Ecclesiastes) to an optimistic view (Proverbs). Psalms, Proverbs, and Job constitute the principal poetic literature of the Hebrew Bible and, in many respects, represent the high point of the Hebrew Bible as literature; in fact, Job must be considered one of the great literary products of man's creative spirit.

Although portions of some of the books of the Ketuvim (*e.g.*, Psalms and Proverbs) were composed before the Babylonian Exile (586–538 BCE), the final form of all the books was postexilic, and Daniel was not written until almost the middle of the 2nd century BCE. The books were not included in the prophetic collection for several reasons: because they did not fit the content or the historical–philosophical framework of that collection, because they were originally seen as purely human and not divine writings, or simply because they were written too late for inclusion. Although some of the books individually were accepted as canonical quite early, the collection of the Ketuvim as a whole, as well as some individual books within it, was not accepted as completed and canonical until well into the 2nd century CE. As noted above, there are several indications that the lapse of time between the canonization of the Prophets and of the Ketuvim was considerable; *e.g.*, the practice of entitling the entire Scriptures "the Torah and the Prophets" and the absence of a fixed name.

The needs of the Hellenistic Jews in Alexandria and elsewhere in the Greek-speaking Diaspora led to the translation of the Bible into Greek. The process of translation began with the Torah about the middle of the 3rd century BCE and continued for several centuries. In the Greek canon, as it finally emerged, the Ketuvim was eliminated as a corpus, and the books were redistributed, together with those of the prophetic collection, according to categories of literature, giving rise to a canon with four divisions: Torah, historical writings, poetic and didactic writings, and prophetic writings. Also, the order of the books was changed, and books not included in the Hebrew Bible were added. The early Christians of both the Eastern and Western churches generally cited and accepted as canonical the Scriptures according to the Greek version. When the Protestants produced translations based upon the Hebrew original text and either excluded or separated out (as Apocrypha) the books not found in the Hebrew Bible, they retained the order and the divisions of the Greek Bible. Thus the Ketuvim is not to be found as a distinct collection in the Christian Old Testament.

Order of the books of the Ketuvim

An ancient tradition, preserved in the Babylonian Talmud, prescribed the following order for the Ketuvim: Ruth, Psalms, Job, Proverbs, Ecclesiastes, Song of Solomon, Lamentations, Daniel, Esther, Ezra (which included Nehemiah), and I and II Chronicles. This sequence was chronological according to rabbinic notions of the authorship of the books. Ruth relates to the age of the judges and concludes with a genealogy of David; the Psalms were attributed, for the most part, to David; Job was assigned to the time of the Queen of Sheba, although the rabbis differed among themselves about the date of the hero; Proverbs, Ecclesiastes, and Song of Solomon were all attributed to Solomon; Lamentations, which was ascribed to Jeremiah, refers to the destruction of Jerusalem and the beginning of the Babylonian Exile; the heroes of Daniel were active until early in the reign of Cyrus II, the king of Persia who ended the exile; Esther pertains to the reign of Xerxes I, later than that of Cyrus but earlier than that of Artaxerxes I, the patron of Ezra, reputed also to have written I and II Chronicles.

Despite this tradition, however, it would appear that the sequence of the Ketuvim was not completely fixed, and there is a great variety in ordering found in manuscripts and early printed editions. The three larger books—Psalms, Job, and Proverbs—have always constituted a group, with Psalms first and the other two interchanging. The order of the five Megillot, or Scrolls (Song of Solomon, Ruth, Lamentations, Ecclesiastes, and Esther), has shown the greatest variations. The order that has crystallized has a liturgical origin; the books are read on certain festival days in Jewish places of worship and are printed in the calendar order of those occasions. Chronicles always appears at either the beginning or the end of the corpus. Its final position is remarkable because the narrative of Ezra and Nehemiah follows that of Chronicles. The final position may have resulted from an attempt to place the books of the Hebrew Bible in a framework (Genesis and Chronicles both begin with the origin and development of the human race, and both conclude with the theme of the return to the land of Israel), but it was more probably the result of the late acceptance of Chronicles into the canon.

Psalms. The Psalms (from Greek *psalmas*, "song") are poems and hymns, dating from various periods in the history of Israel, that were assembled for use at public worship and that have continued to play a central role in the liturgy and prayer life of both Jews and Christians. Known in Hebrew as Tehillim (Songs of Praise), the Psalter (the traditional English term for the Psalms, from the Greek *psalterion*, a stringed instrument used to accompany these songs) consists of 150 poems representing expressions of faith from many generations and diverse kinds of men that formed a total worshipping community. These unsystematic poems epitomize the theology of the entire Hebrew Bible.

Characteristics of Hebrew poetry

Hebrew poetry has much in common with the poetry of most of the ancient Near East, particularly the Canaanite poetic literature discovered at Ras Shamra. Its main features are rhythm and parallelism. The rhythm, which is difficult to determine precisely because the proper pronunciation of ancient Hebrew is unknown, is based upon a system of stressed syllables that follows the thought structure of the poetic line. The line, or stich, is the basic verse unit, and each line of verse is normally a complete thought unit. The most common Hebrew line consists of two parts with three stresses to each part (3/3); thus:

> Have-mercy-on-me,/O-God,/in-your-goodness;
> in-your-great-tenderness/wipe-away/my-faults.
> (Ps. 51:1)

Lines with three or four parts and parts with two, four, or five stresses also occur.

The lines present various kinds of parallelism of members, whereby the idea expressed in one part of a line is balanced by the idea in the other parts. The classical study on Hebrew parallelism was done by Robert Lowth, an 18th-century Anglican bishop, who distinguished three types: synonymous, antithetic, and synthetic. Synonymous parallelism involves the repetition in the second part of what has already been expressed in the first, while simply varying the words.

> Yahweh, do not punish me in your rage,
> or reprove me in the heat of anger.
>
> (Ps. 38:1)

In antithetic parallelism the second part presents the same idea as the first by way of contrast or negation.

> For Yahweh takes care of the way the virtuous go,
> but the way of the wicked is doomed.
>
> (Ps. 1:6)

Synthetic parallelism involves the completion or expansion of the idea of the first part in the second part.

> As a doe longs for running streams,
> so longs my soul for you, my God.
>
> (Ps. 42:1)

Synthetic parallelism is a broad category that allows for many variations, one of which has the picturesque name "staircase" parallelism and consists of a series of parts or lines that build up to a conclusion.

> Pay tribute to Yahweh, you sons of God,
> tribute to Yahweh of glory and power,
> tribute to Yahweh of the glory of his name,
> worship Yahweh in his sacred court.
>
> (Ps. 29:1–2)

Although it is evident that Hebrew poetry groups lines into larger units, the extent of this grouping and the principles on which it is based are uncertain. The acrostic poems are a notable exception to this general uncertainty.

The numeration of the Psalms found in the Hebrew Bible and those versions derived from it differs from that in the Septuagint, the Vulgate, and the versions derived from them. The latter two join Psalms 9 and 10 and 114 and 115 but divide both 116 and 147 into two. The following scheme shows the differences:

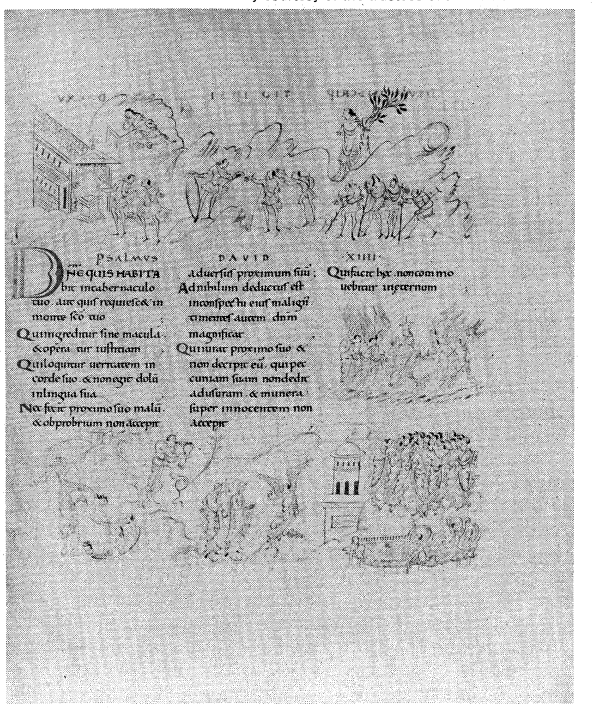
Hebrew	Septuagint–Vulgate
1–8	1–8
9–10	9
11–113	10–112
114–115	113
116	114–115
117–146	116–145
147	146–147
148–150	148–150

Although Roman Catholic versions in the past have used the Septuagint–Vulgate way of numbering, recent translations have followed the Hebrew tradition.

The present form of the Psalter is the result of a lengthy literary history. It is divided into five books (Psalms 1–41; 42–72; 73–89; 90–106; and 107–150), probably in imitation of the five books of the Pentateuch. Psalm 1 serves as an introduction to the whole Psalter, while Psalm 150 is a final doxology (an expression of praise to God); the books are divided from each other by short doxologies that form the conclusions of the last psalm of each of the first four books. This division, however, appears to be artificial. There are indications, cutting across the present divisions, that the book was a compilation of existing collections. That there were several collections existing side by side is seen in the way that certain psalms (*e.g.*, Psalms 14 and 53) duplicate each other almost word for word. At some phase of the Psalter's development there must have been an Elohistic collection (Psalms 42–83) distinguished by the use of the divine name Elohim in place of Yahweh, which is far more common in the rest of the psalms. There appear to be two distinct collections of psalms ascribed to David, one Yahwistic (Psalms 3–41) and the other Elohistic (Psalms 51–72). Further evidence of the book's gradual growth may be seen in the editorial gloss following Psalm 72; it purports to conclude the "prayers of David," although there are more Davidic psalms.

Compilation of existing collections

By courtesy of the trustees of the British Museum



Illustrated text of Psalm 15 (Vulgate Psalm 14) from the Utrecht Psalter of Reims, 9th–12th century. In the British Museum (Harley, MS.603).

The superscriptions found at the head of most of the psalms are obscure but point to the existence of even earlier collections. Psalms are attributed to David, Asaph, and the sons of Korah, among others. It is generally held that Asaph and the sons of Korah indicate collections belonging to guilds of temple singers. Other possible collections include the Songs of Ascents, which were probably pilgrim songs in origin, the Hallelujah Psalms, and a group of 55 psalms with a title normally taken to mean "the choirmaster."

It is evident that the process whereby these various collections were formed and then combined was extremely complex. The investigation of the process is made difficult because individual psalms and whole collections underwent constant development and adaptation. Thus, for example, private prayers became liturgical, songs of local sanctuaries were adapted to use in the Temple, and psalms that became anachronistic by reason of the fall of the monarchy or the destruction of the Temple were reworked to fit a contemporary situation. Such problems complicate the determination of the date and original occasion of the psalm.

Date and authorship

For centuries both Jews and Christians ascribed the whole Psalter to David, just as they ascribed the Pentateuch to Moses and much of the wisdom literature to Solomon. This was thought to be supported by the tradition that David was a musician, a poet, and an organizer of the liturgical cult and also by the attribution of 73 psalms to David in the superscriptions found in the Hebrew Bible. These superscriptions, however, need not refer to authorship. Moreover, it is clear that David could not have written all the psalms attributed to him because some of them presuppose the existence of the Temple in Jerusalem, which was not constructed until later. Contrary to the long-established Davidic authorship tradition, at the end of the 19th century most biblical critics spoke of a Persian date (539–333 BCE) and even of the Maccabean era (mid-2nd century BCE) for the majority of the psalms. In the 20th century the Psalter has been considered to be a collection of poems that reflect all

periods of Israel's history from before the monarchy to the postexilic restoration, and it is thought that David played a central role in the formation of the religious poetry of the Jewish people. Scholars, however, are reluctant to assign precise dates to given psalms.

The most important contribution to modern scholarship on the Psalter has been the work of Hermann Gunkel, a German biblical scholar, who applied form criticism to the psalms. Form criticism is the English name for the study of the literature of the Bible that seeks to separate its literary units and classify them into types or categories (*Gattungen*) according to form and content, to trace the history of these types, and to reconstruct the particular situation in life or setting (*Sitz im Leben*) that gave rise to the various types. This approach does not ignore the personal role of individual composers and their dates, but it recognizes that Hebrew religion, conservative in faith and practice, was more concerned with the typical than with the individual and that it expressed this concern in formal, conventional categories. The study of these types is aided by viewing them in the context of similar literary compositions in the earlier or contemporary cultures of the ancient Near East.

Major types of psalms

Gunkel identified five major types of psalms, each cultic in origin. The first type is the Hymn, which is a song of praise, consisting of an invitation to praise Yahweh, an enumeration of the reasons for praise (*e.g.*, his work of creation, his steadfast love), and a conclusion in which frequently the invitation is repeated. The life setting of the hymns was generally some occasion of common worship. Two subgroups within the hymn type are the Songs of Zion, which glorify Yahweh's presence in the city of Jerusalem, and the Enthronement Songs, which—though their number, setting, and interpretation have been the subject of much debate—acclaim Yahweh's kingship over the whole world.

The second type is the Communal Lament. Its setting was some situation of national calamity, when a period of prayer, fasting, and penitence would be observed. In such psalms Yahweh is invoked, the crisis is described, Yahweh's help is sought, and confidence that the prayer has been heard is expressed.

The Royal Psalms are grouped on the basis not of literary characteristics but of content. They all have as their life setting some event in the life of the pre-exilic Israelite kings; *e.g.*, accession to the throne, marriage, departure for battle. Gunkel pointed out that in ancient Israel the king was thought to have a special relationship to Yahweh and thus played an important role in Israelite worship. With the fall of the monarchy, these psalms were adapted to different cultic purposes.

In the Individual Lament an individual worshipper cries out to Yahweh in time of need. The structure of these psalms includes: an invocation of Yahweh, the complaint, the request for help, an expression of certainty that Yahweh will hear and answer the prayer, and in many cases a vow to offer a thanksgiving sacrifice in praise of the Lord. Three aspects of the individual laments have been the subject of extensive study: the identity of the "enemies" who are often the reason for the complaint; the meaning of the term poor, which is frequently used to describe the worshipper; and the sudden transition in mood to certainty that the prayer has been heard. Psalms of this type form the largest group in the Psalter.

The final major type is the Individual Song of Thanksgiving, which presumably had its setting in the thanksgiving sacrifice offered after a saving experience. These psalms begin and conclude with an exclamation of praise to Yahweh. The body of the psalm contains two elements: the story of the one who has been saved and the recognition that Yahweh was the rescuer.

Gunkel also distinguished several minor types of psalms, including Wisdom Poems, Liturgies, Songs of Pilgrimage, and Communal Songs of Thanksgiving.

For Gunkel, although the types of the psalms were originally cultic, the majority of the poems in the existing Psalter were composed privately in imitation of the cultic poems and were intended for a more personal, "spiritualized" worship. Most biblical scholars since Gunkel have accepted his classifications, with perhaps some modifications, but have focussed increased attention on the setting, the *Sitz im Leben*, in which the psalms were sung. Sigmund Mowinckel, a Norwegian scholar, explained the psalms as wholly cultic both in origin and in intention. He attempted to relate more than 40 psalms to a hypothetical autumnal New Year festival at which the enthronement of Yahweh as the universal king was commemorated; the festival was associated with a similar Babylonian celebration. Artur Weiser, a German scholar, sought the cultic milieu of the Hebrew psalms especially in an annual feast of covenant renewal, which was uniquely Israelite.

Psalms is a source book for the beliefs contained in the entire Hebrew Bible. Yet, doctrines are not expounded, for this is a book of the songs of Israel that describe the way Yahweh was experienced and worshipped. Yahweh is creator and saviour; Israel is his elected people to whom he remains faithful. The enemies of this people are the enemies of Yahweh. In these songs are found the entire range of basic human feelings and attitudes before God—praise, fear, trust, thanksgiving, faith, lament, joy. The book of Psalms has thus endured as the basic prayerbook for Jews and Christians alike.

Proverbs. Proverbs is probably the oldest extant document of the Hebrew wisdom movement, of which King Solomon was the founder and patron. Wisdom literature flourished throughout the ancient Near East, with Egyptian examples dating back to before the middle of the 3rd millennium BCE. It revolved around the professional sages, or wise men, and scribes who were in the service of the court, and it consisted primarily in maxims about the practical, intelligent way one should conduct his life and to some extent in speculations about the very worth and meaning of human life. The most common form of these wise sayings, which were intended for oral instruction especially in the schools run by the sages for the young men at the court, was the *mashal* (Hebrew: "comparison" or "parable," although frequently translated "proverb"). Typically a pithy, easily memorized aphoristic saying based on experience and universal in application, the *mashal* in its simplest and oldest form was a couplet in which a definition was given in two parallel lines related to each other either antithetically or synthetically. Verse 5 of the 15th chapter of Proverbs is an example of a simple antithetic saying:

Wisdom literature in the ancient Near East

> He who spurns his father's discipline is a fool,
> he who accepts correction is discreet.

Other forms of the *mashal*, such as parables, riddles, allegories, and ultimately full-scale compositions developed later. The word *mashal* was derived from a root that meant "to rule," and thus a proverb was conceived as an authoritative word.

The two principal types of wisdom—one practical and utilitarian, the other speculative and frequently pessimistic—arose both within and outside Israel. Practical wisdom consisted chiefly of wise sayings that appealed to experience and offered prudential guidelines for a successful and happy life. Such wisdom is found in a collection of sayings bearing the name of Ptahhotep, a vizier to the Egyptian pharaoh about 2450 BCE, in which the sage counsels his son that the path to material success is by way of proper etiquette, strict discipline, and hard work. Although such instructions were largely materialistic and political, they were moral in character and contributed to a well-ordered society.

Speculative wisdom went beyond maxims of conduct and reflected upon the deeper problems of the value of life and of good and evil. Examples are found in ancient Egyptian and Mesopotamian texts—particularly *Ludlul bel nemeqi*, often called the "Babylonian Job"—in which sensitive poets pessimistically addressed such questions as the success of the wicked, the suffering of the innocent, and, in short, the justice of human life.

Hebrew wisdom, which owed much to that of its neighbours, appeared with the establishment of the monarchy and a royal court and found a patron in Solomon. Through the following centuries the wise men were at

times the object of rebuke by the prophets, who disliked their pragmatic realism. The exile, however, brought a change in Hebrew wisdom; it became deeply religious. The wise men were convinced that religion alone possessed the key to life's highest values. It was this mood that dominated the final shaping of the Hebrew wisdom literature. Though dependent on older materials and incorporating documents from before the exile, the wisdom books in their present form were produced after the exile. In the Hebrew Bible the book of Proverbs offers the best example of practical wisdom, while Job and Ecclesiastes give expression to speculative wisdom. Some of the psalms and a few other brief passages are also representative of this type of literature. Among the Apocrypha, the Wisdom of Solomon and Ecclesiasticus are wisdom books.

Collections in the book of Proverbs

The book of Proverbs is a collection of units originally independent, some of which can be traced back to the era of Solomon. The present form of the book was the result of a long process of growth that was not completed until postexilic times. It consists of two principal collections of early origin called "the proverbs of Solomon" and "proverbs of Solomon which the men of Hezekiah king of Judah copied." Appendixes were added to each of the collections. The whole book was preceded by a long introduction and concludes with a poem praising the ideal wife. In addition to sectional titles, changes in literary form and in subject matter help to mark off the limits of the various units, which can be ordered into nine sections.

The introduction (chapters 1–9) constitutes the youngest unit in the book. It consists of a series of poems or discourses in which a father exhorts his son to acquire wisdom and in which wisdom personified intervenes. These chapters have a more speculative quality than the remainder of the book. They do not treat wisdom simply as a human quality and achievement or as a cultural legacy imparted by teachers and parents; they present it as a universal and abiding reality, transcending the human scene. Wisdom is the first of God's works and participated with him in the creation of the world. A constantly debated aspect of this section concerns the identity of "the loose [strange] woman" who is set over against Wisdom.

The "proverbs of Solomon" (10:1–22:16) consist entirely of parallelistic couplets—the *mashal* in its primitive form. There are 375 aphorisms each complete in itself and arranged in no apparent order. The motivation of this section, in contrast to the preceding, is strongly practical: wisdom is a human achievement by means of which man's life can be fulfilled. The wise are contrasted with fools, and the just with the wicked. It is difficult, however, to establish the nature of the difference, if any, between the wicked and the fool or between the just and the wise.

The "sayings of the wise" (22:17–24:22) consist of longer units or sayings introduced by a preface. The most distinctive feature of this section is its close relationship to a piece of Egyptian writing, "The Instruction of Amenemope," which has been dated within the broad limits of 1000–600 BCE. The Hebrew author apparently used this work as a model—the Egyptian work comprises 30 chapters, and the Hebrew text refers to its "thirty sayings"—and as one of the sources in compiling his own anthology. An additional collection of four wise sayings (24:23–34) forms a supplement to the "sayings of the wise."

The second collection of "proverbs of Solomon" (chapters 25–29) consists of 128 sayings that closely resemble the earlier collection in both spirit and form, although quatrains as well as couplets are included. The scribes of Hezekiah's court (*c.* 700 BCE) are credited with assembling this collection.

The book concludes with four independent units or collections. The "words of Agur" (30:1–14) differs sharply in spirit and substance from the rest of Proverbs; it has much closer affinities with the book of Job, stressing the inaccessibility of wisdom for man. There is no internal evidence, such as a continuous theme, to show that these 14 verses are a single unit; but in the Septuagint they stand together between the "sayings of the wise" and its supplement. The "numerical sayings" (30:15–33) contain elements of riddle and show a special interest in the wonders of nature and the habits of animals. The "instruction of Lemuel" (31:1–9) is an example of the importance of maternal advice to a ruler in the ancient Near East. Lemuel seems to have been a tribal chieftain of northwest Arabia, in the region of Edom. The final section (31:10–31) is an alphabetical poem in praise of the "perfect wife," who is celebrated for her domestic virtues.

Distinctive character of wisdom literature within Hebrew Scriptures

The wisdom movement constituted a special aspect of the religious and cultural development of ancient Israel. As the primary document of the movement, Proverbs bears a clear impress of this distinctive character, so that in many respects it presents a sharp contrast to the outlook and emphases of Israel's faith as attested in the Hebrew Scriptures generally. This contrast also marks Job and Ecclesiastes, however greatly they may differ from Proverbs in other respects.

Proverbs never refers to Israel's history. In the Hebrew Bible as a whole, this history is constantly recalled not so much for social or political reasons as to declare the faith of Israel that God has acted in its history to redeem his people and make known to them the character of his rule. The great themes of the promise to the patriarchs, the deliverance from slavery, the making of the Covenant at Mt. Sinai, the wilderness wandering, and the inheritance of Canaan were celebrated in Israel's worship to tell the story of God's revelation of himself and of his choice of Israel. None of this is alluded to in Proverbs. The implication seems to be that for Proverbs God's revelation of himself is given in the universal laws and patterns characteristic of nature, especially human nature, rather than in a special series of historical events; that is, the revelation of God is in the order of creation rather than in the order of redemption. Moreover, the meaning of this revelation is not immediately self-evident but must be discovered by men. This discovery is an educational discipline that trusts human reason and employs research, classifying and interpreting the results and bequeathing them as a legacy to future generations. The wise are those who systematically dedicate themselves to this discovery of the "way" of God.

Unlike Job and Ecclesiastes, Proverbs (with the exception of the "words of Agur") is optimistic in the sense that it assumes that wisdom is attainable by those who seek and follow it; that is, man can discover enough about the character of God and his law to ensure the fulfillment of his personal life. This character of God is conceived almost entirely in terms of ethical laws, and the rewards for their observance are defined in terms of human values; *e.g.*, health, long life, respect, possessions, security, and self-control.

Because God is apprehended in static terms, rather than dynamic as elsewhere in the Bible, the viewpoint of Proverbs is anthropocentric. Man's destiny depends upon his responsible action. There is no appeal to divine mercy, intervention, or forgiveness; and the divine judgment is simply the inexorable operation of the orders of life as God has established them. Implicit in the book is an aristocratic bias. The wise constitute an elite nurtured by inheritance, training, and self-discipline; fools are those who can never catch up, because of either the determinism of birth or the wasted years of neglect. In its social and cultural attitudes, the book is probably the most conservative in the Bible: wealth and status are most important; obedience to the king and all authorities is inculcated; industry and diligence are fostered, for hunger, poverty, and slavery are the fate of the lazy; and age and accepted conventions are accorded great respect.

Job. The Book of Job is not only the finest expression of the Hebrew poetic genius; it must also be accorded a place among the greatest masterpieces of world literature. The work is grouped with Proverbs and Ecclesiastes as a product of the wisdom movement, even though it contains what might be called an anti-wisdom strain in that the hero protests vehemently against the rationalistic

ethics of the sages. Yet it is the supreme example among ancient texts of speculative wisdom in which a man attempts to understand and respond to the human situation in which he exists.

The Book of Job consists of two separate portions. The bulk of the work is an extended dialogue between the hero and his friends and eventually Yahweh himself in poetic form. The poem is set within the framework of a short narrative in prose form. The book falls into five sections: a prologue (chapters 1 and 2); the dialogue between Job and his friends (3–31); the speeches of Elihu (32–37); the speeches of Yahweh and Job's reply (38–42:6); and an epilogue (42:7–17).

Prose narrative in Job

The prologue and epilogue are the prose narrative. This is probably an old folktale recounting the story of Job, an Edomite of such outstanding piety that he is mentioned by the prophet Ezekiel in conjunction with Noah and Daniel. The name Job was common in antiquity, being found in texts ranging from the 19th to the 14th century BCE. Whether the folktale is preserved in its original oral form or whether it has been retold by the poet of the dialogue is not known. The fact that an Edomite sheikh is commended by the Hebrew God, however, suggests a date before the 6th century BCE, for Jewish distrust of Edomites became intense during the exile, and the archaic language makes a date in the 8th century probable.

Job is pictured as an ideal patriarch who has been rewarded for his piety with material prosperity and happiness. The Satan (Accuser), a member of the heavenly council of Yahweh, acts with Yahweh's permission as an *agent provocateur* to test whether or not Job's piety is rooted in self-interest. Faced with the appalling loss of his worldly possessions, his children, and finally his own health, Job refuses to curse Yahweh. His capacity for trusting Yahweh's goodness has made him an unsurpassed model of patience. Three of Job's friends, whose names identify them also as Edomites, now arrive to comfort him. At this point the poetic dialogue begins. The conclusion of the tale, as given in the epilogue, describes the restoration of Job, who receives double his original possessions and lives to a ripe old age.

The dialogue between Job and his friends

The picture of Job that is presented in the poetic portion is radically different. Instead of the patient and loyal servant of Yahweh, he is an anguished and indignant sufferer, who violently protests the way Yahweh is treating him and displays a variety of moods ranging from utter despair, in which he cries out accusingly against Yahweh, to bold confidence, in which he calls for a hearing before Yahweh. Most scholars have dated this section to the 4th century BCE, but there is a growing tendency to regard it as two centuries earlier, during the period of the exile. This precise dating is based on the fact that the dialogue shows clear literary dependence on Jeremiah, whereas equally obvious connections with Deutero-Isaiah suggest the dependence of the latter on Job.

The poem opens with a heartrending soliloquy by Job in which the sufferer curses the day of his birth. The shocked friends are roused from their silence, and there follow three cycles of speeches (chapters 4–14, 15–21, and 22–27) in which the friends speak in turn. To each such speech Job makes a reply. The personalities of the friends are skillfully delineated, Eliphaz appearing as a mystic in the prophetic tradition, Bildad as a sage who looks to the authority of tradition, and Zophar as an impatient dogmatist who glibly expounds what he regards as the incomprehensible ways of God.

Eliphaz begins the first cycle by recounting a mystical vision that revealed to him the transcendence of God and the fact that all men are by nature morally frail. He suggests that suffering may be disciplinary, although this is irrelevant to Job's plight. Finally, he urges contrite submission to Yahweh. Job chides his friends for failing him in his hour of need and charges God with being his tormentor.

Bildad suggests that the fault may have lain in Job's children and reiterates Eliphaz' call to humble submission. Job then retorts that the doctrine of Yahweh's omnipotence is no answer but a serious problem, because Yahweh appears to be merely omnipotent caprice. He is convinced that if he could only meet Yahweh in open debate he would be vindicated, but he recognizes the need for an impartial third party who could intervene and protect him from Yahweh's overpowering might.

Zophar re-echoes his predecessors' views on Yahweh but goes the full length of accusing Job himself of sin and once more urges Job to a contrition that for him could only be hypocritical. Job continues to insist that Yahweh is capricious and defiantly challenges him but is bewildered when no reply is forthcoming. His longing for death as a welcome release leads him to ask whether man might not hope for a revival after death, but this daring hope is immediately rejected.

The second cycle opens with Eliphaz accusing Job of blasphemy and almost exultantly describing the fate of the wicked. In his reply Job returns to the idea of a third party to the debate. Now, however, this umpire or judge has become an advocate, a counsel for the defense. After Bildad has again elaborated on the fate of the wicked, Job states that a Vindicator, or Redeemer (Go'el), will establish his innocence. The Vindicator of this crucial but sadly corrupted passage (19:25–27) has long been identified with God himself, so that according to some scholars Job "appeals away from the God of orthodox theology to God as He must be." A few scholars, however, recognize the Vindicator as the third party (the "umpire" or "witness") of earlier chapters. It is also unclear whether this vindication will take place before or after Job's death. Then Zophar, though freely admitting that the wicked may indeed enjoy some prosperity, describes how they fall victim to inevitable nemesis. Job maintains that the wicked do not end thus but live on to an old age.

Eliphaz begins the third cycle by accusing Job at last of specific sins and again counsels Job to humble himself before Yahweh. But Job cannot find this God, who seems to be completely indifferent to him. The conclusion of the dialogue is in serious disorder, with speeches placed in Job's mouth that could only have been uttered by the friends. The final speech of Zophar, which is omitted, seems to be represented by a fragment preserved within the third reply of Job.

Chapter 28 is regarded as a later addition by most scholars, because it is hardly in place at this juncture in the dialogue, especially in the mouth of Job. It is a magnificent hymn in praise of wisdom. Chapters 29–31 contain a monologue by Job; in them occurs an adumbration of the highest moral ideal to be found in the Hebrew Bible.

The speeches of Elihu and the speeches of Yahweh

Although a few scholars have maintained that the speeches of Elihu formed part of the original work, most reject this section as a later insertion. The speeches merely reiterate the dogmas of the friends and unduly delay the appearance of Yahweh. Although the section is in poetic form, its style is different from that of the dialogue. Significantly, there is no mention of Elihu in the dialogue or anywhere else in the book, yet the Elihu speeches are familiar with the dialogue, frequently quoting verbatim from it. Chapter 32 is of interest, because it appears to contain the writer's notes and comments on the dialogue, often citing passages from it. Worthy of notice is the writer's emphasis on the disciplinary value of suffering.

The climax of the poem is reached in the speeches of Yahweh, who appears in a majestic theophany—a whirlwind—and reveals himself to Job in three speeches interspersed with two short speeches by Job. Biblical scholars have often questioned whether this section—especially the descriptions of Behemoth (the hippopotamus) and Leviathan (the crocodile) in the second Yahweh speech—is a genuine part of the original poem, but there is no doubt that their presence at this point in the book is a dramatic triumph. Throughout these speeches Yahweh does not offer rational answers to Job's questions and accusations; he raises the discussion to a new perspective. With heavy irony Yahweh puts to Job a series of unanswerable questions about the mysteries of the universe; if, the writer is asking, Job is unable to answer the sim-

ple questions about the divine activity in the marvels of nature, how can Yahweh explain to him the deeper mystery of his dealings with men. Job's personal problem is ignored, yet he finds his answer in this direct encounter with Yahweh:

> I had heard of thee by the hearing of the ear,
> but now my eye sees thee;
> therefore I despise myself,
> and repent in dust and ashes.

Job stands in a new relationship to Yahweh, one no longer based on hearsay but the result of an act of personal faith expressed in repentance.

A few scholars, beginning in the mid-18th century, have attempted to demonstrate the influence of Greek tragedy upon the form of the book. This has not met with acceptance by most critics; its long monologues are not truly dramatic in nature. Neither is it a philosophical discussion in the style of the Platonic dialogues. It is a deeply religious poem with dramatic possibilities. It skillfully blends many genres: folktale, hymn, individual lament, prophetic oracle, and didactic poem.

The author and his basic message

The author remains quite unknown except for a few hints provided by the book itself. That he was a Jew is assumed because of his familiarity with much of the Hebrew literature. Nevertheless, the book does not have a Hebrew setting, it is pervaded with foreign elements, and it shows a special knowledge of Egypt, thus leading many to believe that he was well travelled or lived outside the Holy Land. He certainly was a keen observer of the natural world, and his sensitive feeling for the agony of the sufferer is a compelling argument that he himself had known anguish.

The book touches on many subjects, such as disinterested obedience to God under testing, innocent suffering, social oppression, religious experience and pious suffering, a man's relation to God, and the nature of that God. Scholars have attempted to discover the basic theme or message of the author. Because of the greater difficulty in understanding the Job of the poetic portion, the traditional interpretation looked to the narrative and saw the message of the book as the need for patient bearing and faith in God despite tribulation. When certain poetic passages were thought to point to a belief in the resurrection of the body, Job became not only a patient sufferer but also a prophet of the resurrection. This view, however, does not account for the Job of the poetic portion. Thus, in the 19th century, with the advancement of biblical criticism, scholars began to claim that the author was dealing with the problem of unmerited suffering. The book presents a deep view of the problem of suffering, and Job's experience teaches that man must rest in faith and resign himself to the incomprehensible ways of God.

It would seem, however, that the question raised by Job is both deeper and broader than the question of how to account for the infliction of physical adversity on the innocent. Job's physical suffering is the outward symbol of his intense inward agony, the agony of a man who feels himself lost in a meaningless universe and abandoned even by God. What torments Job—and the author—is the question of the justice of God and the justice and honour of man before God. His passionate pleading of his own righteousness and his calling upon God for a hearing lead him to an encounter with God. This encounter does not answer the question of why the innocent suffer, but it is the only answer to the plea of a man seeking to find his God and to justify himself to him. The complacent believer who has been shattered by suffering, doubt, and despair is confirmed in faith and repents.

The Megillot (the Scrolls). The five books known as the Megillot or Scrolls are grouped together as a unit in modern Hebrew Bibles according to the order of the annual religious festivals on which they are read in the synagogues of the Ashkenazim (central and eastern European Jews and their descendants). They did not originally form a unit and were found scattered in the Bible in their supposed historical position. In the so-called Leningrad Codex of the year 1008 CE, on which the third and subsequent editions of *Biblica Hebraica* edited by Rudolf Kittel are based, the five are grouped together but in a historical order. Nevertheless, their appearance usually follows the order of the liturgical calendar:

Song of Solomon	Pesaḥ (Passover)	March–April
Ruth	Shavuot (Feast of Weeks)	May–June
Lamentations	Tisha be-Av (Fast of Av 9)	July–August
Ecclesiastes	Sukkot (Feast of Tabernacles)	September–October
Esther	Purim (Feast of Lots)	February–March

The five books have little in common apart from their roles in the liturgy. Although the Song of Solomon and Lamentations are poetic in form and Ruth and Esther are stories of heroines, the contrast in the moods and purposes of both pairs sharply distinguishes the books. Ecclesiastes is a product of the Hebrew wisdom movement and exhibits the most pessimistic tone of any book in the Hebrew Bible.

Song of Solomon. The Song of Solomon (also called Song of Songs and Canticle of Canticles) consists of a series of love poems in which lovers describe the physical beauty and excellence of their beloved and their sexual enjoyment of each other. The Hebrew title of the book mentions Solomon as its author, but this seems improbable, primarily because of the late vocabulary of the work. Although the poems may date from an earlier period, the present form of the book is late, perhaps as late as the 3rd century BCE, and its author remains unknown.

Interpretations of the Song of Solomon

The Song of Solomon has been interpreted in many different ways, four of which are noteworthy. The allegorical interpretation takes the book as an allegory of God's love for Israel or of Christ's love for the church. Such a view seems gratuitous and incompatible with the sensuous character of the poems. The dramatic interpretation is based on the dialogue form of much of the book and attempts to find a plot involving either a maiden in Solomon's court and the King or the maiden, the King, and a shepherd lover. The absence of drama in the Semitic literatures and the episodic character of the book make this theory highly improbable. The cultic–mythological interpretation connects the book with the fertility cults of the ancient Near Eastern world. The vigorous condemnation in the Hebrew Bible of such rituals makes it difficult to accept this view, unless it is assumed that the original meaning of the poems was forgotten. The literal interpretation considers the book to be a collection of secular love poems, without any religious implications, that were, perhaps, frequently sung at wedding festivities. According to this commonly accepted view, the poems were received into the biblical canon despite their secular nature and their total lack of mention of God because they were attributed to Solomon and also because they were understood as wedding songs and marriage was ordained by God.

The reasons for the Song of Solomon being read at Passover, which celebrates the Exodus from Egypt, are not entirely clear. Possibly, they include the fact that spring is referred to in the book and that according to the allegorical interpretation the book could refer to God's love for Israel, which is so well evidenced by the events of the Exodus and especially the Covenant at Mt. Sinai.

Ruth. The Book of Ruth is a beautiful short story about a number of good people, particularly the Moabite great-grandmother of David. Though the events described in the work are set in the time of the judges, certain linguistic and other features suggest that the present form dates from postexilic times. It gives the impression, however, of being based on an ancient tradition, perhaps on an ancient written source. In any case, it was certainly grounded on a solid core of fact, for no one would have invented a Moabite ancestress for Israel's greatest king.

The book describes how, during a time of famine, Elimelech, a Bethlehemite, travelled to Moab with his wife, Naomi, and his two sons, Mahlon and Chilion. After his death, the sons married Moabite women, and then they too died, leaving no children. There was thus no one to keep the family line alive and no one to provide for Naomi. Ruth, the widow of Mahlon, dedicated herself to the care of Naomi and insisted on returning with her to her native land and adopting her God. They arrived in

Bethlehem during the harvest, and Ruth went out to work for the two women in the field of Boaz, a wealthy landowner. Naomi urged Ruth to seek marriage with Boaz because he was a kinsman of her late husband, and the firstborn son of such a marriage would count as a son of the deceased. (This resembles the levirate marriage that obliged a man to marry the widow of his deceased brother if the brother died without male issue.) Ruth crept under Boaz' cloak while he slept, and he accepted the implied proposal of marriage. After a nearer kinsman forfeited his claim to Ruth, Boaz married her and a son was born. Thus, loyal Ruth was provided with an excellent husband, the dead Mahlon with a son to keep his name alive, and Naomi with a grandson to support her in her old age.

Many purposes have been assigned to the book: to entertain, to delineate the ancestry of David, to uphold levirate marriage as a means of perpetuating a family name, to commend loyalty in family relationships, to protest the narrowness of Ezra and Nehemiah, the leaders of the postexilic restoration in relation to marriages with non-Jews, to inculcate kindness toward converts to Judaism, to teach that a person who becomes a worshipper of Yahweh will be blessed by him, and to illustrate the providence of God in human affairs. The book may have served all these "purposes," but the author's objective cannot be determined with certainty.

Lamentations of Jeremiah. The Lamentations of Jeremiah consists of five poems (chapters) in the form of laments for Judah and Jerusalem when they were invaded and devastated by the Babylonians in 586 BCE, for the sufferings of the population, and for the poet himself during and after the catastrophe. These grief-stricken laments are intermingled with abject confessions of sin and prayers for divine compassion. The first four poems are alphabetic acrostics; the fifth is not, although like the others it has 22 stanzas, which is the number of letters in the Hebrew alphabet. The formal structure served as a mnemonic device and perhaps was meant to convey the note of wholeness, of Israel's total grief, penitence, and hope. The moving quality of these elegies has suited them for liturgical use. Besides their place in the Jewish liturgy commemorating the anniversary of the destruction of Jerusalem, the laments are employed by the Christian Church to pour out its grief over the Passion and death of Jesus Christ.

Dating the composition of Lamentations

Most critics place the composition of the book before the return of the Jews from exile in 537/536 BCE. Certain passages appear to be word pictures by an eyewitness and would, therefore, have been written shortly after the destruction of Jerusalem. Until the 18th century, the work was universally ascribed to the prophet Jeremiah, and this was supported by a prologue found in the Septuagint and in some manuscripts of the Vulgate. Since that time, however, many scholars have rejected the attribution to Jeremiah chiefly because the ideas and sentiments expressed in Lamentations are unlike those in Jeremiah. Moreover, it is unlikely that the spontaneity and naturalness so characteristic of Jeremiah's utterances could be accommodated to a poetic form as complicated and artificial as that in Lamentations. It is probable that the laments were the product of more than one poet.

Ecclesiastes. The book of Ecclesiastes is a work of the Hebrew wisdom movement, associated by its title and by tradition with King Solomon. It is evident, however, that the book is of much later composition; the author may have identified himself with the famous king and wise man of the past to give greater authority to his work. The language of the book, including the relatively large number of Aramaic forms, and its content point to a date in the early Greek period (later 4th or early 3rd century BCE). That the book was written prior to the 2nd century BCE, however, is shown by its influence on Ecclesiasticus, which was written early in that century, and its appearance among the manuscripts discovered at Khirbat Qumrān, on the northwestern shore of the Dead Sea, where a Jewish community existed in the mid-2nd century.

The name Ecclesiastes is a transliteration of the Greek word that was used in the Septuagint to translate the Hebrew Qohelet, a word connected with the noun *qahal* ("assembly"). Qohelet seems to mean the one who gathers or teaches an assembly; the author used the word as a pseudonym. He appears to be a wisdom teacher writing late in life expressing his skeptical personal reflections in a collection of popular maxims of the day and longer compositions of his own. The book has been described as a sage's notebook of random observations about life. Some interpreters have questioned the unity of authorship, but, given the notebook character of the work, there seems to be little need for questioning the basic integrity of the book.

Although the phrase "vanity of vanities! all is vanity" stressed at both the beginning and the end of the book sums up its theme, it does not convey the variety of tests that the skeptical Qohelet applies to life. He examines everything—material things, wisdom, toil, riches —and finds them unable to give meaning to life. He repeatedly returns to life's uncertainties, to the hidden and incomprehensible ways of God, and to the stark and final fact of death. The only conclusion to this human condition is to accept gratefully the small day-to-day pleasures that God gives to man.

Qohelet stands in sharp contrast to the conventional wisdom schools. He recognizes the relative value of wisdom as against foolishness, but he rejects the oversimplified and optimistic view of wisdom as security for life. He offers a religious skepticism that rejects all facile answers to life's mysteries and God's ways.

Book of Esther. The Book of Esther is a romantic and patriotic tale, perhaps with some historical basis but with so little religious purpose that God, in fact, is not mentioned in it. The book may have been included in the Hebrew canon only for the sake of sanctioning the celebrations of the festival Purim, the Feast of Lots. There is considerable evidence that the stories related in Esther actually originated among Gentiles (Persian and Babylonian) rather than among the Jews. There is also reason to believe that the version given in the Septuagint goes back to older sources than the version given in the Hebrew Bible.

Origin of Esther

Laying the scene at Susa, a residential city of the Persian kings, the book narrates that Haman, the vizier and favourite of King Ahasuerus (Xerxes I; reigned 486–465 BCE), determined by lot that the 13th of Adar was the day on which the Jews living in the Persian Empire were to be slain. Esther, a beautiful Jewess whom the King had chosen as queen after repudiating Queen Vashti, and her cousin and foster father Mordecai were able to frustrate Haman's plans. Haman then schemed to have Mordecai hanged; instead, he was sent to the gallows erected for Mordecai, and Jews throughout the empire were given permission to defend themselves on the day set for their extermination. The governors of the provinces learned in time that Mordecai, who had saved the King from being assassinated by two discontented courtiers, had succeeded to Haman's position as vizier; thus, they supported the Jews in the fight against their enemies.

In the provinces, the Jews celebrated their victory on the following day, but at Susa, where, at Esther's request, the King permitted them to continue to fight on the 14th of Adar, they rested and celebrated their success a day later. Therefore, Esther and Mordecai issued a decree obligating the Jews henceforth to commemorate these events on both the 14th and 15th of Adar.

Theme and language characterize Esther as one of the latest books of the Hebrew Bible, probably dating from the 2nd century BCE. Nothing is known of its author. According to the postbiblical sources, its inclusion in the canon, as well as the observance of the feast of 14th and 15th Adar, still met with strong opposition on the part of the Jewish authorities in Jerusalem as late as the 3rd century CE; yet, despite its lack of specific religious content, the story has become in popular Jewish understanding a magnificent message that the providence of God will preserve his people from annihilation.

Daniel. The Book of Daniel presents a collection of popular stories about Daniel, a loyal Jew, and the

record of visions granted to him, with the Babylonian Exile of the 6th century BCE as their background. The book, however, was written in a later time of national crisis—when the Jews were suffering severe persecution under Antiochus IV Epiphanes (reigned 175–164 BCE), the second Seleucid ruler of Palestine.

The exiled Jews had been permitted to return to their homeland by Cyrus II the Great, master of the Medes and Persians, who captured Babylon in 539 BCE from its last king, Nabonidus, and his son Belshazzar. The ancient Near East was then ruled by the Persians until Alexander the Great brought it under his control in 331. After Alexander's death in 323, his empire was divided among his generals, with Palestine coming under the dominion of the Ptolemies until 198, when the Seleucids won control. Under the Persian and Ptolemaic rulers the Jews seem to have enjoyed some political autonomy and complete religious liberty. But under Antiochus IV Jewish fortunes changed dramatically. In his effort to Hellenize the Jews of Palestine, Antiochus attempted to force them to abandon their religion and practice the common pagan worship of his realm. Increasingly sterner restrictions were imposed upon the Jews, the city of Jerusalem was pillaged, and, finally, in December 167 the Temple was desecrated. The outcome of this persecution was the open rebellion among the Jews, as described in the books of Maccabees. This period of Hellenistic Judaism is treated more fully in JUDAISM, HISTORY OF.

The conflict between the religion of the Jews and the paganism of their foreign rulers is also the basic theme of the Book of Daniel. In Daniel, however, it is regarded as foreseen and permitted by God to show the superiority of Hebrew wisdom over pagan wisdom and to demonstrate that the God of Israel will triumph over all earthly kings and will rescue his faithful ones from their persecutors. To develop this theme the author makes use of a literary and theological form known as apocalypse (from the Greek *apokalypsis*, "revelation" or "unveiling"), which was widely diffused in Judaism and then in Christianity from 200 BCE to 200 CE. Apocalyptic literature professes to be a revelation of future events, particularly the time and manner of the coming of the final age when the powers of evil will be routed in bloody combat and God's kingdom will be established. This revelation usually occurs as a vision expressed in complicated, often bizarre symbolism. The literature is generally pseudonymous, proposed under the name of some authoritative figure of the distant past, such as Daniel, Moses, Enoch, or Ezra. This allows the author to present events that are past history to him as prophecies of future happenings.

Apocalyptic literature

The Book of Daniel, the first of the apocalyptic writings, did not represent an entirely new type of literature. Apocalypse had its beginnings in passages in the works of the prophets. In fact, it has been said that the apocalyptic was really an attempt to rationalize and systematize the predictive side of prophecy. There were significant differences, however. The prophet, for the most part, declared his message by word of mouth, which might subsequently be recorded in writing. The apocalyptist, on the other hand, remained completely hidden behind his message, which he wrote down for the faithful to read. The prophets normally spoke in their own name a message for their own day. The apocalyptists normally wrote in the name of some notable man of the past a message for the time of the age to come.

Like the prophets before them, the apocalyptists saw in the working out of history, which they divided into well-defined periods, a purpose and a goal. The evil in the world might lead men to despair, but the predetermined purpose of God could not be frustrated. A future age of righteousness would replace the present age of ungodliness, and God's purpose would at last be fulfilled. This literature, then, is a mixture of pessimism—times would become worse and worse, and God would destroy this present evil world—and of optimism—out of turmoil and confusion God would bring in his kingdom, the goal of history.

For many centuries the apocalyptic character of the Book of Daniel was overlooked, and it was generally considered to be true history, containing genuine prophecy. In fact, the book was included among the prophetic books in the Greek canon. It is now recognized, however, that the writer's knowledge of the exilic times was sketchy and inaccurate. His date for the fall of Jerusalem, for example, is wrong; Belshazzar is represented as the son of Nebuchadrezzar and the last king of Babylon, whereas he was actually the son of Nabonidus and, though a powerful figure, was never king; Darius the Mede, a fictitious character perhaps confused with Darius I of Persia, is made the successor of Belshazzar instead of Cyrus. By contrast, the book is a not inconsiderable historical source for the Greek period. It refers to the desecration of the Temple in 167 and possibly to the beginning of the Maccabean revolt. Only when the narrative reaches the latter part of the reign of Antiochus do notable inaccuracies appear—an indication of a transition from history to prediction. The book is thus dated between 167 and 164 BCE.

Date, hero, and languages of Daniel

Other considerations that point to this 2nd-century date are the omission of the book from the prophetic portion of the Hebrew canon, the absence of Daniel's name in the list of Israel's great men in Ecclesiasticus, the book's linguistic characteristics, and its religious thought, especially the belief in the resurrection of the dead with consequent rewards and punishments.

The name Daniel would appear to refer to a legendary hero who was used in different ways at different times and who became particularly popular in the storytelling of the Persian and Greek Diaspora as a personification of the practical and theological problems faced by the Jews in that environment. Whether there is any connection between the Daniel of this book and the one mentioned as a wise man without equal in the Book of Ezekiel and as a righteous man in the tale of Aqhat, a Ugaritic text dated from about the middle of the 14th century, is uncertain.

The book is written in two languages: the beginning (1:1–2:4a) and the final chapters (8–12) in Hebrew and the rest in Aramaic. This offers no proof of multiple authorship, however, because the linguistic divisions do not correspond to the division by literary form: chapters 1–6 are stories of Daniel and his friends in exile, and chapters 7–12 are Daniel's apocalyptic visions. Furthermore, there is a singleness of religious outlook, spirit, and purpose throughout. Nevertheless, the problem of the languages has never been satisfactorily answered.

The stories and visions of Daniel

The stories of the first six chapters, which probably existed in oral tradition before the author set them down, begin with the account of how Daniel and his three companions (Hananiah, Mishael, and Azariah, who were given the names Shadrach, Meshach, and Abednego by the Babylonians) came to be living at the Babylonian court and how they remained faithful to the laws of their religion. This is followed by five dramatic episodes calculated to demonstrate the wisdom and might of Israel's God and the unconquerable steadfastness of his loyal people. Thus, through God's gift of wisdom, Daniel excels the professional sages of the pagan court by revealing and interpreting Nebuchadrezzar's dream of a great image, made of four metals, which was shattered by a stone cut without human hand, and then the King's further dream of a tree reduced to a stump, which presaged the punishment of his arrogance by madness, and, finally, the writing on the wall, which spelled Belshazzar's doom at his sacrilegious feast. By trust in God, Daniel's companions, who refused to worship Nebuchadrezzar's golden idol, are miraculously delivered from a fiery furnace, and Daniel himself, thrown into a den of lions for holding fast to his tradition of prayer, is divinely protected.

The last six chapters of the book are apocalyptic. In chapter 7 Daniel is granted a vision of four beasts from the abyss, which are brought under divine judgment, and of "one like a son of man," who is brought before God to be invested with his universal and everlasting sovereignty. The mythological beasts are interpreted as four empires (the Babylonian Empire, the kingdom of the Medes, the Persian Empire, and the empire of Alexander) and the manlike figure as Israel. The vision of a

battle between the ram (Medes and Persians) and the goat (the Greek Empire) in chapter 8 introduces the iniquities of Antiochus IV Epiphanes and is an assurance to the stricken Jews that the end of their tribulation is near. In chapter 9 the author reinterprets the prophecy of Jeremiah that Jerusalem's desolation would end after 70 years. By making these 70 years mean 70 "weeks of years" (*i.e.*, 490 years), the author is again able to focus attention on the period of Antiochus' persecution in the 2nd century and on the imminence of his determined doom. A precise understanding of the author's scheme is not possible, however, because 490 years calculated from the beginning of the exile extends far beyond the time of Antiochus. The remaining chapters provide the fourth commentary on the crisis provoked by the Seleucid tyrant. The greater part of this vision is a sketch of the events that affected the Jews from the Persian period to the time of Antiochus and prepared for his reign of terror. After chapter 11, verse 39, the account of Antiochus' life ceases to correspond with historical fact; an inaccurate prediction of his end is the prelude to the announcement of the end of Israel's tribulation and the inauguration of God's kingdom.

The purpose of the whole book, stories and visions alike, is to encourage Israel to endure under the threat of annihilation and to strengthen its faith that "the Most High rules the kingdom of men" and will in the end give victory to his people and establish his kingdom.

The Chronicler's history

Ezra, Nehemiah, and Chronicles. The final books of the Hebrew Bible are the books of Chronicles and Ezra–Nehemiah, which once formed a unitary history of Israel from Adam to the 4th century BCE, written by an anonymous Chronicler. That these books constituted a single work—referred to as the Chronicler's history, in distinction to the Deuteronomic history and the elements of history from the priestly code of the Torah—appears evident because the same language, style, and fundamental ideas are found throughout and because the concluding verses of II Chronicles are repeated at the beginning of Ezra. The purpose of this history seems to have been to trace the origin of the Temple and to show the antiquity and authenticity of its cult and of the formal, legalistic type of religion that dominated later Judaism.

The history that these books record has already been treated in the historical section of this article and is found in greater detail in JUDAISM, HISTORY OF. The concern in this section will be chiefly with the literary and theological aspects of the books, but their contents can be summarized. In I and II Chronicles the author repeats much of the material from earlier historical books, concentrating upon the history of the kingdom of Judah. The First Book of the Chronicles begins with an extensive genealogy of Israel from Adam to the restoration but is primarily a biography of David that adds further facts to the story as given in Samuel. The Second Book of the Chronicles begins with Solomon and goes through the division of the kingdom to the reign of Zedekiah; once again the Chronicler had access to materials that supplemented the account in I and II Kings. In the Book of Ezra he describes the return of the Jews from the Babylonian Exile and the reconstruction of the Temple. He includes lists of the families who returned and the texts of the decrees under which they returned. In the Book of Nehemiah the reconstruction of the city walls of Jerusalem becomes the basis for a meditation upon the relation between God and his people. This book, too, contains lists of those who participated in the reconstruction, but much of it concentrates upon the description of Nehemiah and his persistence in performing his assignment.

Divisions, author, date, and sources

The fourfold division of the books derives from the Greek and Latin versions; the more basic twofold division into Chronicles and Ezra–Nehemiah is more complex. This original division apparently resulted from the inclusion of the material known as Ezra–Nehemiah in the Hebrew canon before that known as Chronicles because it contained fresh information not found in any other canonical book. When Chronicles was later admitted to the canon, it was placed in order after Ezra–Nehemiah; although the book has retained this position in the Hebrew Bible, the Greek version restored it to its proper sequence. That Chronicles was thus "left aside" may account for the choice of *Paraleipomena* ("Things Omitted") as the Greek title of the book, but the usual and perhaps correct explanation is that Chronicles contains stories, speeches, and observations that were omitted from the parallel accounts in earlier books.

Jewish tradition has identified Ezra as the author of these books, and some modern scholars concur. According to many critics, however, the Chronicler was a Levite cantor in Jerusalem. This position is supported by the author's concern with the Levites and cultic musicians. The date of the work is more difficult to pinpoint. In its final form it has to be later than Ezra, who came to Judah about 400 BCE. An indication of the latest date at which the entire work could have been completed is its silence about the Hellenizing of Judaism that took place after Alexander the Great. This, together with language considerations that point to the late Persian period, has led the majority of commentators to postulate a 4th-century date. Some scholars, however, claim that a time before 300 BCE would be too short to account for the genealogy at the beginning of I Chronicles, which is carried down to the eighth generation after Zerubbabel, one of the leaders of the band that returned from Babylon. Thus, they push the final date to about 200 BCE or even slightly later. It is possible that the 4th-century work of the Chronicler went through a series of minor additions and adaptations until sometime early in the 2nd century, when it reached its final form.

The Chronicler had numerous historical sources—both biblical and extrabiblical—at his disposal. He was closely dependent on the books of Samuel and Kings for all of Chronicles except the first nine chapters. Sometimes he even repeated the actual words of his model, though slight textual variations suggest to some that the Hebrew copy he had before him differed a little from that of the canon and corresponded to that which lay behind the Septuagint. But he was also able to consult the final version of the Torah and the whole of the Deuteronomic history. His use of the personal memoirs of Nehemiah is undisputed; the nature of his Ezra source is less clear, but some have regarded a portion of narrative written in the first person as an autobiographical source. He included many lists, genealogies, census reports, and other official documents that may have been preserved as Temple records. The text refers by name to certain documents representing royal histories and prophetic writings about which, as they have not survived, only speculation is possible.

Theological purpose

The Chronicler made use of all these sources, but he was not shackled by them. Although his work has won increasing respect as a historical document, especially as an indispensable source for the restoration period, his purpose was chiefly theological, not historical. He was convinced of the definitiveness of the divine covenant with David. The holy community that was brought into existence by this covenant, that has been maintained by God through the vicissitudes of history, and that has its worship centred on the Temple in Jerusalem is the true kingdom of God. It is the true Israel and is the Chronicler's only concern. Thus, he mentions the northern kingdom and the kings of Israel only to the extent that they figure in the events of Judah. Loyalty to the Davidic line of succession, to Jerusalem, and to the Temple worship were the central elements in the life of God's people according to this writer. All success and failure were the result of such loyalty or disloyalty. Thus, if a king's reign was long and successful, the Chronicler saw it as the reward of God for a life led in obedience to his will; conversely, a king suffered misfortune only if he had sinned. Significantly, the Chronicler devotes much attention to David's part in the development of the liturgy, especially the organization and functions of the Levites, and omits important but uncomplimentary stories about the King that are found in the Deuteronomic history.

In short, the Chronicler traced the reformed liturgy of his day back to David and laid a solid foundation for the

acceptance and conservation of the religious community that he envisioned—a devout community that worshipped joyfully in the Temple with sacrifice and praise and obeyed the Law of Moses. He knew well that the realization of that community in his day was not perfect and that the future had something better in store, but he seems to have been content to accept the existing Davidic leaders in order not to abandon the dynastic hope because of their shortcomings. These books thus provided an apologia for orthodox Judaism (perhaps in the face of opposition from the Samaritans, the inhabitants of the former northern kingdom), and they offer to the modern reader some insight into the postexilic community in Jerusalem, withdrawn into itself and trying to justify, explain, and preserve its existence and its spirituality.

(R.F.)

V. Intertestamental literature

NATURE AND SIGNIFICANCE

Definitions. A vast amount of Jewish literature written in the intertestamental period (mainly 2nd and 1st centuries BCE) and from the 1st and 2nd centuries CE was preserved, for the most part, through various Christian churches. A part of this literature is today commonly called the Apocrypha (Hidden; hence, secret books; singular apocryphon). At one time in the early church this was one of the terms for books not regarded by the church as canonical (scripturally acceptable), but in modern usage the Apocrypha is the term for those Jewish books that are called in the Roman Catholic Church deuterocanonical works—*i.e.*, those that are canonical for Catholics but are not a part of the Jewish Bible. (These works are also regarded as canonical in the Eastern Orthodox churches.) When the Protestant churches returned to the Jewish canon (Hebrew Old Testament) during the Reformation period (16th century), the Catholic deuterocanonical works became for the Protestants "apocryphal"—*i.e.*, non-canonical.

Meaning of Apocrypha and Pseudepigrapha among Roman Catholics, Protestants, and Jews

In 19th-century biblical scholarship a new term was coined for those ancient Jewish works that were not accepted as canonical by either the Catholic or Protestant churches; such books are now commonly called Pseudepigrapha (Falsely Inscribed; singular pseudepigraphon), *i.e.*, books wrongly ascribed to a biblical author. The term Pseudepigrapha, however, is not an especially well suited one, not only because the pseudepigraphic character is not restricted to the Pseudepigrapha alone—and, indeed, not even all Pseudepigrapha are ascribed to any author, since there are among them anonymous treatises—but also because the group of writings so designated by this name necessarily varies in the different modern collections. Theoretically, the name Pseudepigrapha can designate all ancient Jewish writings that are not canonical in the Catholic Church. The writings of the philosopher Philo of Alexandria (1st century BCE–1st century CE) and the historian Josephus (1st century CE) and fragments of other postbiblical Hellenistic Jewish historians and poets, however, usually are excluded. Rabbinic literature (2nd century BCE–2nd century CE) also is generally excluded; such literature existed for centuries only in oral form (see TALMUD and MIDRASH). The edition of the Pseudepigrapha edited by the British biblical scholar R.H. Charles in 1913, however, contains a translation of *Pirqe Avot* ("Sayings of the Fathers"), an ethical tractate from the Mishna (a collection of oral laws), and even the non-Jewish *Story of Aḥikar* (a folklore hero), though other genuine Jewish writings from antiquity are omitted. Some of the Jewish Pseudepigrapha were discovered only in the last two centuries, and the Dead Sea Scrolls (the first of them discovered in the 1940s), which belong to this category, are not yet all published. Thus, in the broader meaning of the terms, the Apocrypha and Pseudepigrapha are a bloc of Jewish literature written in antiquity from the later Persian period (*c.* 4th century BCE) and not canonized by the Jews.

Texts and versions. A small portion of this literature is preserved in the original languages: Hebrew, Aramaic, and Greek. Most of the Hebrew or Aramaic works, however, exist today only in various translations: Greek, Latin, Syriac, Ethiopian, Coptic, Old Slavonic, Armenian, and Romanian. All the works of the Apocrypha are preserved in Greek, because they have for the Greek Church a canonical value. Those books not considered canonical by the early church have often fallen into oblivion, and their Greek text was often lost; many of the ancient Jewish Pseudepigrapha are today preserved only in fragments or quotations in various languages, and sometimes only their titles are known from old lists of books that were rejected by the church.

Of this literature only the Apocrypha (contained in Latin and Greek Bibles) were read in the liturgical services of the church. The Pseudepigrapha, in their various versions, were in most cases nearly forgotten; and manuscripts of most of them were rediscovered only in modern times, a process that continues. The discovery of the Dead Sea Scrolls at Qumrān in the Judaean desert not only furnished new texts and fragments of unknown and already known Pseudepigrapha but also contributed solutions to problems concerning the origin of other Jewish religious writings (including some Old Testament books), the connection between them, and even their composition and redaction from older sources. The new original texts also strengthened interest in the Jewish literature of the intertestamental period because of its importance for the study both of ancient Judaism and early Christianity. As a result of such discoveries, better critical editions of the Apocrypha and Pseudepigrapha have been published, as well as new studies of their content.

Rediscoveries and new discoveries of texts

The Apocrypha, the texts of which originated mostly in the period before the rise of Christianity, were regarded as canonical in the early church but contain no Christian interpolations. Many of the Pseudepigrapha, however, were interpolated by Christian writers. The nature and the extent of these Christian interpolations is often difficult to define since a Christian interpolator not only changes the text according to his Christian views or introduces specific Christian terminology, but he also may introduce in a Jewish text ideas, motifs, or terminology that are common to both Judaism and Christianity. For these reasons it is sometimes difficult to decide if a passage in a pseudepigraphon, or even sometimes the whole work, is Jewish or Christian.

Persian and Hellenistic influences. Some of the Apocrypha (*e.g.*, Judith, Tobit) may have been written already in the Persian period (6th–4th century BCE), but with these possible exceptions, all the Apocrypha and Pseudepigrapha were written in the Hellenistic period (*c.* 300 BCE–*c.* 300 CE). Yet the influence of Persian culture and religion sometimes can be detected even in comparatively late Jewish works, especially in Jewish apocalyptic literature (see below *Apocalypticism*). The Persian influence was facilitated by the fact that both the Jewish and Persian religions are iconoclastic (against the veneration or worship of images) and opposed to paganism and display an interest in eschatology (doctrines of last times).

Although such an affinity did not exist between Judaism and Hellenistic culture, literary activity among Hellenistic Jews was generally Greek in character: the Greek-writing Jewish authors thought mainly in Greek concepts, used genuine Greek terminology, and wrote many of their works in Greek literary forms.

Though Hellenistic Jewish authors sometimes imitated biblical forms, they learned such forms from their Greek Bible (the Septuagint). Many Greek products written by Jews served as religious propaganda and probably influenced many pagans to become proselytes, or at least to abandon their heathen faith and to become "God fearing." Thus, the Jewish literature written in Greek could be later used by Christianity for similar purposes.

Greek influence on Jewish writings written in Hebrew or Aramaic in Palestine in the intertestamental period was by no means as significant as upon Jewish works written in Greek among the Hellenistic Diaspora (Jews living away from Palestine). In Palestine, religion and culture formed a unity, and the Hellenization of the upper classes in Jerusalem before the Maccabean wars (168–142 BCE) was restricted to some families who had

Greek influence

accepted Greek civilization for practical purposes. Jews in Palestine developed a flourishing autonomous culture based upon religious ideals. Living without interruption in their powerful religious tradition and with their own non-Greek education, the Palestinian Jews were able to produce literary works without significant evidences of Greek influence. The language of this literature was both Aramaic and Hebrew. Under the national revival in the Maccabean period, Hebrew became prevalent as the language of Jewish literature in Palestine; but since Aramaic was a spoken language in Palestine during the whole period, some of the extant literary works of Palestinian Jews in the Maccabean and Roman period probably were originally written also in Aramaic.

Apocalypticism. In intertestamental Jewish literature a special trend developed: namely, apocalypticism. *Apokalypsis* is a Greek term meaning "revelation of divine mysteries," both about the nature of God and about the last days (eschatology). Apocalyptic writings were composed both in Judaism and Christianity; one of them (the Book of Daniel) was accepted in the Jewish canon and another (the book of Revelation) in the New Testament. Other apocalypses form a part of the Pseudepigrapha, and influences of apocalypticism or similar approaches are found in some of the Apocrypha. The sectarian Dead Sea Scrolls are the works of an apocalyptic movement, though not all are written in the style of apocalypses. *The Sibylline Oracles* are, in their Jewish passages, a part of Jewish Hellenistic literature; inasmuch as they contain eschatological prophecies of future doom and salvation, they are apocalyptic, but in their polemics against idolatry and their apology for Jewish faith, they are a product of Jewish Hellenistic propagandistic literature. Because one of the central themes of apocalypticism is that of future salvation, messianic hopes involving the advent of a deliverer are usually the object of intertestamental Jewish apocalypticism.

APOCRYPHAL WRITINGS

Apocryphal works indicating Persian influence. *Esdras.* The "Greek Ezra," sometimes named I (or II or III) Esdras, enjoyed considerable popularity in the early church but lost its prestige in the Middle Ages in the Latin Church. At the reforming Council of Trent (1545–63), the Roman Catholic Church no longer recognized it as canonical and relegated it in the Latin Bible to the end, as an appendix to the New Testament. One of the reasons for its non-canonicity in the West is that the "Greek Ezra" contains parallel material to the biblical books of Chronicles, Ezra, and Nehemiah but differs in textual recension (points of critical revision) and occasionally in the order of the stories. The content of the book is a history of the Jews from the celebration of the Passover in the time of King Josiah (7th century BCE) to the reading of the Law in the time of Ezra (5th century BCE). Though written in an idiomatic Greek, "Greek Ezra" is probably a Greek translation from an unknown Hebrew and Aramaic redaction of the materials contained in the biblical books of Chronicles, Ezra, and Nehemiah. An important part of this book (3:1–5:6), the story of the three youths at the court of Darius, has no parallel in the canonical books. This story concerns a debate between three guardsmen before Darius, king of Persia, about the question of what they consider to be the strongest of all things; the first youth asserts that it is wine, the second says that it is the king, and the third, who is identified with the biblical Zerubbabel (a prince of Davidic lineage who became governor of Judah under Darius), expresses his opinion that "women are strongest, but truth is victor over all things." He is acclaimed as the victor, and, as a reward, he requests that Darius rebuild Jerusalem and its Temple. The story evidently was written in two stages: originally, the competition was about wine, the king, and women, but later, truth was added. Truth is one of the central concepts of Persian religion and the competition itself is before a Persian king; thus it seems likely that the story is Persian in origin and that it became Jewish by the identification of the third youth with Zerubbabel.

The three youths at the court of Darius

Judith. The book of Judith is similar to the biblical Book of Esther in that it also describes how a woman saved her people from impending massacre by her cunning and daring. The name of the heroine occurs already in Gen. 26:34 as a Gentile wife of Esau, but in the book of Judith it evidently has symbolic value. Judith is an exemplary Jewess. Her deed is probably invented under the influence of the account of the 12th-century-BCE Kenite woman Jael (Judg. 5:24–27), who killed the Canaanite general Sisera by driving a tent peg through his head.

The story is clearly fiction, and the anachronisms in it are intentional: they show that the story itself is a mere fiction. The book speaks about the victory of Nebuchadnezzar, "who reigned over the Assyrians at Nineveh" (the name is of the 7th–6th century BCE king of Babylon, Nebuchadrezzar) in the time of an unknown Arphaxad, king of the Medes. Since the western nations of Nebuchadnezzar's empire had refused to come to his aid, the King ordered his commander in chief, Holofernes (a Persian name), to force submission upon the rebellious nations. In subduing these nations Holofernes destroyed their sanctuaries and proclaimed that Nebuchadnezzar alone should henceforth be worshipped as a god. Thus, the Jews, who had recently returned from the Babylonian Captivity (6th century BCE) and rebuilt the Temple, were compelled to prepare for war. Holofernes laid siege to Bethulia (otherwise unknown), described as an important strategic point on the way to Jerusalem. Because of a long siege, the inhabitants wanted to surrender their city, but Judith persuaded the people to delay the surrender for five days. Judith was a virtuous, pious, and beautiful widow. She removed her mourning garments, left the city, entered Holofernes' camp, and was brought before him. On the fourth day, Holofernes decided to seduce Judith and invited her to come into his tent; he then drank more wine than ever before. After he fell into a drunken stupor, Judith cut off his head with his sword and returned with the head to Bethulia. The Jews put Holofernes' head outside the city wall, and the following morning, upon learning of the death of their commander in chief, the Assyrian soldiers dispersed and were pursued by the Jews of Bethulia, who took abundant spoil. The Jews were not threatened again during Judith's lifetime —she lived to be 105—or for long thereafter.

The killing of Holofernes

Many suggestions have been made about the book of Judith's date of composition. Though current scholarly opinion is that the book was written in the warlike patriotic atmosphere of the early Maccabean period (*c.* 150 BCE) by a Palestinian Jew, there are no Maccabean elements in the book. It shows no direct or indirect Greek influences, the deification of kings existed already in the ancient Near East, and the political situation described in the book has nothing in common with the Maccabean period. All the apparently intentional historical mistakes, however, can be understood if it is suggested that the book of Judith was written under Persian rule. Holofernes is, as noted above, a typical Persian name; and the whole political and social situation described in the book fits the Persian world, as do the Jewish life and institutions reflected in the book. Thus, there are no serious indications that the book of Judith is a Maccabean product, and there are many allusions to the time of the Persian rule over Palestine. Only a Greek translation of the book is extant, but, from its style, it is clear that the book was originally written in Hebrew. In his preface to the book of Judith, the Latin biblical scholar Jerome (*c.* 347–419/420 CE) states that he used for his translation a "Chaldaean" (*i.e.*, Aramaic) text and that he also used an older Latin translation from Greek. His translation differs in many points from the original text.

Tobit. The other Jewish short story possibly dating from Persian times is the book of Tobit, named after the father of its hero. From the fragments of the book discovered at Qumrān, scholars now know that the original form of the name was Tobi. Tobit was from the Hebrew tribe of Naphtali and lived as an exile in Nineveh; his son was Tobias. Obeying the tenets of Jewish piety, Tobit buried the corpses of his fellow Israelites who had been

executed. One day, when he buried a dead man, the warm dung of sparrows fell in his eyes and blinded him. His family subsequently suffered from poverty, but then Tobit remembered that he had once left a deposit of silver at Rages (today Teheran) in Media. He sent his son Tobias along with a companion, who was in reality the angel Raphael under the guise of an Israelite, to retrieve the deposit. During the journey, while Tobias was washing in the Tigris, a fish threatened to devour his foot. Upon instructions from Raphael, Tobias caught the fish and removed its gall, heart, and liver, since it was believed that the smoke from the heart and liver had the power to exorcise demons and that ointment made from the gall would cure blindness. On the way he stopped at Ecbatana (in Persia), where Raguel, a member of Tobias' family, lived. His daughter Sarah had been married seven times, but the men had been slain by the demon Asmodeus on the wedding night, before they had lain with her. On the counsel of Raphael, Tobias asked to marry Raguel's daughter, and on the wedding night Tobias put Asmodeus to flight through the stench of the burning liver and heart of the fish. Raphael went to Rages and returned with the deposit. When he returned with his young wife and Raphael to Nineveh, Tobias restored his father's sight by applying the gall of the fish to his eyes. Raphael then disclosed that he was one of God's seven angels and ascended into heaven.

"The Grateful Dead"

The story of the book of Tobit is a historicized and Judaized version of the well-known folktale of "The Grateful Dead" (or "The Grateful Ghost"), in which, after burying a corpse in dangerous circumstances, a young man obtains a bride through the help of the deceased. Asmodeus (in Persian, Aeshma Daeva, the demon of wrath) occurs as a powerful demon in rabbinic literature as well as in folktales. In the Jewish form of the story, "The Grateful Dead" is replaced by the angel Raphael. According to the *Ethiopic Enoch* (20:3; 22:3), Raphael is appointed over the spirits of the souls of the dead (for *Enoch*, see below). Because the cause of this situation is not mentioned in the book of Tobit, the story itself in its Jewish form probably existed before it became the subject of the book of Tobit. The present work is a literary product; the interesting plot gave to the author many occasions to insert religious and moral teachings in the manner of wisdom literature, which is concerned with practical, everyday issues. The book contains prayers, psalms, and aphorisms, most of them put in the mouth of Tobit. It is the oldest Jewish witness of the golden rule (4:15): "And what you hate, do not do to anyone." Eschatological hopes are also described: at the end of time, all Jewish exiles will return, Jerusalem will be rebuilt of precious stones and gold, and all nations will worship the true God. In these eschatological images, however, the figure of the Messiah does not occur.

The religious, social, and literary atmosphere of the book does not contain elements from the Greek period. Thus, the book probably was written already in the Persian period or in the early days of Greek rule (3rd century BCE). The book exists today in three principal recensions, and it is often difficult to determine, in a particular passage, what was the original text. The book was written in Hebrew or Aramaic; the Greek recensions differ, perhaps because they are based on different Semitic versions. These questions may be answered when the Hebrew and Aramaic fragments of the book, which were found among the Dead Sea Scrolls, are published.

The Story of Ahikar. According to the book of Tobit, Ahikar, the cupbearer of the Assyrian king Esarhaddon, was Tobit's nephew; he is a secondary personage in the plot, and his own story is mentioned. Ahikar, is the hero of a Near Eastern non-Jewish work, *The Story of Ahikar*. The book exists in medieval translations, the best of them in Syriac. The story was known in the Persian period in the Jewish military colony in Elephantine Island in Egypt, a fact demonstrated by the discovery of fragmentary Aramaic papyri of the work dating from 450–410 BCE. Thus, the author of the book of Tobit probably knew *The Story of Ahikar*, in which, as in the book of Tobit, the plot is a pretext for the introduction of speeches and wise sayings. Some of Tobit's sayings have close parallels in the words of the wise Ahikar.

Baruch. The apocryphon of Baruch, which is extant in Greek and was included in the Septuagint, is attributed to Baruch, secretary to the Old Testament prophet Jeremiah (7th–6th century BCE). It was Baruch who read Jeremiah's letter to the exiles in Babylon. After hearing his words, the Jews repented and confessed their sins. The first part of the book of Baruch (1:1–3, 8), containing a confession of sins by the Jews following the destruction of Jerusalem and the exiles' prayer for forgiveness and salvation, may date from the Persian or at least from the pre-Maccabean period. This early section was originally written in Hebrew and seems to be very ancient. The other two parts (3:9–4:4 and 4:5–5:9) were written in Greek or freely translated from Hebrew or Aramaic. The first is a praise of wisdom: only Israel received wisdom from God, which is the Law of Moses. The last part of the book of Baruch contains Jerusalem's lament over her desolation and her consolation.

Apocryphal works lacking strong indications of influence. *The Letter of Jeremiah.* The Letter of Jeremiah, like the book of Baruch, was conserved—together with the Greek translation of the Book of Jeremiah—in the Septuagint. The oldest witness of the letter is a fragment of a Greek papyrus, written about 160 BCE and found among the Dead Sea Scrolls at Qumrān. Whether the letter was originally written in Greek or is a translation from Hebrew or Aramaic is difficult to decide. The letter attacks the folly of idolatry as did Jeremiah's letter "to those who were to be taken to Babylon as captives." Though, according to some experts, the idolatry described in the book fits Babylonian cults, the only clear indication of its date is that of the Qumrān fragment.

Prayer of Manasseh. In some manuscripts of the Septuagint and in two later Christian writings, a pseudepigraphic Prayer of Manasseh is contained. This prayer was composed with reference to II Chron. 33:11–18, according to which the wicked Judaean king Manasseh repented and prayed. In the present form the prayer is Greek in origin, but it may have existed in a Hebrew version, of which the Greek is a free adaptation. The prayer was probably composed (or translated) in the 1st century BCE.

Additions to Daniel and Esther. Two of the Old Testament Hagiographa (Ketuvim; see above *The Hebrew canon*)—Daniel and Esther—contain, in their Greek translations, numerous additions.

The Prayer of Azariah and the Song of the Three Young Men. The first addition to Daniel (in Greek and Latin translations Dan. 3:24–68) contains the Prayer of Azariah and the Song of the Three Young Men. These are the prayers of Hananiah, Mishael, and Azariah, the three young men who praised God after they had been placed in the midst of the fiery furnace during a persecution of Jews in Babylon, as told in the Book of Daniel. The first prayer is said by Azariah alone; the second, a thanksgiving prayer, is said by all three after having been saved by God. The two poems are not found in the original Daniel and were never a part of it. They were translated from Hebrew originals or adapted from them. A passage from the second, a liturgical hymn of praise, is a poetic expansion of the doxology that was sung in the Temple when the holy name of God was pronounced. Like the other additions to Daniel, the two prayers were probably composed before 100 BCE.

The wisdom of Daniel

Susanna. The second addition to Daniel, the story of Susanna, and the third one, Bel and the Dragon, are preserved in two Greek versions. In both stories the hero is the wise Daniel. Susanna was the pious and beautiful wife of Joakim, a wealthy Jew in Babylon. Two aged judges became inflamed with love for her. They tried to force her to yield to their lust, and, when she refused, they accused her of committing adultery with a young man, who escaped. She was condemned to death, but when Daniel cross-examined the two elders separately, the first stated that Susanna had been surprised under a mastic tree, the other under a holm tree. Susanna was thus saved and the two false witnesses executed.

The short story, perhaps invented even before the extant Book of Daniel was composed, could very well be added to Daniel (whose name means God is my Judge). The story was written in its present form in Greek, since it contains two Greek puns, but a written Semitic prototype may have existed.

Bel and the Dragon. The third Greek addition to the Book of Daniel is the story of Bel and the Dragon. The Babylonians worshipped the idol of the god Bel and daily provided him with much food, but Daniel proved to the King that the food was in reality eaten by the priests. The priests were punished by death and Bel's temple destroyed. The Babylonians also worshipped a dragon, but Daniel declined to worship him. To destroy the beast, Daniel boiled pitch, fat, and hair together: the dragon ate it and burst asunder. After Daniel's sacrilege of slaying the dragon, the King was forced to cast Daniel into the lions' den, but nothing happened to him. Indeed, he was given a dinner by the prophet Habakkuk, who was brought there by the hair of his head by an angel. On the seventh day the King found Daniel sitting in the den; so he led Daniel out and cast his enemies into the den, where they were devoured.

The two stories are an attack against idolatry. As the addition ends with the story about Daniel in the lions' den, which is also narrated in the canonical Book of Daniel with another motivation, it is probable that this short treatise originated in a tradition that was parallel to the canonical Book of Daniel and that the two stories were translated from a Hebrew or Aramaic original.

Greek additions to Esther. The Hebrew Book of Esther had a religious and social value to the Jews during the time of Greek and Roman anti-Semitism, though the Hebrew short story did not directly mention God's intervention in history—and even God himself is not named. To bring the canonical book up-to-date in connection with contemporary anti-Semitism and to stress the religious meaning of the story, additions were made in its Greek translation. These Greek additions are (1) the dream of Mordecai (Esther's uncle), a symbolic vision written in the spirit of apocalyptic literature; (2) the edict of King Artaxerxes against the Jews, containing arguments taken from classical anti-Semitism; (3) the prayers of Mordecai and of Esther, containing apologies for what is said in the Book of Esther—Mordecai saying that he refused to bow before Haman (the grand vizier) because he is flesh and blood and Esther saying that she strongly detests her forced marriage with the heathen king; (4) a description of Esther's audience with the King, during which the King's mood was favourably changed when he saw that Esther had fallen down in a faint; (5) the decree of Artaxerxes on behalf of the Jews, in which Haman is called a Macedonian who plotted against the King to transfer the kingdom of Persia to the Macedonians; and (6) the interpretation of Mordecai's dream and a colophon (inscription at the end of a manuscript with publication facts), where the date, namely, "the fourth year of the reign of Ptolemy and Cleopatra" (*i.e.*, 114 BCE), is given. This indicates that the additions in the Greek Esther were written in Egypt under the rule of the Ptolemies.

Greek and Roman anti-Semitism

I and II Maccabees. *I Maccabees.* The first two of the four books of Maccabees are deuterocanonical (accepted by the Roman Catholic Church). The First Book of the Maccabees is preserved in the Greek translation from the Hebrew original, the original Hebrew name of it having been known to the Christian theologian Origen of Alexandria. At the beginning, the author of the book mentions Alexander the Great, then moves on to the Seleucid king of Syria, Antiochus Epiphanes (died 164/163 BCE), and his persecution of the Jews in Palestine, the desecration of the Jerusalem Temple, and the Maccabean revolt. After the death of the priest Mattathias, who had refused to obey Antiochus, his son Judas Maccabeus succeeded him and led victorious wars against the Syrian Greeks. Exactly three years after its profanation by Antiochus, Judas captured the Temple, cleansed and rededicated it, and in honour of the rededication initiated an annual festival (Ḥanukka) lasting eight days. After Judas later fell in battle against the Syrian Greeks, his brother Jonathan succeeded him and continued the struggle. Only in the time of Simon, Jonathan's brother and successor, did the Maccabean state become independent. A short mention of the rule of Simon's son John Hyrcanus I (135–104 BCE) closes the book. The author, a pious and nationalistic Jew and an ardent adherent of the family of Maccabees, evidently lived in the time of John Hyrcanus. The book imitates the biblical style of the historical books of the Old Testament and contains diplomatic and other important—though not necessarily authentic—official documents.

Origin of Ḥanukka

II Maccabees. The Second Book of the Maccabees, or its source, was probably written in the same period as I Maccabees. The book is preceded by two letters to the Jews of Egypt: the first from the year 143 BCE and the second one written in 124 BCE commemorating the rededication of the Temple. In the preface of the book, the author indicates that he has condensed into one book the lost five-volume history compiled by Jason of Cyrene. II Maccabees describes the persecution under Antiochus Epiphanes and the Maccabean wars until the victory of Judas Maccabeus over Nicanor, the commander of the Syrian elephant corps, in 161 BCE. The book, written in Greek, is an important document of Hellenistic historiography. Descriptions of the martyrdom of the priest Eleazar and of the seven brothers under Antiochus, in which Greek dramatic style is linked with Jewish religious spirit, became important for Christian martyrology. The book also furnished proof texts for various Jewish and subsequently Christian doctrines (*e.g.*, doctrines of angels and the resurrection of the flesh).

Wisdom literature. *Ecclesiasticus (or Sirach).* There are two deuterocanonical works of the genre known as wisdom literature, one Hebrew and one Greek. The Hebrew work is called Ecclesiasticus, in the Latin Bible and in Greek manuscripts Sophia Iēsou hyiou Sirach (the Wisdom of Jesus the Son of Sirach); the original Hebrew title was probably Hokhmat Yeshuaʿ Ben-Syra, the Proverbs of Ben-Sira. Written in Hebrew about 185 BCE, it was translated into Greek by the author's grandson in Egypt. A Syriac translation also was made. Portions (about three-fifths) of the Hebrew text were found in medieval copies in a synagogue of Cairo and a part of the book in a fragment of a scroll from Massada in Palestine (written *c.* 75 BCE). Small Hebrew fragments also were found among the Dead Sea Scrolls; one of them, the Psalms scroll, contains a large part of a poem about wisdom that is a part of the appendix (chapter 51) and that was not written by the author. The Proverbs of Ben-Sira are often quoted in rabbinic literature.

The book is written in the poetical style of the wisdom books of the Old Testament (*e.g.*, Proverbs, Job) and deals with the themes of practical and theoretical morality. The religious and moral position of the author is conservative—he does not believe in the afterlife, but he reflects the contemporary religious positions. He identifies wisdom, the origin of which is divine, with "the Law which Moses commanded," an idea that became important for later Judaism. He also reflects contemporary debates about freedom of will and determinism, and, though realistic in his basic opinions, he sometimes expresses eschatological hopes of salvation for his people. His piety is ethical, though lacking in asceticism; and he invites his readers to enjoy life, which is short (in this point some Greek influence is palpable, but it is not very deep). At the end of the book the author praises, in chronological order, "the fathers of old," from the beginning of history to his contemporary, the high priest Simon, whose appearance in the Temple is poetically described. After some verses comes the colophon with the author's name—the last chapter being an appendix not composed by the author.

The themes of intertestamental wisdom literature

The Wisdom of Solomon. The other deuterocanonical wisdom book, the Wisdom of Solomon, was written in Greek, though it purports to have been written by King Solomon himself. The hypothesis that the first half of the book was translated from Hebrew seems to be without foundation and probably came into existence be-

cause, in this section, the author imitated in Greek the Old Testament poetical style. The Wisdom of Solomon was probably written in Alexandria (Egypt) in the 1st century BCE.

The book has three parts. The first (chapters 1–5) concerns the contrast between pious and righteous Jews and the wicked, sinful, and mundane Jews who persecute the righteous; the lot of the righteous is preferable to the sorrows and final condemnation of the sinners. In the second part (chapters 6–9) Solomon speaks about the essence of wisdom and how he attained it. In the third part (chapters 10–19) the author proves the value of wisdom by telling—not in an exact chronological order—how, in the history of Israel from the beginning until the conquest of Palestine, God exalted Israel and punished the heathens, the Egyptians, and the Canaanites. He also describes the folly of heathenism and its origins in human aberrations.

The author fuses Judaism and Hellenism both in style and in thought. Though he imitates biblical style, he is also influenced by Greek rhetoric. He also freely uses Greek philosophical and other terms and is influenced by Jewish apocalyptic literature. Some close parallels to the Dead Sea sect (at Qumrān), both in eschatology and in anthropology (doctrines about man), can be found in the Wisdom of Solomon.

THE PSEUDEPIGRAPHAL WRITINGS

Works indicating a Greek influence. *The Letter of Aristeas.* An important document of Jewish Hellenistic literature is *The Letter of Aristeas,* a pseudepigraphon ascribed to Aristeas, an official of Ptolemy II Philadelphus, a Greek monarch of Egypt in the 3rd century BCE. The letter is addressed to his brother and gives an account of the translation of the Pentateuch (first five books of the Old Testament) into Greek, by order of Ptolemy. According to the legend, reflected in the letter, the translation was made by 72 elders, brought from Jerusalem, in 72 days. The letter, in reality written by an Alexandrian Jew about 100 BCE, attempts to show the superiority of Judaism both as religion and philosophy. It also contains interesting descriptions of Palestine, of Jerusalem with its Temple, and of the royal gifts to the Temple.

Translation of the Pentateuch into Greek

IV Maccabees. Another Jewish Hellenistic work combining history and philosophy is *The Fourth Book of Maccabees.* The theme of the book, reflecting the views of the Greek Stoics, is "whether the Inspired Reason is supreme ruler over the passions." This thesis is demonstrated by the martyrdom of the elderly scribe Eleazar and the unnamed seven brothers and their mother, taken from II Macc. 6:18–7:41. The idea of the expiatory force of martyrdom is stressed more in *IV Maccabees* than in its source. The author probably lived in the 1st century BCE and may have been from Antioch (in Syria), where the tombs of the Maccabean martyrs were venerated by the Jews.

III Maccabees. The Greek book called *The Third Book of Maccabees* itself has nothing to do with the Maccabean period. Its content is a legend, a miraculous story of deliverance, which is also independently told—in another historical context—by Josephus (*Against Apion* II, 5). In III Maccabees the story takes place during the reign of Ptolemy IV Philopator (reigned 221–203 BCE). The central episode of the book is the oppression of Egyptian Jews, culminating with an anti-Jewish decree by the King. The Jews who were registered for execution were brought into the hippodrome outside of Alexandria; the King had ordered 500 elephants to be drugged with incense and wine for the purpose of crushing the Jews, but by God's intercession "the beasts turned round against the armed hosts [of the king] and began to tread them under foot and destroy them." The Jews fixed annual celebrations of this deliverance. The book was probably written at the end of the 1st century BCE by an Alexandrian Jew in a period of high anti-Jewish tension.

The Lives of the Prophets. The little book called the *Lives of the Prophets* is a collection of Jewish legends about Old Testament prophets. It is preserved in Greek and in versions and recensions in various languages, all based on the Greek. The purpose of the work was to furnish to the readers of the Bible further information about the prophets. The collection evidently passed through Christian hands since it includes an assumed prophecy of Jeremiah about the birth of Christ. Thus, the date of composition of the supposed original Jewish work and the question as to whether it was originally written in Hebrew or Greek are difficult to resolve. Scholars are inclined toward a 1st-century-CE date in Palestine—with the exception of the life of "Jeremiah," which is Egyptian in origin.

The Ascension of Isaiah. According to the *Lives of the Prophets,* Jeremiah was stoned to death and Isaiah was sawn asunder. These two legends are reflected in two originally Jewish works. The *Ascension of Isaiah,* in which the martyrdom of Isaiah is narrated, is as a whole extant only in Ethiopic, translated from a Greek original, which itself is also known from fragments. The book contains important Christian passages from the 1st century CE, but the story about Isaiah's martyrdom is most likely based upon a Jewish written source. According to this legend, Isaiah was killed by the wicked king Manasseh, who served Beliar-Sammael, the chief of the evil spirits, instead of God. Isaiah, with his followers, had fled to the wilderness, but upon being captured he was sawn asunder with a wooden saw, and his followers fled to the region of Tyre and Sidon. The activity of Beliar is known also from the writings of the sect that preserved the Dead Sea Scrolls and similar writings, and the story itself resembles in some way the history of the Dead Sea sect; but no fragment of the Jewish part of the book was found among the Dead Sea Scrolls. The original *Martyrdom of Isaiah* was written probably in Hebrew or Aramaic before the 1st century CE.

The legendary death of Isaiah

Paralipomena of Jeremiah. In the last chapter of the Greek text of the *Paralipomena* (chronicles) *of Jeremiah,* there is a hint of the Christian part of the *Ascension of Isaiah:* the people stoned Jeremiah to death because he, like Isaiah before him, prophesied the coming of Christ. In a parallel legend (preserved in Arabic), both the violent death of Jeremiah and the Christian motif are lacking. The book begins shortly before and ends shortly after the Babylonian Exile and contains mostly otherwise unknown legends. The legend about the long sleep of Abimelech (the biblical Ebed-melech—an Ethiopian eunuch who rescued Jeremiah from a cistern), who slept and so did not see the destruction of Jerusalem by the Babylonians—is based upon a legendary understanding of Psalm 126:1; a similar legend about another person is preserved in the Talmud (the authoritative rabbinical compendium of Jewish law, lore, and commentary). The book is basically Jewish, and the last chapter was Christianized. The Jewish work was probably written at the end of the 1st century CE or at the beginning of the 2nd, originally in either Hebrew, Aramaic, or Greek.

The Testament of Job. Though there are scholars who think that the *Testament of Job* was once written in Hebrew or Aramaic, it is more probable that the existing Greek text of the book is the original or even a rewritten later version of a Greek work; a fragment of an older form is probably preserved in the Greek translation of Job (2:9). Job is identified, according to some Jewish traditions, with the biblical Jobab (king of Edom), and his (second) wife is Dinah, Jacob's daughter. Job knew by revelation that, for destroying an idol, he would undergo suffering but that a happy end would be the final outcome. Thus, in contrast to the biblical Book of Job, this work does not deal with the question of God's righteousness but places great emphasis on resurrection and eternal life. These special motifs in the book indicate that the book probably was written by a member of an unknown Jewish group that upheld a high mystical spirituality. The extreme "pietistic" tendency of the book is noted in the exaggeration of Job's love for suffering and of his charity to the poor. At the end of the book Job's soul was taken to heaven in a heavenly chariot. The book was probably written before 70 CE.

Life of Adam and Eve. The many Christian legends in many languages about the lives of Adam and Eve probably have their origin in a Jewish writing (or writings) about the biblical first man and woman. The most important of these works are the Latin *Vita Adae et Evae* (*Life of Adam and Eve*) and a Greek work closely parallel to it, named erroneously by its first editor the *Apocalypse of Moses.* The narrative runs from the Fall to the deaths of Adam and Eve. The religious message in the story involves the repentance of Adam and Eve after their expulsion from paradise—and the description of their deaths does not show any traces of the idea of original sin, which was important in later Christian theology. Nonetheless, there are definitely Christian passages in the various versions, and the treatment of Adam in the literature of the Ebionites (an early Jewish Christian sect) shows an affinity for the story. Thus, the Jewish source probably was composed in the 1st century CE in Jewish circles that influenced the Ebionites. The original language of this supposed source is unknown.

Apocalyptic and eschatological works. *III Baruch.* Apocalyptic literature was much concerned about sources of information about the heavenly world and about the places of the damned and saved souls. In later Jewish and early Christian apocalypses, in which the hero undertakes a heavenly trip and sees the secrets that are hidden from others, these sources of information are highly significant. *III Baruch,* a book written in Greek—in which Baruch, the disciple of the prophet Jeremiah, visits the universe and sees its secrets and the places of the souls and of the angels—is such an apocalypse. In the Greek text the number of heavens visited by Baruch is five, but it is possible that originally he was said to have seen seven heavens. There are Christian passages in the book, but it seems to have been a Jewish work from the 1st century CE later rewritten by a Christian.

Interests of apocalyptic literature

II Enoch. Similar in content is *II Enoch,* or *The Book of the Secrets of Enoch,* which is preserved only in an Old Slavonic translation. The oldest text does not contain any Christian additions nor any passage from which it could be concluded that the book was written in Greek. Thus, the book could have been written originally in Hebrew or Aramaic, probably in the 1st century CE. The hero who visits the heavens is the biblical Enoch (son of Jared). The author of the book knew at least some of the treatises contained in *I Enoch.* The book also contains the story of the miraculous birth of the biblical priest-king Melchizedek.

The Psalms of Solomon. Other Jewish apocalypses or books containing eschatological elements did not deal with the mysteries of celestial worlds but rather with the political aspect of apocalyptic thought and with the last days and the messianic age. This latter theme is one of the important motifs of the *Psalms of Solomon,* a book written originally in Hebrew; only the Greek translation of the *Psalms* is preserved. The title is evidently a later addition—the author himself apparently had no intention to give the impression that his 18 psalms were composed by the biblical king Solomon. The *Psalms of Solomon* were written in Jerusalem about the middle of the 1st century BCE, and, though persons are not named, they reflect the dramatic events of the Jewish history of that period, especially the Roman general Pompey's conquest of Jerusalem in 63 BCE and his violent death in Egypt. In Psalm 17, the author denounces the Hasmonean dynasty as illegal and describes the coming of the Davidic Messiah (a kingly saviour from the line of David). His religious opinion resembles the teachings of the Pharisees (a sect that espoused a reinterpretation of Jewish laws and customs), especially in his faith in the resurrection of the body and in the question of free will, though he most likely was not a Pharisee but rather a member of the community of Ḥasidim, a Jewish pietistic group that had joined the Maccabean revolt from its beginning.

The Assumption of Moses. The *Assumption of Moses* originally contained apocalyptic material—no longer extant—in the form of a legend. According to Origen, the dispute between the archangel Michael and the devil for the body of Moses was narrated in the *Assumption of Moses.* This legend, which has parallels in the rabbinic literature, probably formed the end of the *Assumption of Moses,* the first part of which was discovered in a Latin manuscript. The Latin version was translated from Greek, but the original language was Semitic, probably Hebrew.

The main content of the preserved part is Moses' prophecy about the future, from his time until the Kingdom of Heaven will be revealed. According to the custom of apocalyptic literature, names of persons and groups are not mentioned, but from the last events hinted at in the book it can be assumed that it was written at the beginning of the 1st century CE, while Jesus was alive. In its older version, the book apparently was written at the beginning of the Maccabean revolt, some years before the Book of Daniel; after a description of the pre-Maccabean Hellenistic priests (chapter 5) and before the description of the persecutions by Antiochus Epiphanes (chapter 8), chapters 6–7 contain Jewish history from the time of the later Hasmonean rulers to the time of the sons of Herod —as well as polemics against leading religious circles, which are accused of religious hypocrisy, as are the Pharisees in the Christian Gospels. The author of these chapters (6–7), a contemporary of Jesus, evidently erroneously identified the wicked pre-Maccabean priests with the wicked late Maccabean priestly rulers and also interpreted Antiochus Epiphanes as a kind of eschatological Antichrist. No messianic figure is mentioned in the eschatological description of the Kingdom of God: God himself and his angel will bring the salvation.

The coming of the Kingdom of Heaven

The Sibylline Oracles. The Sibylline Oracles is a collection of oracles in Greek verse containing pagan, Jewish, and Christian material from various periods. It comprised 15 books (books IX, X, and XV are lost), of which 4,240 verses are extant. Sibyl is the name (or title) of a legendary ancient pagan prophetess. In the Hellenistic period, eastern nations fabricated Sibylline oracles as propagandistic literature against Greek and, later, against Roman occupation. The political anti-Roman and anti-pagan tone is typical of the Jewish and Christian parts of Sibylline oracles; they also contain religious propaganda for the respective religion. Because Jewish parts used pagan material and Christian authors interpolated Jewish parts or used Jewish material, it is sometimes difficult to decide what verses are pagan, Jewish, or Christian. *The Sibylline Oracles* perhaps became a part of Jewish (and Christian) apocalyptic literature because of their emphasis on eschatology. The oldest Jewish "Sibyl" is contained in the third book: it dates from about 140 BCE and describes the coming of the Messiah. Book IV was written by a Jew about 80 CE: the eruption of Vesuvius (79) is viewed as a divine punishment for the massacre of Jews in the Roman war (70). Book V was written by a Jew about 125.

II Esdras (or IV Esdras). Two important apocalyptic pseudepigrapha (II Esdras and the *Apocalypse of Baruch*), in which the political and eschatological aspects are central to the aim of the books, were written in Palestine at the end of the 1st century CE as a consequence of the catastrophic destruction of the Second Temple in Jerusalem (70). Both were written as if they reflected the doom that befell the people of Israel after the destruction of the First Temple (586 BCE) by the Babylonians. II Esdras (or IV Esdras) was written in Hebrew, but only various translations from a lost Greek version are preserved. The Latin version (in which chapters 1–2 and 15–16 have been added by a Christian hand) at one time was printed at the end of the Latin Bible. The book consists of six visions attributed to the biblical Ezra (who is, at the beginning of the book, erroneously identified with Salathiel, the father of Zerubbabel, a leader of the returning exiles from Babylon). The tragedy of his nation evokes in the heart of the author questions about God's righteousness, the human condition, the meaning of history, and the election of Israel; "Ezra" does not find consolation and full answer in the words of the angel who was sent to him, which also contain revelations about the last days. In the fourth vision "Ezra" sees a mourning woman; she disappears and a city (the New

Questions about the meaning of history

Jerusalem) stands in her place. In the fifth vision a monstrous eagle appears, the symbol of the Roman Empire, and a lion, the symbol of the Messiah. The final victory of the Messiah is described in the last vision of the man (Son of man) coming from the sea. In chapter 14 "Ezra" is described as dictating 94 books: 24 are the books of the Hebrew Bible, and the other 70 are esoteric.

The Apocalypse of Baruch. The *Apocalypse of Baruch* was written about the same time as II (IV) Esdras, and the less profound *Apocalypse* probably depends much upon II Esdras. The *Apocalypse of Baruch* survives only in a Syriac version translated from Greek; originally the book was composed in Hebrew or Aramaic and is ascribed to Baruch, the disciple of Jeremiah and a contemporary of the destruction of the First Temple. If II Esdras asks questions about important problems of human history and the tragic situation of Israel after the destruction of the Second Temple, the *Apocalypse of Baruch* apparently was written to give a positive, traditional answer to these doubts.

Pseudepigrapha connected with the Dead Sea Scrolls. There are three Pseudepigrapha that are closely connected with the writings of the Dead Sea sect: the *Book of Jubilees*, the Ethiopic *Book of Enoch*, and the *Testaments of the Twelve Patriarchs*. It is not accidental that fragments of the two first books and of two sources of the third were found among the Dead Sea Scrolls.

The Book of Jubilees. From the fragments of the *Book of Jubilees* among the Dead Sea Scrolls, scholars note that the book was originally written in biblical Hebrew. The whole book is preserved in an Ethiopic version translated from Greek.

The book is written in the form of a revealed history of Israel from the creation until the dwelling of Moses on Mt. Sinai, where the content of the book was revealed to Moses by "the angel of the presence." The *Book of Jubilees* in fact is a legendary rewriting of the book of Genesis and a part of Exodus. One of the main purposes of the author is to promote, in the form of divine revelation, a special sectarian interpretation of Jewish law. All the legal prescriptions noted in the book were practiced by the Dead Sea sect; in connection with the solar calendar of 52 weeks, one of the Dead Sea Scrolls even mentions the *Book of Jubilees* as the source. The (unpublished) *Temple Scroll*, a book of sectarian prescriptions that paraphrases—also as divine revelation—a part of the Mosaic Law and was composed by the Dead Sea sect before 100 BCE (*i.e.*, in the same period as the *Book of Jubilees*), closely resembles some parts of the *Book of Jubilees*. Thus, the *Book of Jubilees* could be accepted by the Dead Sea sect and apparently was written in the same circles, immediately before the sect itself came into existence. The apocalyptical hopes expressed in the book are also identical to those of the Dead Sea sect.

The Book of Enoch. Another book that was written during the period of the apocalyptic movement in which the Dead Sea sect came into existence is the *Book of Enoch*, or *I Enoch*. It was completely preserved in an Ethiopic translation from Greek, and large parts from the beginning and end of the Greek version have been published from two papyri. Aramaic fragments of many parts of the book were found among the Dead Sea Scrolls, as were Hebrew fragments of the *Book of Noah*, either one of the sources of *Enoch* or a parallel elaboration of the same material. Passages of the *Book of Noah* were included in *Enoch* by its redactor (editor). Scholars generally agree that the somewhat haphazard redaction of the book was made in its Greek stage, when a redactor put together various treatises of the Enochic literature that were written at various times and reflected various trends of the movement.

The *Book of Noah*

Besides the passages from the *Book of Noah*, five treatises are included in the *Book of Enoch*. The hero of all of them is the biblical Enoch. The first treatise (chapters 1–36) speaks about the fall of the angels, who rebelled before the Flood, and describes Enoch's celestial journeys, in which divine secrets were revealed to him. It was probably written in the late 2nd century BCE.

The second part of the *Book of Enoch* is the "Parables" (or Similitudes) of Enoch (37–71). These three eschatological sermons of Enoch refer to visions; their original language was probably Hebrew rather than Aramaic. This treatise is an important witness for the belief in the coming of the Son of man, who is expressly identified with the Messiah; in chapters 70–71, which are probably a later addition, the Son of man is identified with Enoch himself. The treatise probably dates from the 1st century BCE.

As Aramaic fragments from the Dead Sea Scrolls show, the astronomical book entitled "The Book of the Heavenly Luminaries" (chapters 72–82) is in the present form abbreviated in the *Book of Enoch*. All these astronomical mysteries were shown to Enoch by the angel Uriel. The treatise propagates the same solar calendar that is also known from the *Book of Jubilees* and from the Dead Sea sect. This treatise was probably written before the year 100 BCE.

The fourth treatise (chapters 83–90) contains two visions of Enoch: the first (chapters 83–84), about the Flood, is in reality only a sort of introduction to the second one ("the vision of seventy shepherds"), which describes the history of the world from Adam to the messianic age; the personages of the visions are allegorically described as various kinds of animals. The symbolic description of history continues to the time of Judas Maccabeus; then follows the last assault of Gentiles and the messianic period. Thus, the treatise was written in the early Hasmonean period, some time after the biblical Book of Daniel.

The fifth treatise (chapters 91–107) contains Enoch's speech of moral admonition to his family. The moral stress and the social impact is similar to parts of Jesus' teaching; even the form of beatitudes (blessings) and woes is present. The treatise shows some affinities to the Dead Sea Scrolls, but the author was not a member of the Dead Sea sect; he opposes the central teaching of the sect, the doctrine of predestination (98: 4–5). The treatise apparently was written at the end of the 1st century BCE. Chapter 105, lacking in the Greek version, is a late interpolation, probably of Christian origin.

The author of the treatise himself apparently incorporated into it a small apocalypse, the "Apocalypse of Weeks" (93:1–10; 91:12–17); in it the whole of human history is divided into ten weeks; seven of them belong to the past and the last three to the future.

Testaments of the Twelve Patriarchs. The third pseudepigraphon that shows important affinities with the Dead Sea sect is the *Testaments of the Twelve Patriarchs*, the last speeches of the 12 sons of the Hebrew patriarch Jacob. In its extant form, containing Christian passages, the book was written in Greek. Fragments of two original Semitic sources of the book were found among the Dead Sea Scrolls: the Aramaic "Testament of Levi" (fragments of it were also discovered in Aramaic in the medieval Geniza, or synagogue storeroom, in Cairo) and a Hebrew fragment of the "Testament of Naphtali." A Hebrew "Testament of Judah," which was used both by the *Book of Jubilees* and the *Testaments of the Twelve Patriarchs* in their description of the wars of the sons of Jacob, also probably existed.

Whether Hebrew and Aramaic prototypes for all the 12 testaments of the patriarchs existed is difficult to ascertain. The present book was originally written in Greek. In it each of the sons of Jacob before his death gives moral advice to his descendants, based upon his own experience. All the testaments, with the exception of Gad, also contain apocalyptic predictions.

Between the *Testaments of the Twelve Patriarchs* and the Dead Sea sect there is a historical and ideological connection. The sources of the book were found among the scrolls, the source of the "Testament of Levi" is quoted in a sectarian writing (the Damascus Document), a dualistic outlook is common to the book and the sect, and the devil is named Belial in both. There are, however, important differences: in regard to the nature of the dualism between good and evil, there is in the *Testaments* the concept of the good and bad inclination, known from rabbinical literature, which does not exist in the scrolls;

Connection between the *Testaments of the Twelve Patriarchs* and the Dead Sea sect

though the sect believed in an afterlife of souls, the *Testaments* reflect the belief in the resurrection of the body; there are no traces of the doctrine of predestination in the testaments, a doctrine that is so important for the sect. Only the "Testament of Asher" preaches, as did the Dead Sea sect, hatred against sinners; the other testaments stress, as does rabbinic literature and especially Jesus, the precept of love for God and neighbour. Thus, it is probable that the testaments of the patriarchs were composed in circles in which doctrines of the Dead Sea sect were mitigated and combined with some rabbinic doctrines. A similar humanistic position, founded both on doctrines of the Dead Sea sect and of the Pharisees, is typical of Jesus' message, and there are important parallels between his message and the *Testaments of the Twelve Patriarchs.*

QUMRAN LITERATURE (DEAD SEA SCROLLS)

New literary documents from the intertestamental period were found in the caves of Qumrān in the vicinity of the Dead Sea in the 1940s, but only a portion of them has yet been published. All the Dead Sea Scrolls were written before the destruction of the Second Temple; with the exception of small Greek fragments, they are all in Hebrew and Aramaic. The scrolls formed the library of an ancient Jewish sect, which probably came into existence at the end of the 2nd century BCE and was founded by a religious genius, called in the scrolls the Teacher of Righteousness. Scholars have tried to identify the sect with all possible groups of ancient Judaism, including the Zealots and early Christians, but it is now most often identified with the Essenes; all that the sectarian scrolls contain fits previous information about the Essenes, and the Dead Sea Scrolls help scholars to interpret the descriptions about the Essenes in ancient sources.

Apocryphal and pseudepigraphal writings. The importance of the discovery is very great; the scrolls of books of the Old Testament caused a new evaluation of the history of the text of the Hebrew Bible; fragments of the Apocrypha (Sirach and Tobit) and of already known and unknown Pseudepigrapha enlarge knowledge about Jewish literature of the intertestamental period, and the properly sectarian scrolls are important witnesses about an ancient sect that influenced, in some points, the origins of Christianity.

Among the previously unknown Pseudepigrapha were large parts of an Aramaic scroll, the *Genesis Apocryphon*, which retells stories from Genesis in the manner of a number of apocryphal books. The chapters that are preserved are concerned with Lamech, his grandfather Enoch, Noah, and Abraham, and the narrators in the scroll are the respective biblical heroes. There is a close affinity between this scroll and the *Book of Jubilees* and *Book of Enoch*, fragments of these books having been also found among the Dead Sea Scrolls. Another pseudepigraphon that resembles the Dead Sea sect in spirit is the *Testaments of the Twelve Patriarchs;* fragments of two of its sources, namely, the Aramaic "Testament of Levi" and a Hebrew "Testament of Naphtali," are extant in the Qumrān library. All these books were composed in an apocalyptic movement in Judaism, in the midst of which the Dead Sea sect originated. It is sometimes difficult to ascertain if a work was written within the sect itself or if it represents the broader movement. The largest scroll, the *Temple Scroll*, is as yet unpublished. It describes—by the mouth of God himself and in Hebrew—not the Temple of the last days but the Temple as it should have been built. There are strong ties between the *Temple Scroll* and the *Book of Jubilees* and the prescriptions in it fit the conceptions of the sect; the work was composed by the sectarians themselves.

The *Temple Scroll*

Pesharim. An important source of knowledge about the history of the Dead Sea sect is the *pesharim* ("commentaries"; singular *pesher*). The sectarian authors commented on the books of Old Testament prophets and the book of Psalms and in the commentaries explained the biblical text as speaking about the history of the sect and of events that happened in the time of its existence. According to the manner of apocalyptic literature in the *pesharim*, persons and groups are not named with their proper names but are described by symbolic titles—*e.g.*, the Teacher of Righteousness for the founder of the sect. The most important sectarian commentaries are the *pesharim* on Habakkuk and on Nahum.

The War of the Sons of Light Against the Sons of Darkness. One of the most interesting Dead Sea Scrolls is *The War of the Sons of Light Against the Sons of Darkness*, a description of the eschatological war between the Sons of Light—*i.e.*, the sect—and the rest of mankind, first with the other Jews and then with the Gentiles. At the end the Sons of Light will conquer the whole world, and in this war they will be helped by heavenly hosts; the Sons of Darkness, aided by the devil Belial and his demonic army, and, finally, all wicked ones will be destroyed. The work contains prayers and speeches that will be uttered in the eschatological war as well as military and other ordinances. Thus, the book also could be called the *Manual of Discipline* for the last war.

Books of ordinances. Other books of ordinances of the sect have been preserved, containing prescriptions and other material. Three such compositions are written on one scroll: the *Manual of Discipline*, the *Rule of the Congregation*, and the manual of *Benedictions*. The *Manual of Discipline* is the rule (or statement of regulations) of the Essene community; the most important part of this work is a treatise about the special theology of the sect. The *Rule of the Congregation* contains prescriptions for the eschatological future when the sect is expected to be the elite of the nation. The manual of *Benedictions*, preserved only in a fragmentary state, contains benedictions that are to be said in the eschatological future.

Another sectarian book of ordinances is the Damascus Document (the Zadokite Fragments). The work was already known from two medieval copies before the discovery of the Dead Sea Scrolls, but fragments of it also were found in Qumrān, and the connection between this work and the Dead Sea sect is evident. The Damascus Document was written in a community in Damascus, which was not as rigidly organized as the Essenes. The work contains the rules of this community and reminiscences of the sect's history. Some scholars think that "Damascus" is only a symbolical name for Qumrān.

Hodayot. One of the most important Essene works is the *Hodayot* ("Praises")—a modern Hebrew name for the *Thanksgiving Psalms*. This scroll contains sectarian hymns of praise to God. In its view of the fleshly nature of man, who can be justified only by God's undeserved grace, it resembles St. Paul's approach to the same problem. Some scholars think that the work, or a part of it, was written by the Teacher of Righteousness.

The *Thanksgiving Psalms*

Among other fragments of scrolls liturgical texts of prayers were found, as well as fragments of horoscopes written in a cryptic script. (D.Fl.)

VI. New Testament canon, texts, and versions

THE NEW TESTAMENT CANON

Conditions aiding the formation of the canon. The New Testament consists of 27 books, which are the residue, or precipitate, out of many 1st–2nd-century-AD writings that Christian groups considered sacred. In these various writings the early church transmitted its traditions: its experience, understanding, and interpretation of Jesus as the Christ and the self-understanding of the church. In a seemingly circuitous interplay between the historical and theological processes, the church selected these 27 writings as normative for its life and teachings—*i.e.*, as its canon (from the Greek *kanōn*, literally, a reed or cane used as a measuring rod and, figuratively, a rule or standard). Other accounts, letters, and revelations—*e.g.*, the *Didachē* (Teaching of the Twelve Apostles), *Gospel of Peter*, *First Letter of Clement*, *Letter of Barnabas*, *Apocalypse* (Revelation) *of Peter*, *Shepherd of Hermas*—exist, but through a complex process the canon was fixed for both the Eastern and Western churches in the 4th century. The canon contained four Gospels (Matthew, Mark, Luke, and John), Acts, 21 letters, and one book of a strictly revelatory character, Revelation. These were not necessarily the oldest writings, not all equally revelatory, and not all directed to the church at large.

Significance of the Old Testament

The Old Testament in its Greek translation, the Septuagint (LXX), was the Bible of the earliest Christians. The New Covenant, or Testament, was viewed as the fulfillment of the Old Testament promises of salvation that were continued for the new Israel, the church, through the Holy Spirit, which had come through Christ, upon the whole people of God. Thus, the Spirit, which in the Old Testament had been viewed as resting only on special charismatic figures, in the New Testament became "democratized"—*i.e.*, was given to the whole people of the New Covenant. In postbiblical Judaism of the first Christian centuries, it was believed that the Spirit had ceased after the writing of the Book of Malachi (the last book of the Old Testament canon) and that no longer could anyone say "Thus saith the Lord," as had the prophets, nor could any further holy writ be produced.

The descent of the Spirit on the community of the Messiah (*i.e.*, the Christ) was thus perceived by Christians as a sign of the beginning of the age to come, and the church understood itself as having access to that inspiration through the Spirit. Having this understanding of itself, the church created the New Testament canon not only as a continuation and fulfillment of the Old Testament but also as qualitatively different, because a new age had been ushered in. These 27 books, therefore, were not merely appended to the traditional Jewish threefold division of the Old Testament—the Law (Torah), the Prophets (Nevi'im), and the Writings (Ketuvim)—but rather became the New Testament, the second part of the Christian Bible, of which the Old Testament is the first.

Because of a belief that something almost magical occurs—with an element of secrecy—when a transmitted oral tradition is put into writing, there was, in both the Old and New Testaments, an expression of reluctance about committing sacred material to writing. When such sacred writings are studied to find the revealed word of God, a settled delimiting of the writings—*i.e.*, a canon—must be selected. In the last decade of the 1st century, the Synod of Jamnia (Jabneh), in Palestine, fixed the canon of the Bible for Judaism, which, following a long period of flux and fluidity and controversy about certain of its books, Christians came to call the Old Testament. A possible factor in the timing of this Jewish canon was a situation of crisis: the fall of Jerusalem and reaction to the fact that the Septuagint was used by Christians and to their advantage, as in the translation of the Hebrew word *'alma* ("young woman") in chapter 7, verse 14, of Isaiah—"Behold, a young woman shall conceive and bear a son, and shall call his name Immanuel"—into the Greek term *parthenos* ("virgin").

As far as the New Testament is concerned, there could be no Bible without a church that created it; yet conversely, having been nurtured by the content of the writings themselves, the church selected the canon. The concept of inspiration was not decisive in the matter of demarcation because the church understood itself as having access to inspiration through the guidance of the Spirit. Indeed, until *c.* AD 150, Christians could produce writings either anonymously or pseudonymously—*i.e.*, using the name of some acknowledged important biblical or apostolic figure. The practice was not believed to be either a trick or fraud. Apart from letters in which the person of the writer was clearly attested—as in those of Paul, which have distinctive historical, theological, and stylistic traits peculiar to Paul—the other writings placed their emphases on the message or revelation conveyed, and the author was considered to be only an instrument or witness to the Holy Spirit or the Lord. When the message was committed to writing, the instrument was considered irrelevant, because the true author was believed to be the Spirit. By the mid-2nd century, however, with the delay of the final coming (the Parousia) of the Messiah as the victorious eschatological (end-time) judge and with a resulting increased awareness of history, increasingly a distinction was made between the apostolic time and the present. There also was a gradual cessation of "authentically pseudonymous" writings in which the author could identify with Christ and the Apostles and thereby gain ecclesiastical recognition.

Acceptance of anonymous or pseudonymous writings

The process of canonization. The process of canonization was relatively long and remarkably flexible and detached; various books in use were recognized as inspired, but the Church Fathers noted, without embarrassment or criticism, how some held certain books to be canonical and others did not. Emerging Christianity assumed that through the Spirit the selection of canonical books was "certain" enough for the needs of the church. Inspiration, it is to be stressed, was neither a divisive nor a decisive criterion. Only when the canon had become self-evident was it argued that inspiration and canonicity coincided, and this coincidence became the presupposition of Protestant orthodoxy (*e.g.*, the authority of the Bible through the inspiration of the Holy Spirit).

The need for consolidation and delimitation. Viewed both phenomenologically and practically, the canon had to be consolidated and delimited. Seen historically, however, there were a number of reasons that forced the issue of limiting the canon. Oral tradition had begun to deteriorate in post-apostolic times, partly because many or most of the eyewitnesses to the earliest events of Jesus' life and death and the beginning of the church had died. Also, the oral tradition may simply have suffered in transmission. Papias (died *c.* 130), a bishop of Hieropolis, in Asia Minor, was criticized by Irenaeus (died *c.* 200), a bishop of Lugdunum (now Lyons, France) as not having truly been an eyewitness of the Apostles and of the Lord, whose words he claimed to transmit. Papias had said, "For I did not suppose that the things from the books would aid me so much as the things from the living and continuing voice." Eusebius (*c.* 260–*c.* 340), a church historian, reported these comments in his *Ecclesiastical History* and pointed out inconsistencies in Papias' recollections, doubted his understanding, and called him "a man of exceedingly small intelligence." Large sections of oral tradition, however, which were probably translated in part from Aramaic before being written down in Greek—such as the Passion (suffering of Christ) narrative, many sayings of Jesus, and early liturgical material—benefitted by the very conservativism implicit in such traditions. But because the church perceived its risen Lord as a living Lord, even his words could be adjusted or adapted to fit specific church needs. Toward the end of the 1st century, there was also a conscious production of gospels. Some gospels purported to be words of the risen Lord that did not reflect apostolic traditions and even claimed superiority over them. Such claims were deemed heretical and helped to push the early church toward canonization.

Faced with heresy and claims to late revelations, the early church was constrained to retain the historical dimension of its faith, the *ephapax*, or the "once for all," revelation of God in Jesus Christ.

Impulse toward canonization from heretical movements. Gnosticism (a religious system with influence both on Judaism and Christianity) tended to foster speculation, cutting loose from historical revelation. In defense the orthodox churches stressed the apostolic tradition by focussing on Gospels and letters from apostolic lives and distinguished them from Gnostic writings, such as the *Gospel of Truth* (mentioned by Irenaeus) and now found in Coptic translation in a collection of Gnostic writings from Egypt; it is a Coptic manuscript of a Valentinian Gnostic speculation from the mid-2nd century—*i.e.*, a work based on the teachings of Valentinus, a Gnostic teacher from Alexandria. In the same collection is the *Gospel of Thomas* in Coptic, actually a collection of sayings purporting to be the words of the risen Christ, the living Lord. This "gospel" also occurred in Greek (*c.* 140), and warnings against it as heretical were made by the Church Fathers in the 2nd to the 4th centuries.

Gnosticism

In a general prophetic apocalyptic mood, another heresy, Montanism, arose. This was an ecstatic enthusiastic movement claiming special revelation and stressing "the age of the spirit." Montanus (died *c.* 175) and two prophetesses claimed that their oracular statements contained new and contemporary authoritative revelations. This break with the apostolic time caused vigorous response. An anti-Montanist reported that "the false prophet is

one who speaks in ecstasy after which follow freedom . . . and madness of soul."

The single most decisive factor in the process of canonization was the influence of Marcion (flourished *c.* 140), who had Gnostic tendencies and who set up a "canon" that totally repudiated the Old Testament and anything Jewish. He viewed the Creator God of the Old Testament as a cruel God of retribution and the Jewish Law. His canon consisted of *The Gospel*, a "cleaned up" Luke (the least Jewish), and the *Apostolikon*, (ten Pauline letters with Old Testament references and analogies edited out, without Hebrews, I and II Timothy, and Titus). This restrictive canon acted as a catalyst to the formation of a canon more in line with the thought of the church catholic (universal).

Late-2nd-century canons. By the end of the 2nd century, Irenaeus used the four canonical Gospels, 13 letters of Paul, I Peter, I and II John, Revelation, *Shepherd of Hermas* (a work later excluded from the canon), and Acts. Justin Martyr (died *c.* 165), a Christian apologist, wrote of the reading of the Gospels, "the memoirs of the Apostles," in the services, in which they were the basis for sermons. In his writings he quoted freely from the Gospels, Hebrews, the Pauline Letters, I Peter, and Acts.

The influence of Tatian's *Diatessaron*

Justin's Syrian pupil, Tatian (*c.* 160), although he quotes from John separately, is best known for his *Diatessaron* (literally, "through four" [gospels], but also a musicological term meaning "choral" "harmony"), which was a life of Christ compiled from all four Gospels but based on the outline and structure of John. This indicates both that Tatian was aware of four gospel traditions and that their canonicity was not fixed in final form at his time in Syria. Although Tatian was later declared a heretic, the *Diatessaron* was used until the 5th century and influenced the Western Church even after four separated gospels were established.

The first clear witness to a catalog of authoritative New Testament writings is found in the so-called Muratorian Canon, a crude and uncultured Latin 8th-century manuscript translated from a Greek list written in Rome *c.* 170–180, named for its modern discoverer and publisher Lodovica Antonio Muratori (1672–1750). Though the first lines are lost, Luke is referred to as "the third book of the Gospel," and the canon thus contains [Matthew, Mark] Luke, John, Acts, 13 Pauline letters, Jude, two letters of John, and Revelation. Concerning the *Apocalypse of Peter*, it notes that it may be read, although some persons object; it rejects the *Shepherd of Hermas* as having been written only recently in Rome and lacking connection with the apostolic age. The Wisdom of Solomon (a Jewish intertestamental writing), is included in the accepted works as written in Solomon's honour.

Criteria of canonicity

Some principles for determining the criteria of canonicity begin to be apparent: apostolicity, true doctrine (*regula fidei*), and widespread geographical usage. Such principles are indicated by Muratori's argument that the Pauline Letters are canonical and universal—the Word of God for the whole church—although they are addressed to specific churches, on the analogy of the letters to the seven churches in Revelation; in a prophetic statement to the whole church, seven specific churches are addressed, then the specific letters of Paul can be read for all. Thus, the catholic status of the Pauline letters to seven churches is vindicated on the basis of the revelation of Jesus Christ to John, the seer and writer of Revelation. Wide usage in the church is indicated in calling Acts the Acts of *all* the Apostles and in the intention of the "general address"—*e.g.*, "To those who are called," in Jude—of the Catholic (or general) Letters—*i.e.*, I and II Peter, I, II, and III John, James, and Jude. The criterion of accordance with received teaching is plain in the rejection of heretical writings. The Muratorian Canon itself may have been, in part, a response to Marcion's heretical and reductive canon.

The criteria of true doctrine, usage, and apostolicity all taken together must be satisfied, then, in order that a book be judged canonical. Thus, even though the *Shepherd of Hermas*, the *First Letter of Clement*, and the *Didachē* may have been widely used and contain true doctrines, they were not canonical because they were not apostolic nor connected to the apostolic age, or they were local writings without support in many areas.

During the time of the definitive formation of the canon in the 2nd century, apparent differences existed in the Western churches (centred in or in close contact with Rome) and those of the East (as in Alexandria and Asia Minor). It is not surprising that the Roman Muratorian Canon omitted Hebrews and accepted and held Revelation in high esteem, for Hebrews allows for no repentance for the baptized Christian who commits apostasy (rejection of faith), a problem in the Western Church when it was subjected to persecution. In the East, on the other hand, there was a dogmatic resistance to the teaching of a 1,000-year reign of the Messiah before the end time—*i.e.*, chiliasm, or millenarianism—in Revelation. There was also a difference in the acceptance of Acts and the Catholic Letters. With the continued expansion of the church, particularly in the 2nd century, consolidation was necessary.

Canonical standards of the 3rd and 4th centuries. Clement of Alexandria, a theologian who flourished in the late 2nd century, seemed to be practically unconcerned about canonicity. To him, inspiration is what mattered, and he made use of the *Gospel of the Hebrews*, the *Gospel of the Egyptians*, the *Letter of Barnabas*, the *Didachē*, and other extracanonical works. Origen (died *c.* 254), Clement's pupil and one of the greatest thinkers of the early church, distinguished at least three classes of writings, basing his judgment on majority usage in places that he had visited: (1) *homologoumena* or *anantirrhēta*, "undisputed in the churches of God throughout the whole world" (the four Gospels, 13 Pauline Letters, I Peter, I John, Acts, and Revelation); (2) *amphiballomena*, "disputed" (II Peter, II and III John, Hebrews, James, and Jude); and (3) *notha*, "spurious" (*Gospel of the Egyptians*, *Thomas*, and others). He used the term "scripture" (*graphē*) for the *Didachē*, the *Letter of Barnabas*, and the *Shepherd of Hermas*, but did not consider them canonical.

Standards of Origen and Eusebius

Eusebius shows the situation in the early 4th century. Universally accepted are: the four Gospels, Acts, 14 Pauline letters (including Hebrews), I John, and I Peter. The disputed writings are of two kinds: (1) those known and accepted by many (James, Jude, II Peter, II and III John, and (2) those called "spurious" but not "foul and impious" (*Acts of Paul*, *Shepherd of Hermas*, *Apocalypse of Peter*, *Letter of Barnabas*, *Didachē* and possibly the *Gospel of the Hebrews*); finally there are the heretically spurious (*e.g.*, *Gospel of Peter*, *Acts of John*). Revelation is listed both as fully accepted ("if permissible") and as spurious but not impious. It is important that Eusebius feels free to make authoritative use of the disputed writings. Thus canon and authoritative revelation are not yet the same thing.

Determination of the canon in the 4th century. Athanasius, a 4th-century bishop of Alexandria and a significant theologian, delimited the canon and settled the strife between East and West. On a principle of inclusiveness, both Revelation and Hebrews (as part of the Pauline corpus) were accepted. The 27 books of the New Testament—and they only—were declared canonical. In the Greek churches there was still controversy about Revelation, but in the Latin Church, under the influence of Jerome, Athanasius' decision was accepted. It is notable, however, that, in a mid-4th-century manuscript called Codex Sinaiticus, the *Letter of Barnabas* and the *Shepherd of Hermas* are included at the end but with no indication of secondary status, and that, in the 5th-century Codex Alexandrinus, there is no demarcation between Revelation and *I* and *II Clement*.

The Syriac canon

In the Syriac Church, Tatian's *Diatessaron* was used until the 5th century, and in the 3rd century the 14 Pauline Letters were added. Because Tatian had been declared a heretic, there was a clear episcopal order to have the four separated Gospels when, according to tradition, Rabbula, bishop of Edessa, introduced the Syriac version known as the Peshitta—also adding Acts, James, I Peter, and I John—making a 22-book canon. Only

much later, perhaps in the 7th century, did the Syriac canon come into agreement with the Greek 27 books.

Developments in the 16th century. With the advent of printing and differences between Roman Catholics and Protestants, the canon and its relationship to tradition finally became fixed. During the Counter-Reformation Council of Trent (1545–63), the canon of the entire Bible was set in 1546 as the Vulgate, based on Jerome's Latin version. For Luther, the criterion of what was canonical was both apostolicity, or what is of an apostolic nature, and "was Christum treibet"—what drives toward, or leads to, Christ. This latter criterion he did not find in, for example, Hebrews, James, Jude, and Revelation; even so, he bowed to tradition, and placed these books last in the New Testament.

TEXTS AND VERSIONS

Textual criticism. *The physical aspects of New Testament texts.* To establish the reliability of the text of ancient manuscripts in order to reach the text that the author originally wrote (or, rather, dictated) involves the physical aspects of the texts: collection, collation of differences or variant readings in manuscripts, and comparison in matters of dating, geographical origins, and the amount of editing or revision noted, using as many copies as are available. Textual criticism starts thus with the manuscripts themselves. Families of manuscripts may be recognized by noting similarities and differences, degrees of dependence, or stages of their transmission leading back to the earliest text, or autograph. The techniques used in textual studies of ancient manuscripts are the same whether they deal with secular, philosophical, or religious texts. New Testament textual criticism, however, operates under unique conditions because of an abundance of manuscripts and the rather short gap between the time of original writing and the extant manuscripts, shorter than that of the Old Testament.

Significance of the large number of New Testament manuscripts

Compared with other ancient manuscripts, the text of the New Testament is dependable and consistent, but on an absolute scale there are far more variant readings as compared with those of, for example, classical Greek authors. This is the result, on the one hand, of a great number of surviving manuscripts and extant manuscript fragments and, on the other, of the fact that the time gap between an oral phase of transmission and the written stage was far shorter than that of many other ancient Greek manuscripts. The missionary message—the kerygma (proclamation)—with reports of the Passion, death, Resurrection, and Ascension of Jesus Christ and collections of his deeds and sayings was, at first, oral tradition. Later it was written down in Gospel form. The letters of Paul, Apostle to the Gentiles who founded or corresponded with churches, were also collected and distributed as he had dictated them. All autographs of New Testament books have disappeared. In sharp contrast to the fact that the oldest extant full manuscript of a work by the Greek philosopher Plato (died 347 BC) is a copy written in 895—a gap of more than 1,000 years bridged by only a few papyrus texts—there was a time gap of less than 200 or 300 years between the original accounts of the New Testament events and extant manuscripts. In fact, a small (about 2.5 inches by 3.5 inches [6.4 by 8.9 centimetres]) papyrus fragment with verses from the 18th chapter of the Gospel According to John can be dated *c.* 120–130; this earliest known fragment of the New Testament was written 40 years or less after the presumed date of the production of that Gospel (*c.* 90).

Excluding papyri found preserved in the dry sands, as in Egypt (where the Gospel According to John was evidently popular judging from the large number of fragments found there), the approximate number of New Testament manuscripts dating from the 3rd to 18th centuries are: 2,000 of the four Gospels; 400 of Acts, Pauline, and Catholic letters together; 300 of Pauline letters alone; 250 of Revelation; and 2,000 lectionaries—*i.e.*, collections of gospel (and sometimes Acts and letter) selections, or pericopes, meant to be used in public worship. Quotations from the Church Fathers—some of which are so extensive as to include almost the whole New Testament—account for more than 150,000 textual variants. Of the quotations in the Fathers, however, it is difficult to make judgments because the quotations may have been intended to be exact from some particular text traditions, but others may have been from memory, conflations, harmonizations, or allusions. Of the many New Testament manuscripts to date, however, only about 50 contain the entire 27 books of the New Testament. The majority have the four Gospels, and Revelation is the least well attested. Prior to the printing press (15th century), all copies of Bibles show textual variations.

Types of writing materials and methods. In Hellenistic times (*c.* 300 BC–*c.* AD 300), official records were often inscribed on stone or metal tablets. Literary works and detailed letters were written on parchment or papyrus, though short or temporary records were written or scratched on potsherds (ostraca) or wax tablets. Scrolls were made by gluing together papyrus sheets (made from the pith of the papyrus reed) or by sewing together parchment leaves (made from treated and scraped animal skins); they were written in columns and read by shifting the roll backward and forward from some wooden support on one or both ends. Such scrolls were used for literary or religious works and seldom exceeded 30 feet (nine metres) in length because of their weight and awkwardness in handling.

Scrolls and codices

In contrast, the church used not scrolls but the codex (book) form for its literature. A codex was formed by sewing pages of papyrus or parchment of equal size one upon another and vertically down the middle, forming a quire; both sides of the pages thus formed could be written upon. In antiquity, the codex was the less honourable form of writing material, used for notes and casual records. The use of the book form testifies to the low cultural and educational status of early Christianity—and, as the church rose to prominence, it brought "the book" with it. Not until the time of the Roman emperor Constantine in the 4th century, when Christianity became a state religion, were there parchment codices containing the whole New Testament.

Some very early New Testament manuscripts and fragments thereof are papyrus, but parchment, when available, became the best writing material until the advent of printing. The majority of New Testament manuscripts from the 4th to 15th centuries are parchment codices. When parchment codices occasionally were deemed no longer of use, the writing was scraped off and a new text written upon it. Such a rewritten (*rescriptus*) manuscript is called a palimpsest (from the Greek *palin*, "again," and *psaō*, "I scrape"). Often the original text of a palimpsest can be discerned by photographic process.

In New Testament times there were two main types of Greek writing: majuscules (or uncials) and minuscules. Majuscules are all capital (uppercase) letters, and the word uncial (literally, $\frac{1}{12}$ of a whole, about an inch) points to the size of their letters. Minuscules are lowercase manuscripts. Both uncials and minuscules might have ligatures making them into semi-connected cursives. In Greco-Roman times minuscules were used for the usual daily writing. In parchments from the 4th to the 9th centuries, both majuscules and minuscules were used for New Testament manuscripts, but by the 11th century all the manuscripts were minuscules.

Problems of the continuous flow of letters

In these early New Testament manuscripts, there were no spaces between either letters or words, rarely an indication that a word was "hyphenated," no chapter or verse divisions, no punctuation, and no accents or breathing marks on the Greek words. There was only a continuous flow of letters. In addition, there were numerous (and sometimes variable) abbreviations marked only by a line above (*e.g.*, $\overline{\text{IC}}$ for IHCOUC, or Jesus, and $\overline{\text{KC}}$ for *kyrios*, or Lord. Not until the 8th–9th century was there any indication of accents or breathing marks (both of which may make a difference in the meaning of some words); punctuation occurred sporadically at this period; but not until the Middle Ages were the texts supplied with such helps as chapters (*c.* 1200) and verses (*c.* 1550).

Occasionally, the parchment was stained (*e.g.*, purple), and the ink was silver (*e.g.*, Codex Argenteus, a 5th–6th-

century Gothic translation. Initial letters were sometimes illuminated, often with red ink (from which comes the present English word rubric, based on the Latin for "red," namely *ruber*).

Types of manuscript errors. Since scribes either copied manuscripts or wrote from dictation, manuscript variants could be of several types: copying, hearing, accidental, or intentional. Errors in copying were common, particularly with uncial letters that looked alike. In early manuscripts OC (for *hos*, "[he] who"), for example, might easily be mistaken for the traditional abbreviation of God: $\overline{\Theta C}$ (for ΘEOC, *theos*). Dittography (the picking up of a word or group of words and repeating it) and haplography (the omission of syllables, words, or lines) are errors most apt to occur where there are similar words or syllables involved. In chapter 17, verse 15, of John, in one manuscript the following error occurs: "I do not pray that thou shouldest take *them from the* [world, but that thou shouldst keep *them from the*] evil one" becomes "I do not pray that thou shouldst take them from the evil one." This is obviously a reading that omitted the words between two identical ends of lines—*i.e.*, an error due to *homoioteleuton* (similar ending of lines).

Especially in uncial manuscripts with continuous writing, there is a problem of word division. An English example may serve to illustrate: GODISNOWHERE may be read "God is now here" or "God is nowhere." Internal evidence from the context can usually solve such problems. Corrections of a manuscript either above the line of writing or in the margin (and also marginal comments) may be read and copied into the text and become part of it as a gloss.

Errors of hearing are particularly common when words have the same pronunciation as others but differ in spelling (as in English: "their, there"; "meet, meat"). This kind of error increased in frequency in the early Christian Era because some vowels and diphthongs lost their distinctive sound and came to be pronounced alike. For example, the Greek vowels *ē*, *i*, and *u* and the diphthongs *ei*, *oi*, and *ui* all sounded like the ēē (as in "feet"). Remarkable mistransiations can occur as, for example, in I Corinthians, chapter 15, verse 54: "Death is swallowed up in victory"—becomes by itacism (pronunciation of the Greek letter ē) "Death is swallowed up in conflict" (*neikos*). Another problem of itacism is the distinction between declensions of the 1st and 2nd persons in the plural ("we" and "you") in Greek, which can sound the same (*hemeis*, "we"; *humeis*, "you"), because the initial vowels are not clearly differentiated. Such errors can cause interpretative difficulties.

A different category of error occurs in dictation or copying, when sequences of words, syllables, or letters in a word are mixed up, synonyms substituted in familiar passages, words read across a two- (or more) column manuscript instead of down, or assimilated to a parallel. Intentional changes might involve corrections of spelling or grammar, harmonizations, or even doctrinal emendations, and might be passed on from manuscript to manuscript. Paleographers—*i.e.*, scientists of ancient writing—can note changes of hands in manuscript copying or the addition of new hands such as those of correctors of a later date.

Significance and reliability of paleography

Paleography, a science of dating manuscripts by typological analysis of their scripts, is the most precise and objective means known for determining the age of a manuscript. Script groups belong typologically to their generation; and changes can be noted with great accuracy over relatively short periods of time. Dating of manuscript material by a radioactive-carbon test requires that a small part of the material be destroyed in the process; it is less accurate than dating from paleography.

Attempts to approximate an original manuscript and critical scholarship. Textual criticism of the Greek New Testament attempts to come as near as possible to the original manuscripts (which did not survive), based on reconstructions from extant manuscripts of various ages and locales. Assessment of the individual manuscripts and their relationships to each other can produce a fairly reliable text from various readings that may have been the result of copying and recopying of manuscripts. It is not always age that matters. Older manuscripts may be corrupt, and a reading in a later manuscript may in reality be ancient. No single witness or group of witnesses is reliable in all its readings.

Erasmus' edition

When Erasmus, the Dutch Humanist, prepared the Greek text for the first printed edition (1516) of the New Testament, he depended on a few manuscripts of the type that had dominated the church's manuscripts for centuries and that had had its origin in Constantinople. His edition was produced hastily, he even translated some parts for which he did not have a Greek text from Jerome's Latin text (Vulgate). In 1522 Cardinal Ximines, a Spanish scholarly churchman, published his Complutensian Polyglot at Alcalá (Latin: Complutum), Spain, a Bible in which parallel columns of the Old Testament are printed in Hebrew, the Vulgate, and the Septuagint (LXX), together with the Aramaic Targum (translation or paraphrase) of Onkelos to the Pentateuch with a translation into Latin. The Greek New Testament was volume 5 of this work, and the text tradition behind it cannot be determined with any accuracy. During the next decades new editions of Erasmus' text profited from more and better manuscript evidence and the printer Robert Estienne of Paris produced in 1550 the first text with a critical apparatus (variant readings in various manuscripts). This edition became influential as a chief witness for the *Textus Receptus* (the received standard text) that came to dominate New Testament studies for more than 300 years. This *Textus Receptus* is the basis for all the translations in the churches of the Reformation, including the King James Version.

Large extensive New Testament critical editions prepared by the German scholars C. von Tischendorf (1869–72) and H. von Soden (1902–13) had Sigla (signs) for the various textual witnesses; they are complex to use and different from each other. The current system, a revision by an American scholar, C.R. Gregory (adopted in 1908), though not uncomplicated has made uniform practice possible. A more pragmatic method of designation and rough classification was that of the Swiss scholar J.J. Wettstein's edition (1751–52). His textual apparatus was relatively uncomplicated. He introduced the use of capital Roman, Greek, or Hebrew letters for uncials and Arabic numbers for minuscules. Later, a Gothic P with exponents came into use for papyri and, in the few cases needed, Gothic or Old English O and T with exponents for ostraca and talismans (engraved amulets). Lectionaries are usually designated by an italicized lowercase *l* with exponents in Arabic numbers.

Known ostraca—*i.e.*, broken pieces of pottery (or potsherds) inscribed with ink—contain short portions of six New Testament books and number about 25. About nine talismans date from the 4th to 12th centuries; they are good-luck charms with a few verses on parchment, wood, or papyrus. Four of these contain the Lord's Prayer. These short portions of writing, however, are hardly of significance for a study of the New Testament textual tradition.

Texts and manuscripts. In referring to manuscript text types by their place of origin, one posits the idea that the major centers of Christendom established more or less standard texts: Alexandria; Caesarea and Antioch (Eastern); Italy and Gallia plus Africa (Western); Constantinople, the home for the Byzantine text type or the *Textus Receptus*. While such a geographical scheme has become less accurate or helpful, it still serves as a rough classification of text types.

Characteristics of major codices

Uncials. The main uncials known in the 17th–18th centuries were: A, D, D^p, E^a, and C. A, Codex Alexandrinus, is an early-5th-century manuscript containing most of the New Testament but with lacunae (gaps) in Matthew, John, and II Corinthians, plus the inclusion of the extracanonical I and II Clement. In the Gospels, the text is of the Byzantine type, but, in the rest of the New Testament, it is Alexandrian. In 1627 the A uncial was presented to King Charles I of England by the Patriarch of Constantinople; it has been in the British Museum, in London, since 1751.

D, Codex Bezae Cantabrigiensis, is a 5th-century Greco-Roman bilingual text (with Greek and Latin pages facing each other). D contains most of the four Gospels and Acts and a small part of III John and is thus designated D^{ea} (e, for *evangelia*, or "gospels"; and a for *acta*, or Acts). In Luke, and especially in Acts, D^{ea} has a text that is very different from other witnesses. Codex Bezae has many distinctive longer and shorter readings and seems almost to be a separate edition. Its Acts, for example, is one-tenth longer than usual. D represents the Western text tradition. D^{ea} was acquired by Theodore Beza, a Reformed theologian and classical scholar, in 1562 from a monastery in Lyons (in France). He presented it to the University of Cambridge, England, in 1581 (hence, Beza Cantabrigiensis).

D^{p}, Codex Claramontanus, of the same Western text type although not remarkably dissimilar from other known texts, contains the Pauline Letters including Hebrews. D^{p} (p, for Pauline epistles) is sometimes referred to as D_2. Beza acquired this 6th-century manuscript at about the same time as D^{ea}, but D^{p} was from the Monastery of Clermont at Beauvais (hence, Claramontanus). It is now in the Bibliothèque Nationale, in Paris.

E^{a}, Codex Laudianus, is a bilingual Greco-Latin text of Acts presented in 1636 by Archbishop Laud, an Anglican churchman, to the Bodleian Library at Oxford. It is a late-6th- or early-7th-century manuscript often agreeing with D^{ea} and its Western readings but also having a mixture of text types, often the Byzantine.

C, Codex Ephraemi Syri rescriptus, is a palimpsest. Originally written as a biblical manuscript in the 5th century, it was erased in the 12th century, and the treatises or sermons of Ephraem Syrus, a 4th-century Syrian Church Father, were written over the scraped text. It was found *c.* 1700 by the French preacher and scholar Pierre Allix; and Tischendorf, using chemical reagents, later deciphered the almost 60 percent of the New Testament contained in it, publishing it in 1843. The text had two correctors after the 5th century but is, on the whole, Byzantine and reflects the not too useful common text of the 9th century.

Although there are numerous minuscules (and lectionaries), their significance in having readings going back to the first six centuries AD was not noted until textual criticism had become more refined in later centuries.

The main uncials and some significant minuscules that were discovered and investigated in the 19th century changed the course of the textual criticism and led the way to better manuscript evidence and methods of dealing with it. This has continued into the 20th century. The main new manuscript witnesses are designated ℵ or S, B, W, and Θ.

Newer uncial codices

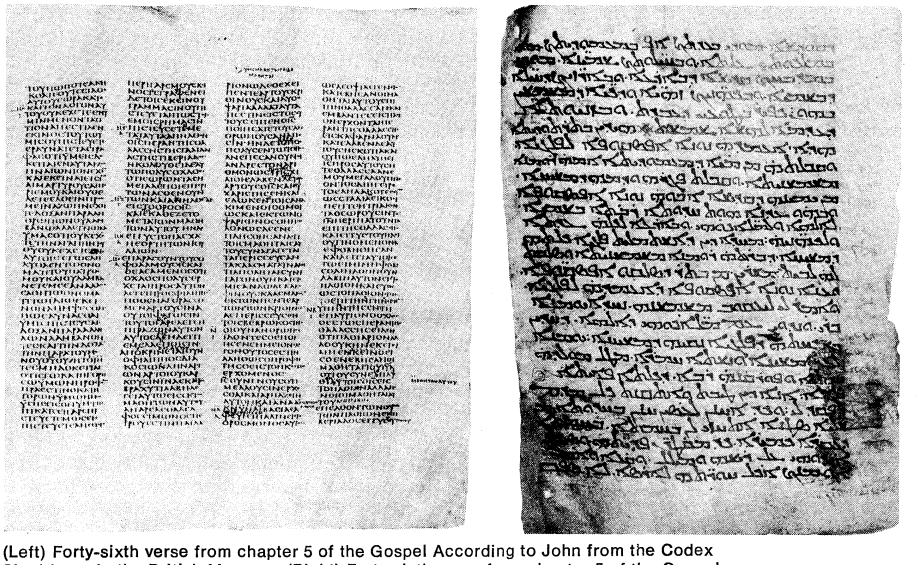
ℵ or S, Codex Sinaiticus, was discovered in 1859 by Tischendorf at the Monastery of St. Catherine at the foot of Mt. Sinai (hence, Sinaiticus) after a partial discovery of 43 leaves of a 4th-century biblical codex there in 1844. Though some of the Old Testament is missing, a whole 4th-century New Testament is preserved, with the *Letter of Barnabas* and most of the *Shepherd of Hermas* at the end. There were probably three hands and several later correctors. Tischendorf convinced the monks that giving the precious manuscript to Czar Alexander II of Russia would grant them needed protection of their abbey and the Greek Church. Tischendorf subsequently published ℵ (S) at Leipzig and then presented it to the Tzar. The manuscript was in Leningrad until 1933, during which time the Oxford University Press in 1911 published a facsimile of the New Testament from photographs of the manuscript taken by Kirsopp Lake, an English biblical scholar. The manuscript was sold in 1933 by the Soviet regime to the British Museum for £100,000. The text type of ℵ is in the Alexandrian group, although it has some Western readings. Later corrections representing attempts to alter the text to a different standard probably were made about the 6th or 7th century at Caesarea.

B, Codex Vaticanus, a biblical manuscript of the mid-4th century in the Vatican Library since before 1475, appeared in photographic facsimile in 1889–90 and 1904. The New Testament lacks Hebrews from chapter 9, verse 14, on, the Pastorals, Philemon, and Revelation. Because B has no ornamentation, some scholars think it slightly older than ℵ. Others, however, believe that both B and ℵ, having predominantly Alexandrian texts, may have been produced at the same time when Constantine ordered 50 copies of the Scriptures. As an early representation of the Alexandrian text, B is invaluable as a most trustworthy ancient Greek text.

W, Codex Washingtonianus (or Freerianus), consists of the four Gospels in the so-called Western order (Matthew, John, Luke, and Mark, as D^{ea}). It was acquired in Egypt by C.L. Freer, an American businessman and philanthropist (hence, the Freer-Gospels), in 1906 and is now in the Freer Gallery of Art of the Smithsonian Institution, in Washington, D.C. Codex Washingtonianus is a 4th–5th-century manuscript probably copied from several different manuscripts or textual families. The Byzantine, Western (similar to Old Latin), Caesarean, and Alexandrian text types are all represented at one point or another. One of the most interesting variant readings is a long ending to the Gospel According to Mark following a reference to the risen Christ (not found in most manuscript traditions).

Θ, Codex Koridethianus, is a 9th-century manuscript

By courtesy of the (left) trustees of the British Museum, (right) Cambridge University Press; photograph, John Ray



(Left) Forty-sixth verse from chapter 5 of the Gospel According to John from the Codex Sinaiticus. In the British Museum. (Right) Forty-sixth verse from chapter 5 of the Gospel According to John from the Codex Palimpsestus Sinaiticus, Old Syriac text.

taking its name from the place of the scribe's monastery, Koridethi, in the Caucasus Mountains, near the Caspian Sea. Θ contains the Gospels; Matthew, Luke, and John have a text similar to most Byzantine manuscripts, but the text of Mark is similar to the type of text that Origen and Eusebius used in the 3rd–4th centuries, a Caesarean type. The manuscript is now in Tiflis, capital city of the Georgian S.S.R.

Minuscules. Although there are many minuscules, most of them come from the 9th century on; a few, however, shed significant light on earlier readings, representing otherwise not well attested texts or textual "families." In the early 20th century, the English scholar Kirsopp Lake (hence, Lake group) discovered a textual family of manuscripts known as Family 1:1, 118, 131, and 209 (from the 12th to 14th centuries) that have a text type similar to that of Θ, a 3rd–4th-century Caesarean type. At the end of the 19th century, W.H. Ferrar, a classical scholar at Dublin University (hence, the Ferrar group), found that manuscripts 13, 69, 124, and 346—and some minuscules discovered later (from the 11th to 15th centuries)—also seemed to be witnesses to the Caesarean text type. Manuscript 33, the "Queen of the Cursives," is a 9th–10th-century manuscript now at the Bibliothèque Nationale, in Paris; it contains the whole New Testament except Revelation and is a reliable witness to the Alexandrian text (similar to B) but, in Acts and the Pauline Letters, shows influence of the Byzantine text type.

Textual families

Lectionaries range from the 5th to the 6th century on; some early ones are uncials, though many are minuscules. Scholarly work with lectionary texts is only at its beginning, but the textual types of lectionaries may preserve a textual tradition that antedates its compilation and serves to give examples of the various text forms.

Papyri. The earliest New Testament manuscript witnesses (2nd–8th centuries) are papyri mainly found preserved in fragments in the dry sands of Egypt. Only in the recent decades of this century have the relatively recently discovered New Testament papyri been published. Of those cataloged to date, there are about 76 New Testament manuscripts with fragments of various parts of the New Testament, more than half of them being from the 2nd to 4th centuries. All the witnesses prior to 400 are of Egyptian provenance, and their primitive text types, though mainly Alexandrian, establish that many text types existed and developed side by side. One of the most significant papyrus finds is p^{52}, from *c.* 130 to 140, the earliest extant manuscript of any part of the New Testament. P^{52} consists of a fragment having on one side John 18:31–33 and on the other John 18:37–38, indicating that it was a codex, of which the text type may be Alexandrian. It is now in the John Rylands Library at Manchester.

The Chester Beatty papyri

In the early 1930s, British mining engineer A. Chester Beatty acquired three 3rd-century papyri from Egypt; they were published in 1934–37. Known as p^{45}, p^{46}, and p^{47}, they are, for the most part, in his private library in Dublin.

P^{45}, Beatty Biblical Papyrus I (and some leaves in Vienna), contains 30 leaves of an early- or mid-3rd-century codex of Matthew, Mark, Luke, John, and Acts. Each Gospel is of a different text type, and, although the leaves are mutilated, the Alexandrian text appears to predominate (particularly in Acts, in which a short non-Western text prevails); the whole may be thought of as pre-Caesarean.

P^{46}, Beatty Biblical Papyrus II (and Papyrus 222 at the University of Michigan), consists of 86 leaves of an early-3rd-century (*c.* 200) codex quire containing the Pauline Letters in the following order: Romans, Hebrews, I and II Corinthians, Ephesians, Galatians, Philippians, Colossians, and I Thessalonians. Although some of the leaves are quite mutilated, the text type of p^{46} appears to be Alexandrian. P^{47}, Beatty Biblical Papyrus III, is from the late 3rd century. It contains Rev. 9:10–17:2. It is the oldest, but not the best, text of Revelation and agrees with A, C, and ℵ.

Other early significant papyri are: p^{66}, p^{48}, p^{72}, p^{75}, and p^{74}. P^{66}, also known as Papyrus Bodmer II, contains in 146 leaves (some of which have lacunae) almost all of the Gospel According to John, including chapter 21. This codex, written before 200, is thus merely one century removed from the time of the autograph, the original text. Its text, like that of p^{45}, is mixed, but it has elements of an early Alexandrian text. P^{66} and the other Bodmer papyri, which Martin Bodmer, a Swiss private collector, acquired from Egypt, were published 1956–61. They are in the private Bodmer library at Cologny, near Geneva. P^{48} is a late-3rd-century text of Acts now in a library in Florence. It contains Acts 23:11–17, 23–29 and illustrates a Greek form of the Western text in Egypt in the 3rd century. The papyri of p^{72} Papyri Bodmer VII and VIII, are also from the 3rd century. VII contains a manuscript of Jude in a mixed text, and VIII contains I and II Peter. In I Peter the Greek was written by a scribe whose native language was Coptic; there are numerous examples of misspellings and itacisms that when corrected leave a text similar to the Alexandrian witnesses.

The papyri of p^{75}, Papyri Bodmer XIV and XV, are 2nd–3rd-century codices containing most of Luke and of John, with John connected to Luke on the same page (unlike the Western order of the Gospels). The text coincides most with B but also has affinities with p^{66} and p^{45} as a predecessor of Alexandrian form.

P^{74}, Bodmer Papyrus XVII, is a 6th–7th-century text of Acts and the Catholic Letters. Acts show affinities with ℵ and A and no parallels with the Western text.

These and other papyri witness to the state of the early text of the New Testament in Egypt, indicating that no *one* text dominated and that text types of different origin flourished side by side.

Versions. *Early versions.* Even with all these witnesses, there remain problems in the Greek text. These include variants about which there is no settled opinion and some few words for which no accurate meaning can be found because they occur only once in the New Testament and not in prior Greek works. Very early translations of the New Testament made as it spread into the non-Greek-speaking regions of the missionary world, the so-called early versions, may provide evidence for otherwise unknown meanings and reflections of early text types.

The Koine, Latin, Syriac, Coptic, and other versions

In the Eastern half of the Mediterranean, Koine (common, vernacular) Greek was understood, but, elsewhere, other languages were used. Where Roman rule dominated, Latin came into use—in North Africa, perhaps in parts of Asia Minor, Gaul, and Spain (*c.* 3rd century). Old Latin versions had many variants, and these translations, traditionally known as the *Itala*, or Old Latin (O.L.), are designated in small letters of the Roman alphabet. The African versions were further from the Greek than were those made in Europe.

In dealing with the New Testament, Jerome prepared a Latin recension of the Gospels using a European form of the Old Latin and some Greek manuscripts. Though the completed Latin translation at the end of the 4th century was produced by no one editor or compiler—a commonly accepted Latin text, the Vulgate, emerged. A reworked official critical edition was a concern of the Council of Trent (1545–63), and in 1592 the Clementine Vulgate, named after Pope Clement VIII, became the authoritative edition. Since Vatican II (1962–65), an ecumenical group of biblical scholars using the best available manuscript witnesses has been engaged in the preparation of a critically sound revision of the Vulgate.

At Edessa (in Syria) and western Mesopotamia neither Latin nor Greek was understood. Therefore, Syriac (a Semitic language related to Aramaic) was used. Old Syriac was probably the original language of the *Diatessaron* (2nd century), but only fragments of Old Syriac manuscripts survive. The Peshitta (common, simple) Syriac (known as syr^{pesh}) became the Syrian 22-book Vulgate of the New Testament, and, at the end of the 4th century, its text was transmitted with great fidelity. The Philoxenian (syr^{phil}) and Harclean (syr^{hare}) versions followed in the 6th–7th centuries and contained all 27 of the New Testament books. The Palestinian (similar to Palestinian Aramaic) Syriac (syr^{pal}) may date to the

5th century but is known chiefly from 11th- to 12th-century lectionaries and is quite independent of other Syriac versions, reflecting a different text type.

In Egypt, in the later Hellenistic period, the New Testament was translated into Coptic—in the south (Upper Egypt) the Sahidic (cop^{sah}), and in the north (Lower Egypt) the Bohairic (cop^{boh}), the two principal dialects. By the 4th century, the Sahidic version was known, and the Bohairic somewhat later. The Coptic versions are fairly literal and reflect a 2nd–3rd-century Alexandrian Greek text type with some Western variants.

A Gothic version was made from the Byzantine text type by a missionary, Ulfilas (late 4th century); an Armenian version (5th century) traditionally was believed to have been made from the Syriac but may have come from a Greek text. Related perhaps to the Armenian was a Georgian version; and an Ethiopic version (*c.* 6th–7th century) was influenced both by Coptic and later Arabic traditions. In the various versions there is evidence of geographical spread, of the history of the underlying text traditions used, and of how they were interpreted in the early centuries.

The many readings in the Greek, Latin, and Syriac Fathers, who can be dated and located, can, to some extent, shed light on the underlying New Testament texts they quoted or used.

Elucidation of meanings of unknown Greek words

Another use both of the versions and patristic quotations is elucidation of the meaning of hitherto unknown Greek words in the New Testament.

An example is *epiousios* in the Lord's Prayer as given in verse 11 of chapter 6 of Matthew and verse 3, chapter 11, of Luke. The traditional translation in the Western Church is "daily" (referring to bread). From the Old Latin, Jerome, the early Syriac versions, and a retroversion of the Lord's Prayer into a proposed Aramaic substratum, the meaning is either "daily" or, more likely, "for the morrow"; and modern translations include this meaning in footnotes, including the suggestion that it may refer to eucharistic bread. The Greek is possibly a coined compound word that, on the basis of its component parts, yields "for the morrow" or "that which is coming soon." Such latter treatment is not conjectural emendation but rather creative analysis in context, where no Greek variants help. The biblical scholar, in possession of many variants, usually uses conjecture only as a means of last resort, and any conjecture must be both intrinsically suitable and account for the reading considered corrupt in the transmitted text.

Later and modern editions. New Testament editions in the 18th century did not question the *Textus Receptus* (*T.R.*), despite new manuscript evidence and study, but its limitations became apparent. E. Wells, a British mathematician and theological writer (1719), was the first to edit a complete New Testament that abandoned the *T.R.* in favour of more ancient manuscripts; and English scholar Richard Bentley (1720) also tried to go back to early manuscripts to restore an ancient text, but their work was ignored. In 1734 J.A. Bengel, a German Lutheran biblical theologian, stressed the idea that not only manuscripts but also families of manuscript traditions must be differentiated, and he initiated the formulation of criteria for text criticism. J.J. Wettstein's edition (1730–51) had a wealth of classical and rabbinic quotations, but his theory on text was better than the text itself. A German Lutheran theologian, J.S. Semler (1767), further refined Bengel's classification of families.

Modern criteria and classifications of texts

J.J. Griesbach (1745–1812), a German scholar and student of Semler, adapted the text-family classification to include Western and Alexandrian text groups that preceded the Constantinopolitan groupings. He cautiously began to alter texts according to increasingly scientific canons of text criticism. These are, with various refinements, still used, as, for example, that "the difficult is to be preferred to the easy reading," and "the shorter is preferable to a longer"—both of which reason (with many other factors) that correction, smoothing, or interpretation leads to clearer and longer readings.

In the 19th century, classical philologist Karl Lachmann's critical text (1831) bypassed the *T.R.*, using Greek manuscripts prior to the 4th century. C. von Tischendorf's discovery of ℵ (S) and his New Testament text (8th edition, 1864) collated the best manuscripts and had the richest critical apparatus thus far.

Two English biblical scholars, B.F. Westcott and F.J.A. Hort of Cambridge, using ℵ and B, brought out an edition in 1881–82 and classified the text witnesses into four groupings: Neutral (B, ℵ, the purest and earliest Eastern text); Alexandrian (a smoothed Neutral text as it developed in Alexandria); Western (D, Old Syrian, O.L., the Western Fathers with glosses that caused many readings to be rejected); and Syrian (A^e and the Byzantine tradition as it later developed). Such a "family tree" clearly showed the *T.R.* (Syrian) and, hence, the King James Version based upon it as an inferior text type; and the Revised Standard Version is based on such superior text types as B and ℵ.

Another critical edition (1902–13) was made by H. von Soden, a German biblical scholar who presupposed recensions to which all manuscripts can lead back. The importance of his work is in his enormous critical apparatus rather than in his theoretical groupings. B.H. Streeter, an English scholar, revised Westcott and Hort's classification in 1924. Basically, he challenged the concept of any uncontaminated descent from originals and made the observation (already alluded to in the evolution of papyrus evidence) that even the earliest manuscripts are of mixed text types. Yet, Streeter grouped texts in five families: Alexandrian, Caesarean, Antiochene, European Western, and African Western—parts of which all led into the Byzantine text and had become the *T.R.*

Unlikelihood of reconstructing an autograph

Despite grouping, it is clear that no reading backward from text families can reach an autograph. A strictly local text theory is useless in view of the papyrus evidence that there were no "unmixed" early texts. The use of external evidence cannot push beyond the boundary of the 3rd century. This insight brought about a new perspective. Only by using the canons of the internal evidence of readings can the best texts be determined, evaluating the variants from case to case—namely, the eclectic method. In modern times, therefore, the value of text families is primarily that of a step in the study of the history of the texts and their transmission. The eclectic method of reconstruction of an earliest possible New Testament text will yield the closest approximation of the historical texts put together into the New Testament canon. (For other, later and modern versions, see above *Old Testament canon, texts, and versions.*)

VII. New Testament history

THE JEWISH AND HELLENISTIC MATRIX

The historical background of the New Testament and its times must be viewed in conjunction with the Jewish matrix from which it evolved and the Hellenistic (Greek cultural) world into which it expanded during a period of Jewish religious propaganda. It is difficult, however, to separate the phenomena of the Jewish and Hellenistic backgrounds, because the Judaism out of which the church arose was a part of a very Hellenized world. The conquests of Alexander the Great culminated in 331 BC, and the subtle but strong influence of Greek culture, language, and customs that was spread by his conquests united his empire. Jews in both Palestine and the Diaspora (Dispersion) were, however, affected by Hellenism, as in ideas of cosmic dualism and rich religious imagery derived in part from Eastern influence as a result of the Greek conquests. Greek words were transliterated into Hebrew and Aramaic even in connection with religious ideas and institutions as, for example, synagogue (religious assembly), Sanhedrin (religious court), and paraclete (advocate, intercessor). It could be argued that the very preoccupation with ancient texts and tradition and the interpretation thereof is a Hellenistic phenomenon. Thus, what may appear as the most indigenous element in the activity of the Jewish scribes, sages, and rabbis (teachers)—*i.e.*, textual scholarship—has its parallels in Hellenistic culture and is part of the general culture of the times. The thought worlds merged, confronted each other, and communicated with each other.

The Hasmonean kingdom. After Alexander's death the empire was split, and first the Ptolemies, an Egyptian dynasty, and then the Seleucids, a Syrian dynasty, held Palestine. Antiochus IV Epiphanes, a 2nd-century-BC Seleucid king, desecrated the Temple in Jerusalem; a successful Jewish revolt under the Maccabees, a priestly family, resulted in its purification and in freedom from Syrian domination in 164 BC. This began the Hasmonean (Maccabean) dynasty, which appropriated both the powers of king and high priest. This reign, which created dissatisfaction on the part of other groups who considered their own claims falsely usurped, lasted until internecine strife brought it to an end. John Hyrcanus II, a 1st-century-BC Hasmonean king, appealed to Rome for help, and Pompey, a Roman general, intervened, bringing Palestine under Roman rule in 63 BC. John Hyrcanus, given the title of ethnarch was later executed for treason (30 BC), thus ending the Hasmonean line, but Jewish independence came to an end by Roman occupation.

Roman intervention in Palestine

Rule by the Herods. The Herods who followed were under the control of Rome. Herod the Great, son of Antipater of Idumaea, was made king of Judaea, having sided with Rome, and he ruled with Roman favour (37–4 BC). Though he was a good statesman and architect, he was hated by the Jews as a foreigner and semi-Jew. Jesus was born a few years before the end of his reign, and "the slaughter of the innocents," young children of Bethlehem who were killed as possible pretenders to Herod's throne, was attributed to Herod. After his death, Palestine was divided among three of his sons: Philip was made tetrarch of Iturea (the northeast quarter of the province) and ruled from 4 BC until AD 37. Herod Antipas became tetrarch of Galilee and Peraea until AD 39 and, like his father, was a builder, rebuilding Sepphoris and Tiberias before he was banished. Herod Antipas had John the Baptist beheaded and treated Jesus with contempt at Jesus' trial before him, before sending him back to Pontius Pilate, the Roman procurator (AD 26–36) at the time of Jesus' Crucifixion. Archelaus was made ethnarch of Judaea, Samaria, and Idumaea but was removed by AD 6 for his oppressive rule, and Judaea then became an imperial province, governed by procurators responsible to the emperor.

Two other Herods are mentioned in the New Testament: Agrippa I (called "Herod the king," AD 37–44) had James, the brother of John, killed and had Peter arrested; and the last of the Herods, Agrippa II, king of Trachonitis (*c.* AD 50–100), welcomed the procurator Festus (*c.* AD 60–62), who replaced Felix (*c.* AD 52–60) for the trial of Paul.

Roman occupation and Jewish revolts. In AD 66–70 there was a Jewish revolt while Nero was emperor of Rome (54–68). When he died and was succeeded by Vespasian, his former army commander (69–79), the siege and final destruction of Jerusalem occurred (AD 70). Before this event, Jewish Christians had fled, perhaps to Pella, and Yohanan ben Zakkai, a leading Jewish rabbi, with a group of rabbinical scholars, fled to Yavneh, where they established an academy that gave leadership to the Jews. Under both the emperors Trajan (98–117) and Hadrian (117–138), Jews in Egypt and Mesopotamia rebelled and again fought unsuccessfully against Rome in Palestine for forbidding the practice of religious rites, and, under Simeon Bar Kokhba (or Bar Koziba), a Jewish revolutionary messianic figure, the final Jewish war was waged (132–135). After this defeat Jerusalem became a Roman colony; a temple to Jupiter was erected there, and Jews were prevented from entering the city until the 4th century.

The effects of the fall of Jerusalem

When the Romans had entered Palestine in 63 BC, they practiced a relatively humane occupation until *c.* AD 66–70. They did not interfere with religious practices unless they considered them a threat to Rome, and their rights of requisition were precise and limited.

JEWISH SECTS AND PARTIES

From both the New Testament and extrabiblical material the main religious groups or parties in Palestinian Judaism may be discerned. Such descriptions, however, may be somewhat biassed or apologetic. Philo, an Alexandrian Jewish philosopher (died *c.* AD 40), Josephus, a Jewish apologist to the Romans (died *c.* 100), and sectarian writings found at Qumrān near the Dead Sea in 1947 that date back to about *c.* 200 BC and end about AD 70 all provide data about the respective Jewish religious groups in Palestine in the 1st century BC and the 1st century AD. The Pharisees (typically Jesus' opponents, although his ideas may have been close to their own), the Sadducees, and the Zealots are mentioned in the New Testament. The Essenes were described by Philo and Josephus, but new evidence from their own writings makes their group better understood (*i.e.*, the Dead Sea Scrolls from Qumrān).

The Pharisees. The Pharisees (possibly spiritual descendants of the Ḥasidim [Pious Ones], who were the exponents of Maccabean revolt) were strict adherents to the Law. Their name may come from *parush*—*i.e.*, "separated" from what is unclean, or what is unholy. They were deeply concerned with the Mosaic Law and how to keep it, and they were innovators in adapting the Law to new situations. They believed that the Law was for all the people and democratized it—even the priestly laws were to be observed by all, not only by the priestly class —so that they actually had a belief in a priesthood of all believers. They included Oral as well as Written Law in their interpretations. Though they did not accept the Roman occupation, they kept to themselves, and by pious acts, such as giving alms and burying the dead, they upheld the Law. Their interpretations of Law were sometimes considered casuistic because they believed they must find interpretations that would help all people to keep the Law. Their underlying hope was eschatological: in the day when Israel obeyed the Torah, the Kingdom would come. The Pharisees were called "smooth interpreters" by their opponents, but their hope was to find a way to make the living of the Law possible for all people. In their meal fellowship (*ḥavura*) they observed the laws strictly and formed a nucleus of obedient Israel. The Pharisees believed in the resurrection of the dead and had a developed angelology.

Pharisaic interpretations of the Law

The Sadducees. The Sadducees, more conservative and static, consisted mainly of the old priesthood and landed aristocracy and, perhaps, some Herodians. They were collaborators with Rome. They did not believe in resurrection because they found no Old Testament enunciation of such a doctrine. In a way, they seemed to respect the Pharisees in legal matters; but both the Pharisees—because they were a bourgeois rather than a popular movement—and the Sadducees—because they were aristocrats —rejected the *ʿam ha-aretz* (People of the Land), who were no party but simply the poor, common people whom they considered ignorant of the Law.

The Zealots. The Zealots were revolutionaries who plotted actively against the Roman oppression. That the Pharisees did not react in this way was perhaps because of their belief in Providence: what happens is the will of God, and their free will is expressed in the context of trust and piety in conjunction with an eschatological hope of winning God's Kingdom through obedience to Law.

The Essenes. Though the Essenes of the Dead Sea Scrolls are not mentioned in the New Testament, they are described by Philo, Eusebius, a 4th-century Christian historian, and Josephus. With publication of the Essenes' own sectarian writings since the 1950s, however, they have become well-known. They did not have any really new ideas, but their founder, the Teacher of Righteousness, believed that he knew the interpretation of the prophets for his time in a way that was not even known to the prophets of their own day. Their withdrawal into desert seclusion was in opposition to the ruling powers in the city and the Temple of Jerusalem. They lived apart from society in constant study of the Scriptures and with a firm belief that they were the elect of Israel living in the end of days and to whom would come messianic figures—a messiah of David (royal) and a messiah of Aaron (priestly). Membership in their group and acceptance or rejection of its founder determined their

Eschatological views of the Essenes

place in the age to come. After a long period of probation and initiation, a man became a member of this elect community that had strict rules of community discipline that would seal or destroy his membership in their New Covenant. Ritual lustrations preceded most liturgical rites, the most important one of which was participation in a sacred meal—an anticipation of the messianic banquet, to which only the fully initiated members in good standing were admitted and which was presided over by representatives of the Davidic and Aaronic messiahs. From what is known of them, their communities were celibate, living "in the presence of the angels" and thus required to be in a state of ritual purity. Their laws were strict, their discipline severe, and—unlike Pharisees, Sadducees, and Zealots—they were not simply different parties within Judaism but a separate eschatological sect. The Pharisees did have lodges and a common meal, but membership in the Pharisaic party did not, as it did with the Essenes, guarantee a place in the age to come; and the attitude of the Pharisees to a leader or founder was not, as it was to the Essenes, one of the bases on which such place could be attained. Thus, the Essenes—as the early Jewish Christians—were an eschatological Jewish sect. They believed that they alone, among those living in the end time, would be saved. The apocalypticism of the Essenes and the early Christians had many similarities, but the Christians had a higher eschatological intensity because they already knew who the Messiah would be when he came in the future at the Parousia (the "Second" Advent), and they also had a recollection of the earthly Jesus, knowledge of the risen Lord, and the gift of the Spirit upon the church. Both communities lived in an era wherein the cosmic battle of God versus Satan-Belial was taking place, but the Christian community already had the traditions of Jesus' victory over Satan and the experience of his Resurrection. Both Essenes and Christians were sects with tightly knit organizations, but the church had a historically based messiah. The Essenes probably were killed or forced to flee from their wilderness community *c.* AD 68, yet some of their ideas can still be traced in the ministry of John the Baptist (who might have been an Essene) and in the thought world of the New Testament (see also JUDAISM, HISTORY OF).

Similarities and differences between Essenes and Christians

THE RELIGIOUS SITUATION IN THE GRECO-ROMAN WORLD OF THE 1ST CENTURY AD

Hellenistic religions. With the expansion of Christianity into the Hellenistic world either to Jews or increasingly to Gentiles, there were various reasons why the Christian message that spread, for example by Paul, met the needs of the Hellenistic Age and world. There was no lack of religions, but there was a crisis of upheaval, unrest, and uncertainty and a desire to escape from mortality and the domination of unbending fate. There was also a desire to win personal knowledge of the universe and a dignified status within it—*i.e.*, a religious identity crisis. City-states with their cults of civic gods were unstable, because men changed from place to place and the gods of the city were distant from individual needs and anxieties. After Alexander's conquests, the resulting religious syncretism did not meet individual needs and longings that were increasingly becoming conscious. Many Gentiles turned to Judaism, at least as "god fearers," and later to Christianity. There were also "mystery religions," the secrets of which were known only to the initiate, which may have arisen from Eastern fertility cults with their dying and rising gods and were transformed in the Hellenistic Age to cults of a saviour god whose dying and rising gives personal immortality. Such mystery cults often provided meaningful relationships with fellow initiates (see also HELLENISTIC RELIGIONS; MYSTERY RELIGIONS).

The religious crisis of the 1st century AD

Astrology. There were elements in the Greek world that may have come from the East, partly Egyptian and Babylonian, which gave rise to astrology. The basic conviction of astrology was that the heavenly bodies were deities that in a direct way control life and events on earth. An older idea of *tychē*, or "fate," originally signified the chance element in the universe, a capriciousness that increased insecurity. Astrology transformed this into a fate or destiny in which everything is strictly regulated by celestial deities. Man's problem, then, is that of finding security from overwhelming powers outside human control. One way is to "read a horoscope." Because the heavenly deities are systematic and orderly according to astronomic observation, this order and regularity can be exploited to see how and in what way events will happen and can perhaps be used or avoided. Another way is to deal with such forces through magic. From the Hellenistic period many magical papyri with formulas for dealing with sicknesses, demons, and other adverse forces have been found. Magic attempts to manipulate and control what affects the world by a kind of participation in the event (see also ASTROLOGY; MAGIC).

Philosophical solutions. Solutions were also sought in philosophy. Socrates, a 5th-century-BC Greek philosopher, was largely concerned with the search for the "good," the good life. After Plato and Aristotle, however, philosophical systems sought to supply man's longing for inward security and stability. These were sought not by an in-depth understanding of reality but by ad hoc constructions—a new dogmatism for providing infallible plans and attaining immediate security—that the age demanded. Those philosophies were crude constructions that gave shelter and were defended by an unyielding dogmatism as absolute truths; if they were proved false, they would remove their promised security. Epicureanism, founded by the Greek philosopher Epicurus (341–270 BC), was basically a philosophy of escape, and its goal was serenity and tranquillity, a negative concept characterized by absence of fear, pain, and struggle. Fate, providence, and the afterlife were eliminated to deny the anxieties they provoked in terms of control, reward, or judgment. Epicurus attempted to meet this crisis by adopting a completely material view of the universe, including the soul, and thereby eliminating interference by deities both in life and after death. He did believe in the gods; but they, too, lived in their own perfect tranquillity, away from the universe. The Epicurean was both self-reliant and at peace with the absence of pain. There was also emphasis on friendship and the development of close communities.

Goals of escape or acceptance

Zeno, a 3rd-century-BC philosopher, was the founder of Stoicism. Stoicism was a rule of life that held that all reality was material but was animated by a rational principle that was at the same time both the law of the universe and of the human soul. The wise man then could accept and learn to live a life in conformity to this permeating reason without letting anything affect him. He responded to duty and accepted it.

Cynicism was a philosophy that maintained a cosmic view of life with a method of dealing with crisis by reducing man's needs to a minimum. Later in the Hellenistic period, a group of Stoic–Cynic preachers arose and, in New Testament times, wandered around calling men to repent and change their lives from sin to virtue (see also PHILOSOPHY, HISTORY OF WESTERN).

ADAPTATION OF THE CHRISTIAN MESSAGE TO THE HELLENISTIC RELIGIOUS SITUATION

The Christian message adapted itself to this Hellenistic situation of crisis and proved a successful answer: Jesus was proclaimed as Lord and Saviour, Baptism was practiced as a form of initiation and a passage from death to new life, and the Lord's Supper was celebrated as a sacral meal. The obvious difference between Christianity and the mystery religions is that a historical person, Jesus, forms the center of cult and devotion; his titles came from his Jewish background. Adaptation took place out of the Jewish matrix of Christianity—and Hellenistic terms that were meaningful were also used, such as illumination and regeneration. Such terms are not to be found in the earliest origins of Christianity but in the communication of the Christian message to a new environment. Among the religious and philosophic needs of the time was that of a cult that provided for the needs of the individual along with a community of worship. Christ

as Lord was viewed as universal, and his teachings made the universe understandable, as well as providing a basis for ethics. In a period of expansion, all religions are to some extent syncretistic, as is the case of Christianity in the 2nd century. Such a phenomenon belongs to a religion in a time of strength. Though universal, however, Christ was believed to have an exclusive claim, and in this there was security and relief for the anxieties of the period. The church was more than a philosophy; it had a social and enduring structure. It also reached out to all men—not only to those regarded as the best of men. It called them to a new life and gave them a new home and community, the church.

THE LIFE OF JESUS

Difficulties in establishing a chronology of the life of Jesus

Though the fact that Jesus was a historical person has been stressed, significant, too, is the fact that a full biography of accurate chronology is not possible. The New Testament writers were less concerned with such difficulties than the person who attempts to construct some chronological accounts in retrospect. Both the indifference of early secular historians and the confusions and approximations attributable to the simultaneous use of Roman and Jewish calendars make the establishment of a chronology of Jesus' life difficult. That the accounts of Matthew and Luke do not agree is a further problem. Thus, only an approximate chronology may be reconstructed from a few somewhat conflicting facts. The points of reference are best taken from knowledge of the history of the times reflected in the passages.

According to Matthew, Jesus was born near the end of the reign of Herod the Great, thus before 4 BC. In Luke chapter 2, verses 1 to 2, Jesus is said to have been born at the time of a census when Quirinius was governor of Syria. Such a census did occur, but in AD 6–7. Because this was after Herod's death and not in agreement with a possible date of Jesus' baptism, this late date is unlikely. There may have been an earlier census under another governor; an inscription in the Lateran Museum records an unnamed governor who twice ruled Syria, and the suggestion has been made that this was, indeed, Quirinius and that in an earlier time a reported census according to Roman calculation might have been carried out *c.* 8 BC, one of a series of such. With such speculation and the combined evidence of Matthew and Luke, an approximate year of birth might be 7–6 BC.

In Luke chapter 3, verse 23, it is stated that Jesus' ministry began when he was about 30 years of age. This would not come within the dates of the procuratorship of Pontius Pilate (AD 26–36), and the age might simply approximate a term for Jesus' having arrived at maturity. In Luke several dates are implied to assist in dating the Baptism of Jesus: the 15th year of Tiberius (*c.* 29, according to his accession as co-emperor with Augustus), while Pontius Pilate was in office (during 26–36), while Herod Antipas was tetrarch (4 BC–AD 39) and Herod Philip tetrarch (4 BC–AD 37). These limits make a speculation of Jesus' Baptism and the start of his ministry *c.* AD 27/28.

The duration of Jesus' ministry can be an average of the one year, as indicated in the Synoptic Gospels (Matthew, Mark, and Luke) or about three years as indicated in John, based on various cycles of harvests and festivals. This would be about two years. Because Jesus was crucified before 36 and his ministry started about 27/28, he then was crucified about AD 30 (see also JESUS CHRIST).

THE CHRONOLOGY OF PAUL

For the chronology of Paul's ministry, there are also some extra-biblical data: According to Josephus, Herod Agrippa I was made ruler of all Palestine by the emperor Claudius in AD 41 and reigned for three years. His death was thus in AD 44. A famine in Claudius' reign took place when Tiberius Alexander was procurator of Judaea (*c.* 46–48), and Egyptian papyri suggest (by reference to high wheat prices) that the date of the famine was about 46. The Gallio inscription at Delphi (in Greece) gives a date for Gallio, proconsul of Achaia when Paul was at Corinth. It notes that Claudius was acclaimed emperor for the 26th time. This would bring the date of being declared emperor to about 52 and Gallio's term of office (about one year) to about 51–52.

The chronology of Paul's missionary journeys

The chronology of Paul's missionary journeys and the dates of his letters have been the object of an investigation made difficult by the fact that the account in Acts does not agree with Paul's own letters, which are, of course, more reliable.

With the help of external references, some degree of absolute chronology might be sought—with several years' margin both because of uncertainty as to extra-biblical dating and much ambiguity about internal evidence. Although Paul would be in a better position to know his own situation, often his letters are, in their present form, combined fragments from various times (see below *The Second Letter of Paul to the Corinthians; The Letter of Paul to the Phillipians*). A chronology can be reached by comparing Paul's accounts of his journeys and sojourns with those reported in Acts. Given references in Acts and the Gallio inscription, it is possible to place Paul in Corinth in AD 51, and, since he was there for 18 months, it can be assumed that he began his missionary work sometime in 49 (he had previously been in Thessalonica and Philippi and in Troas and Asia Minor). This probably fits in with the "expulsion" of Jews from Rome about AD 49, thus indicating that Paul met Priscilla and Aquila, two Roman Jewish Christians, in Corinth at this time. This indicates that he was at an "apostolic conference" at Jerusalem sometime shortly before this (a comparison of chapters 13 and 15 of Acts with chapters 1 and 2 of Galatians shows that the author of Acts made two visits out of the one recorded by Paul), either in 49 or 48.

Though the dates in Galatians 1 and 2 are uncertain—not indicating whether they refer to 17 years *in toto* or only 14 years, because half years were equated with whole ones—they do establish the call of Paul to become a Christian in 31 or about 34–35. Working in the other direction, it is known that Paul wrote to the Thessalonians from Corinth, thus indicating a date of about 50 as probable for the writing of I Thessalonians.

From Corinth, Paul went to Ephesus, where, according to Acts, he remained (probably in prison) for three years. This would place him in Ephesus during the period 52–55, thus allowing time for a journey from Corinth via Ephesus to Antioch and then back to Ephesus. A sequence given in Acts, chapters 16 and 18, shows two possibilities for Paul to have been in Galatia that work in agreement with Galatians, chapter 4, verse 13, demonstrating that Galatians was written from Ephesus about 53–54. Ephesus can also be the location from which came I Cor., Phil., and probably Philem.

II Corinthians appears to have been written from Macedonia during 55. From the dating of the periods of Felix and Festus in office at Caesarea (mid-50s) and from the events in Felix' time of office, it is probable that Paul was in prison under Felix by 56.

Thus, data of Acts 18 and 20 regarding the journey and sojourn at Corinth can be correlated with data in Romans 15 to place the epistle to the Romans about 56, before the journey back to Jerusalem, ending in the arrest of Paul in 56. The two years of Acts 24:27 can then be explained as the time during which Paul was in prison at Caesarea, so that in 58 Paul was before Festus and sent to Rome.

That Paul was then in Rome for two more years is established in Acts chapter 28, verse 30. It can be concluded that Paul died sometime after 60, possibly during or before the Neronian persecution of 64 (*cf. I Clem.* 5). All this does not resolve the question of a possible Spanish journey nor give precise dates and locations for II Thessalonians, Colossians, Ephesians, or the Pastoral Letters (see also PAUL THE APOSTLE, SAINT).

VIII. New Testament literature

INTRODUCTION TO THE GOSPELS

Meaning of the term gospel. From the late AD 40s and until his martyrdom in the 60s, Paul wrote letters to the churches that he founded or guided. These are the earliest Christian writings that the church has, and in them

Earliest concept of "gospel"

he refers to "the gospel." In Romans chapter 1, verse 1, he says: "Paul, a servant of Jesus Christ, called to be an apostle, set apart for the gospel of God . . ." and goes on to describe this "gospel" in what was already by that time traditional language, such as: "promised beforehand through his prophets in the holy scriptures, the gospel concerning his Son, who was descended . . . our Lord" (Rom. 1:1–4). This gospel is the power of God for salvation to everyone who has faith ". . . for in it the righteousness of God is revealed through faith for faith . . ." (1:17). In I Corinthians Paul had reminded his congregation in stylized terms of "the gospel" he had brought to them. It consisted of the announcement that Jesus had died and risen according to the Scriptures.

Thus, the "gospel" was an authoritative proclamation (as announced by a herald, *kēryx*), or the kerygma (that which is proclaimed, *kērygma*). The earthly life of Jesus is hardly noted or missed, because something more glorious—the ascended Lord who sent the Spirit upon the church—is what matters.

In the speeches of Peter in Acts, the transition from kerygma to creed or vice versa is almost interchangeable. In Acts 2 Jesus is viewed as resurrected and exalted at the right hand of God and made both Lord and Christ. In Acts 3 Peter's speech proclaims Jesus as the Christ having been received in heaven to be sent at the end of time as judge for the vindication and salvation of those who believe in him. Here the proclaimed message, the gospel, is more basic than an overview of Jesus' earthly life, which in Acts is referred to only briefly as "his acting with power, going about doing good, and healing and exorcising" (10:38ff.). Such an extended kergyma can be seen as a transition from the original meaning of gospel as the "message" to gospel meaning an account of the life of Jesus.

The term gospel has connotations of the traditions of Jesus' earthly ministry and Passion that were remembered and then written in the accounts of Matthew, Mark, Luke, and John. They are written from the post-Resurrection perspective and they contain an extensive and common Passion narrative as they deal with the earthly ministry of Jesus from hindsight. And so the use of the term gospel for Matthew, Mark, Luke, and John has taken the place of the original creedal–kerygmatic use in early Christianity. It is also to be noted that, in the

Theological presuppositions that helped to shape the Gospels

Evangelists' accounts, their theological presuppositions and the situations of their addressees molded the formation of the four canonical Gospels written after the Pauline Letters. The primary affirmations—of Jesus as the Christ, his message of the Kingdom, and his Resurrection—preceded the Evangelists' accounts. Some of these affirmations were extrapolated backward (much as the Exodus event central in the Old Testament was extrapolated backward and was the theological presupposition for the patriarchal narratives in Genesis). These stories were shaped by the purpose for their telling: religious propaganda or preaching to inspire belief. The kerygmatic, or creedal, beginning was expanded with material about the life and teaching of Jesus, which a reverence for and a preoccupation with the holy figure of Jesus demanded out of loving curiosity about his earthly ministry and life.

The English word gospel is derived from the Anglo-Saxon *godspell* ("good story"). The classical Greek word *euaggelion* means "a reward for bringing of good news" or the "good news" itself. In the emperor cult particularly, in which the Roman emperor was venerated as the spirit and protector of the empire, the term took on a religious meaning: the announcement of the appearance or accession to the throne of the ruler. In contemporary Greek it denoted a weighty, authoritative, royal, and official message.

In the New Testament, no stress can be placed on the etymological (root) meaning of *eu* ("good"); in Luke chapter 3, verse 18 (as in other places), the word means simply authoritative news concerning impending judgment.

Form criticism. In the Pauline writings, as noted above, gospel, kerygma, and creed come close together from oral to written formulas that were transmitted about the Christ event: Jesus' death and Resurrection. In the apostolic Fathers (early 2nd century), the transition was made from oral to written tradition; the translation of the presumed Aramaic traditions had taken place before the Gospel material had been committed to writing. By the time of Justin Martyr (*c.* 155), these writings were called Gospels and referred to in the plural; they contain the words, deeds, and Passion narratives—*i.e.*, the present four Gospels compiled and edited by the Evangelists according to their various needs and theological emphases. Justin also referred to these as "memoirs of the Apostles."

Patterns noted in the Gospels

Such a Gospel began with a missionary announcement concerning a cosmic divine figure, a man with divine characteristics who would bring salvation and hope to the world. The earthly historical Jesus, however, was the criterion of the proclamation—being both the content of the church's proclamation and the object of its faith.

The identification of basic patterns in the history of oral and written traditions—the stage of tradition prior to any literary form and particularly as the traditions passed from an oral to a written form—and the determination of their creative milieu, or their situations and functions in various places and under various circumstances, are tasks of form criticism. Through such study, small independent units may be isolated in a postulated more primitive form than they were before being incorporated into more extended accounts. The term *Sitz-im-Leben* refers to the "Sitz im Leben der Kirche"—*i.e.*, the situation in the life of the church in which the material was shaped and adjusted to the needs at hand. Only through such studies is it possible to progress tentatively to an assessment of a "Sitz im Leben Jesu."

Both Jews and Gentiles could use "biographies," often for propaganda purposes. Philo and Josephus recounted the wonderful lives and deeds of Old Testament heroes such as Moses; and there are miraculous tales of the prophets Elijah and Elisha told in order that faith might be inspired or justified. A miracle worker (*theios anēr*, "divine man") and stories about him comprised an aretalogy (from *aretē*, "virtue"; also manifestation of divine power, miracle). Aretalogies were frequently used to represent the essential creed and belief of a religious or philosophical movement. The Life of Apollonius of Tyana, a Neo-Pythagorean philosopher and wonder-worker (transmitted by the Greek writer Philostratus), was widely read. He was depicted as having performed miracles and as being possessed of divine cosmic power not as an exception but as an example to men who have the possibility of sharing such power (*cf.* Matt. 9:8). There were tales of Heracles, the Greek hero, and a whole literature of Alexander the Great as wonder-workers, divine men.

Types of forms and genres

Though the pericopes (small units) of which the Gospels are constituted include many forms, or genres, they are mainly divided into narratives (including legends, miracle stories, exorcisms, healings, and tales) and sayings (prophetic and apocalyptic sayings, proverbs and wisdom sayings, parables, church discipline and rules for the community, Christological sayings, such as the so-called "I am" sayings [*e.g.*, "I am the bread of life"], in John, revelations, and legal sayings). Some stories may simply be the background for a pithy saying; these latter are sometimes called paradigmatic sayings, and the pronouncement stories are their vehicles of transmission. The forms have many different names, but form criticism started with Homeric form analysis (taking oral tradition into account), which was applied to Old Testament studies by Hermann Gunkel, a German biblical scholar, and applied to the New Testament, on the basis of the German classical philologist Eduard Norden's stylistic studies, by such biblical scholars as Rudolf Bultmann and Martin Dibelius.

Form criticism asks and answers questions about what shaped the preliterary tradition and the earliest written traditions into blocks as they are found in the Gospels. This may be a historical context (as a missionary situation), a need for admonition (as church-discipline sections), or for the transmission of teaching in a faithful

way (as in a "school," be it Matthean, Pauline, or Johannine). One large block of the material, however, is to all intents and purposes the same (although differing in details) in all four canonical Gospels: the Passion narrative. In the Synoptic Gospels there is also a basic nucleus in the sayings about Jesus that are mysterious, prophetic, and apocalyptic and that point to the significance of Jesus as the Christ who has come in history in the person of Jesus of Nazareth.

Such form-critical studies were centred on the smaller units of tradition (pericopes) that make up the Gospels, and their intention was partly to assess relative age and authenticity of such traditions. In more recent times the tools of form criticism have been applied to a more synthetic method that could be used to determine the relation between a genre of literature and the Christological and theological perspectives that made such genres natural. A presentation of Jesus material in the form of more or less disconnected sayings (as in the so-called Q Source, composed of independent sayings, behind Matthew and Luke, and in the *Gospel of Thomas;* see below *The two- and four-source hypotheses*) tends to fit a Christology in which Jesus is viewed as a teacher of Wisdom, an envoy of Wisdom, or as Wisdom herself. The collections of wonder stories (aretalogies) grew out of a Christology of Jesus as the divine man. Another type of Jesus material with independent existence seems to have been "revelations," or "apocalypses," in which Jesus Christ speaks to his followers. This is seen, for example, in Mark 13, I Thessalonians, chapter 4, the canonical book of Revelation to John, and the noncanonical *Didache* 16.

These genres of material now represented in the canonical Gospels are amply represented also in the non-canonical writings from the first Christian centuries. The discovery of a Gnostic library of Coptic writings at Naj' Ḥammādī, in Egypt, in the 1940s gave scholars a new opportunity to compare the canonical Gospels with the Jesus material of these various types, some of them having been called and used as gospels (such as the *Gospel of Thomas*). In the light of such a wider spectrum of material, it appears that the gospel form for which Mark is the earliest witness became a criterion for the orthodox transmission of the Christian message about Jesus. By making the confession of Jesus as the crucified and risen Lord (the earliest kerygma and "gospel" as found in Paul and Acts) the form of an extensive Passion account prefaced by a limited amount of narrative and teaching, Mark set the stage for a faith that anchored faith in Jesus Christ in the events of the earthly life of Jesus. This form of the "gospel" became the standard within which the other commonly accepted Gospels grew. It became the criterion for later creedal statements concerning Jesus Christ as true God and true man. By such a criterion, gospels that seemed to disregard his humanity (*e.g., Gospel of Thomas*, the *Gospel of Peter*) were judged heretical.

Significance of the Passion account

THE SYNOPTIC PROBLEM

Early theories about the Synoptic problem. Since the 1780s, Matthew, Mark, and Luke have been referred to as the Synoptic Gospels (from *synoptikos*, "seen together"). The extensive parallels in structure, content, and wording of Matthew, Mark, and Luke make it even possible to arrange them side by side so that corresponding sections can be seen in parallel columns. John Calvin, the 16th-century Reformer, wrote a commentary on these Gospels as a harmony. Such an arrangement is called a "synopsis," or Gospel harmony, and, by careful comparison of their construction, compilation, and actual agreement or disagreement in wording or content, literary- or source-critical relationships can be seen. Augustine, the great 4th–5th-century Western theologian, considered Mark to be an abridged Matthew, and, until the 19th century, some variation of this solution to literary dependency dominated the scene. It still recurs from time to time.

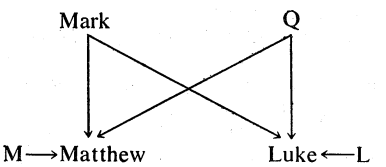
The Synoptic problem is one of literary or of source criticism and deals with the written sources after compilation and redaction. Matthew was the Gospel most used for the selections read in the liturgy of the church, and other Gospels were used to fill in the picture. One attempted solution to the problem of priority was the proposed existence of an Aramaic primitive gospel, now lost, as the first Gospel from which a later Mark in Greek was translated and arranged. The Greek Mark would thus be first based on a prior Semitic Matthew, and later both Mark and Matthew would be translations dependent on Matthew, and Luke dependent on both. The preservation of an ecclesiastical priority of Matthew breaks down because of the literary word-for-word agreement in Matthew, Mark, and Luke. This agreement occurs to far too great an extent to be accounted for in translations and revisions, not to mention the agreement in the order of the various pericopes as viewed in a synoptic parallel arrangement.

For similar reasons, a fragment theory holding that the Gospels were constructed of small written collections brought together in varying sequences cannot stand the test of actual structure—but it has the merit of stressing compilation of sources.

In 1789 J.J. Griesbach, a German biblical scholar, hypothesized that the Synoptics had not developed independently, but in his "usage-hypothesis" he recognized that there must be literary dependency. He thought that Mark used Matthew as well as Luke, but this could not account for the close relationship of Matthew and Luke. His basic concept of literary dependency, however, paved the way for K. Lachmann who observed in 1835 that Matthew and Luke agree only when they also agree with Mark and that, where material is introduced that is not in Mark, it is inserted in different places. This, it is held, can only be explained on the basis of the priority of Mark and its use as the patterning form of Matthew and Luke. This insight led to a so-called two-source hypothesis (by two German biblical scholars, Heinrich Holtzmann in 1863, and Bernhard Weiss in 1887–88), which, with various modifications and refinements of other scholars, is the generally accepted solution to the Synoptic problem.

The two- and four-source hypotheses. The two-source hypothesis is predicated upon the following observations: Matthew and Luke used Mark, both for its narrative material as well as for the basic structural outline of chronology of Jesus' life. Matthew and Luke use a second source called Q (from German *Quelle*, "source"), not extant, for the sayings (logia) found in common in both of them. Thus, Mark and Q are the main components of Matthew and Luke. In both Matthew and Luke there is material peculiar to each of their Gospels; this material is probably drawn from some other sources, which may be designated M (material found only in Matthew's special source) and L (material found only in Luke's special source). This is known as the four-document hypothesis, which was elaborated in 1925 by B.H. Streeter, an English biblical scholar. The placement of Q material in Luke and Matthew disagrees at certain points according to the needs and theologies of the addressees of the gospels, but in Matthew the Marcan chronology is the basic scheme into which Q is put. Mark's order is kept, on the whole, by Matthew and Luke, but, where it differs, at least one agrees with Mark. After chapter 4 in Matthew and Luke, not a single passage from Q is in the same place. Q was a source written in Greek as was Mark, which can be demonstrated by word agreement (not possible, for example, with a translation from Aramaic, although perhaps the Greek has vestiges of Semitic structure form). A diagram might thus be:

The source called Q



In approximate figures, Mark's text has 661 verses, more than 600 of which appear in Matthew and 350 in Luke. Only *c.* 31 verses of Mark are found nowhere in Matthew

or Luke. In the material common to all three Synoptics, there is very seldom verbatim agreement of Matthew and Luke against Mark, though such agreement is common between Matthew and Mark or Luke and Mark or where all three concur.

The postulated common saying source of Matthew and Luke, Q, would account for much verbatim agreement of Matthew and Luke when they include sayings absent from Mark. The fact that the sayings are used in different ways or different contexts in Matthew and Luke is an indication of a somewhat free way in which the editors could take material and mold it to their given situations and needs. An example of this is the parable in Matthew and Luke about the lost sheep (Matt. 18:10–14, Luke 15:3–7). The basic material has been used in different ways. In Matthew, the context is church discipline—how a brother in Christ who has lapsed or who is in danger of doing so is to be gently and graciously dealt with—and Matthew shapes it accordingly (the sheep has "gone astray"). In Luke, the parable exemplifies Jesus' attitude toward sinners and is directed against the critical Pharisees and scribes who object to Jesus' contact with sinners and outsiders (the sheep is "lost").

Variations in the use of certain sayings of Jesus

Another example of two passages used verbatim in Luke and Matthew is Jesus' lament over Jerusalem. In Luke (13:34–35; the lament over Jerusalem) Jesus refers to how they will cry "Blessed be the King who comes in the name of the Lord" when he enters Jerusalem (Lk. 19:38). In Luke, the passage is structured into the life of Jesus and refers to his triumphal entry into Jerusalem, "Blessed is he who comes in the name of the Lord"). In Matthew (23:37–39) this same lament is placed after the entry into the city (21:9) and thus refers to the fall of Jerusalem and the Last Judgment. Apparently, Luke has historicized a primarily eschatological saying.

Since the 1930s, scholars have increasingly refined sources, postulated sources behind sources, and many stages of their formation. The premise of the two- (or four-) source hypothesis is basic and provides information as to literary sources; further refinement is of interest only to the specialist. Another movement in synoptic research—and also research including John—is that which concentrates rather on the treatment of gospels as a whole, formally and theologically, with patterns or cycles to be investigated. It may be significant that the latest and best regarded Greek synopsis is that of the German scholar Kurt Aland, *Synopsis Quattuor Evangeliorum* (1964; *Synopsis of the Four Gospels*, 1972), which includes the Gospel According to John and, as an appendix, the *Gospel of Thomas*, as well as ample quotations from noncanonical gospels and Jesus' sayings preserved in the Church Fathers.

THE SYNOPTIC GOSPELS

The Gospel According to Mark. The Gospel According to Mark is the second in canonical order of the Gospels and is both the earliest gospel that survived and the shortest. Probably contemporaneous with Q, it has no direct connection with it. The Passion narrative comprises 40 percent of Mark, and, from chapter 8, verse 27, onward, there is heavy reference forward to the Passion.

Though the author of Mark is probably unknown, authority is traditionally derived from a supposed connection with the Apostle Peter, who had transmitted the traditions before his martyr death under Nero's persecution (*c.* 64–65). Papias, a 2nd-century bishop in Asia Minor, is quoted as saying that Mark had been Peter's amanuensis (secretary) who wrote as he remembered (after Peter's death), though not in the right order. Because Papias was from the East, perhaps the Johannine order would have priority, as is the case in the structure of the Syrian scholar Tatian's *Diatesseron* (harmony of the Gospels).

Traditional connection with Peter

Attempts have been made to identify Mark as the John Mark mentioned in Acts 12 or as the disciple who fled naked in the garden (Mark 14). A reference to "my son, Mark," in I Peter is part of the same tradition by which Mark was related to Peter; thus the Evangelist's apostolic guarantor was Peter.

The setting is a Gentile church. There is no special interest in problems with Jews and little precision in stating Jewish views, arguments, or terminology. Full validity is given the worship of the Gentiles. In further support of a Gentile setting and Roman provenance is the argument that Mark uses a high percentage of so-called Latinisms —*i.e.*, Latin loanwords in Greek for military officers, money, and other such terms. Similar translations and transliterations, however, have been found in the Jerusalem Talmud, a compendium of Jewish law, lore, and commentary, which certainly was not of Roman provenance. The argument from Latinisms must be weighed against the fact that Latin could be used anywhere in the widespread Roman Empire. In addition, for the first three centuries the language of the church of Rome was Greek—so the Gentile addressees might just as well have been Syrian as Roman. The Latinisms—as well as the Aramaisms—are rather an indication of the vernacular style of Mark, which was "improved" by the other Evangelists.

Simplicity and directness of style

Mark is written in rather crude and plain Greek, with great realism. Jesus' healing of a blind man is done in two stages: first the blind man sees men, but they look like trees walking, and only after further healing activity on Jesus' part is he restored to see everything clearly. This concrete element was lost in the rest of the tradition. It is also perhaps possible that this two-stage healing is a good analogy for understanding Mark theologically: first, through enigmatic miracles and parables in secret, and only later, after recognition of Jesus as the Christ, is there a gradual clarification leading to the empty tomb. In chapter 3, verse 21, those closest to Jesus call him insane ("he is beside himself"), a statement without parallel in the other Gospels.

In Mark, some Aramaic is retained, transliterated into Greek, and then translated—*e.g.*, in the raising of Jairus' daughter (5:41) and in the healing of the deaf mute (7:34). The well-known *abba*, Father, is retained in Mark's account of Jesus' prayer in Gethsemane. In the two miracle stories, the Aramaic may have been retained to enhance the miracle by the technique of preserving Jesus' actual words. And a cry of Jesus on the Cross is given in Aramaized Hebrew.

The stories in Mark are woven together with simple stereotyped connectives, such as the use of *kai euthus* ("and immediately," "straightway"), which may be thought of as a Semitic style (as a typical simple connective in the Old Testament narrative style). More likely, however, this abruptness indicated that the compiler-redactor of Mark has used geography and people simply as props or scenes to be used as needed to connect the events in the service of the narrative.

Except for the Passion narrative, there is little chronological information. References in chapters 13 and 14 appear to presuppose that the Jerusalem Temple (destroyed in AD 70) still stood (in Matthew and Luke this is no longer the case); but the context of chapter 13, the "Little Apocalypse," is so interwoven with eschatological traditions of both the Jewish and Christian expectations in the 1st century that it cannot serve with certainty as a historical reference. To some extent, however, chapter 13 does help to date Mark—the priority of which has already been established from literary criticism—because it is in good agreement with the traditions that Mark was written after the martyrdom of Peter. Mark may thus be dated somewhere after 64 and before 70, when the Jewish war ended.

Structure of the Gospel

The organization and schematizing of Mark reveals its special thrust. It may be roughly divided into three parts: (1) 1:1–8:26—the Galilean ministry—an account of mighty deeds (an aretalogy); (2) 8:27–10:52—discussions with his disciples centred on suffering; and (3) 11:1–16:8 —controversies, Passion, death, the empty tomb, and the expected Parousia in Galilee.

"The beginning of the Gospel" in the first words of Mark apparently refers to John the Baptist, who is clearly described as a forerunner of the Messiah who calls the people to repentance. Jesus never calls himself the Messiah (Christ). After Jesus' Baptism by John, the

heavens open, the Spirit descends, and a heavenly voice proclaims Jesus as God's beloved son with whom He is well pleased. Already in this account there is a certain secrecy, because it is not clear whether the onlookers or only Jesus witnessed or heard. Jesus was then driven by the Spirit into the wilderness, the place of demons and struggle, to be tempted by Satan, surrounded by wild beasts (the symbols of the power of evil and persecution) and ministered to by angels. Here again he is in secret, alone. The opening of the struggle with Satan is depicted, and the attendance by angels is a sign of Jesus' success in the test.

Many references to persecution in Mark point toward Roman oppression and a martyr church that was preoccupied with a confrontation with the Satanic power behind the world's hostility to Jesus and his message. There was stress on the underlying fact that the church must witness before the authorities in a hostile world. Much of the martyrological aspect of Mark's account is grounded in his interpretation of the basic function of Jesus' Passion and death and its implication that the Christian life is a life of suffering witness.

The proximity of the Kingdom of God

What Jesus preached in Galilee at the beginning of his ministry was that the time is fulfilled and the Kingdom of God is "at hand"; *i.e.*, very very near—therefore repent! (1:15). In Matthew this same message is that of both John the Baptist (3:2) and Jesus (4:17). This sets the stage; and the miraculous ministry in Galilee about which the followers are enjoined to secrecy points not so much to Jesus as the wonder-worker as to the great scheme of pushing back the frontier of Satan. Toward the end of this first section, the Pharisees ask Jesus for a sign, and he answers in no uncertain terms that no sign will be given (8:12). In the Synoptic Gospels the miracles are never called "signs" (as in John); and no sign is to be given prior to the cosmological, eschatological signs from heaven that belong to the end: darkening of the Sun and Moon and extreme tribulations that in postbiblical Jewish eschatology—the mood of the first Christian century—is a sign of the coming of the heavenly Son of man to judge the world.

Parables are a revelatory mode of expression; they are not just illustrations of ideas or principles. Jesus, the revealer, tells his disciples that the secret of the Kingdom of God is given to them but that to the outsider everything is in parables (or riddles) *in order that* they may not hear and understand lest they repent and be forgiven (4:10–12). This mystery and hiddenness is particularly related to the parables about the coming of the kingdom. Yet, even Jesus' disciples did not recognize him as the Messiah, although his miracles were such that only a messianic figure could perform them: forgiving sins on earth, casting out demons, raising the dead, making the deaf hear and the stammerer (the dumb) speak, and the blind to see—all fulfillments of Old Testament prophecy concerning the Messiah. Only the demons, supranatural beings, recognize Jesus. There is a constant campaign against Satan from the temptation after Jesus' Baptism until his death on the Cross, and, in each act of healing or exorcism, there is anticipated the ultimate defeat of Satan and the manifestation of the power of the new age. In all this Mark stresses the need for secrecy and Peter's confession of Jesus as the Christ (8:29) is told in Mark as the opportunity to motivate an acceptance of the admonition "not to tell" by reference to the necessity of suffering.

Emphasis on suffering and the messianic secret

This strong emphasis on the necessity of suffering—in the life of Jesus and in the life of the disciples—before the hour of victory gives the best explanation to what scholars have called the secrecy motif in Mark—*i.e.*, the constant stress on not telling the world about Jesus' messianic power.

According to William Wrede, a German scholar, the messianic secret motif was a literary and apologetic device by which the Christological faith of the early church could be reconciled with the fact that Jesus never claimed to be the Messiah. According to Wrede, Mark's solution was: Jesus always knew it but kept it a secret for the inner group. After Peter's confession at Caesarea Philippi, Jesus began to speak of a *suffering* Son of man. The Son of man in Jewish apocalyptic was a glorious, transcendent, heavenly figure who would come victorious on clouds of glory to judge the world at the end of time. Suffering was not part of this picture. E. Sjöberg (1955) has interpreted the messianic secret not as a literary invention but as an understanding both that the Messiah would appear without recognition except by those who are chosen and to whom he reveals himself and that he must suffer. For outsiders, then, he remains a mystery until the age to come. Even his disciples did not understand the necessity of suffering. Only in the light of Resurrection faith—the hope of the Parousia and final victory over Satan—could they understand that he had to suffer and die to fulfill his mission and how they, too, must suffer.

Martyrological aspects in Mark can be noted from the beginning. Already according to 2:20 Jesus' disciples are not to fast until "when the bridegroom is taken away from them and then they will fast. . . ." In Mark 8 to 10, there is great concentration on discussions with the disciples. The theme is suffering, and repeatedly they are reminded that there is no way of coming to glory except through suffering. Three Passion predictions meet either with rejection, fear, or confusion. In the Transfiguration (9:2–13; in which three disciples—Peter, James, and John —see Jesus become brighter and Elijah and Moses, two Old Testament prophets, appear) there is the same emphasis. The tension between future glory and prior suffering is the more striking when the Transfiguration is recognized as a Resurrection appearance, placed here in an anticipatory manner. The disciples are reminded of an association of Elijah with John the Baptist and his fate. This is also a hidden epiphany (manifestation)—the triumphal enthroned king closely juxtaposed with suffering and death.

After the third Passion prediction, in chapter 10, two of the disciples ask for places of honour when Jesus is glorified. He reminds them that suffering must precede glory for "The Son of man also came not to be served but to serve, and to give his life as a ransom for many." It is worth noting that this is the only reference to the death of Christ as a ransom or sacrifice but that Mark does not dwell on the Christological implications, but uses the saying for ethical purposes. Even so, the Marcan text gives one of the important building blocks for Christological growth and reflection on the suffering Son of man.

Just as Jesus' public ministry in Mark started with the calling of disciples, so the central part of the Gospel calls them to participate through suffering in his own confrontation with the power of Satan.

Emphasis on the Passion

In the last section of the Gospel, the scene is shifted to Jerusalem, where Jesus is going to die. His entry is described as triumphal and openly messianic and is accompanied by acted-out parables in a judgment of a barren fig tree, casting money changers out of the Temple, and in a parable of a vineyard in which the beloved son of the owner is killed. There is an increasing conflict and alienation of the authorities. Chapter 13, the "Little Apocalypse," made up of a complex arrangement of apocalyptic traditions, serves as instruction to the disciples and thence to the church that they must endure through tribulation and persecution until the end time. Thus, although the setting is Jerusalem, the orientation is toward Galilee, the place where the Parousia is expected (16:7). The Holy Spirit will come to those who must witness in the situation of trial before governors and authorities (13:11); in the final eschatological trials only by God's intervention can anyone endure unless the time be shortened for the elect. Because this chapter is shaped as a discourse that precedes the Passion narrative, it serves as a farewell address, a type of testament including apocalyptic sayings and warnings to the messianic community at the end of the "narrative" before the Passion —as do most testament forms (admonitions given before death to those beloved who will remain behind).

The Cross is both the high point of the Gospel and its lowest level of abject humiliation and suffering. A cry of

dereliction and agony and the cosmic sign of the rending of the Temple veil bring from a Gentile centurion acknowledgment of Jesus as Son of God. The disciples reacted to the scandal of the Cross with discouragement, although already the scene is set for a meeting in Galilee. There are no visions of the risen Lord, however, in the best manuscripts (verses 9–20 are commonly held to be later additions), and Mark thus remains an open-ended Gospel. The Resurrection is neither described nor interpreted. Not exultation but rather involvement in the battle with Satan is the inheritance until the victorious coming in glory of the Lord—a continual process with the empty tomb pointing to hope of the final victory and glory, the Parousia in Galilee. The Gospel ends on the note of expectation. The mood from the last words of Jesus to the disciples remains: What I say to you, I say to all: Watch!

The Gospel According to Matthew. Matthew is the first in order of the four canonical Gospels and is often called the "ecclesiastical" Gospel, both because it was much used for selections for pericopes for the church year and because it deals to a great extent with the life and conduct of the church and its members. Matthew gave the frame, the basic shape and colour, to the early church's picture of Jesus. Matthew used almost all of Mark, upon which it is to a large extent structured, some material peculiar only to Matthew, and sayings from Q as they serve the needs of the church. This Gospel expands and enhances the stark description of Jesus from Mark. The fall of Jerusalem (AD 70) had occurred, and this dates Matthew later than Mark, *c.* 70–80.

Matthew as the "ecclesiastical" Gospel

Although there is a Matthew named among the various lists of Jesus' disciples, more telling is the fact that the name of Levi, the tax collector who in Mark became a follower of Jesus, in Matthew is changed to Matthew. It would appear from this that Matthew was claiming apostolic authority for his Gospel through this device but that the writer of Matthew is probably anonymous.

The Gospel grew out of a "school" led by a man with considerable knowledge of Jewish ways of teaching and interpretation. This is suggested by the many ways in which Matthew is related to Judaism. It is in some ways the most "Jewish" Gospel. Striking are 11 "formula quotations" ("This was to fulfill what was spoken by the prophet . . .") claiming the fulfillment of Old Testament messianic prophecies.

The structure of Matthew

The outstanding feature of Matthew is its division into five discourses, or sermons, following narrative sections with episodes and vignettes that precede and feed into them: (1) chapters 5–7—the Sermon on the Mount—a sharpened ethic for the Kingdom and a higher righteousness than that of the Pharisees; (2) chapter 10—a discourse on mission, witness, and martyrological potential for disciples with an eschatological context (including material from Mark 13); (3) chapter 13—parables about the coming of the Kingdom; (4) chapter 18—on church discipline, harshness toward leaders who lead their flock astray and more gentleness toward sinning members; and (5) chapter 23–25—concerned with the end time (the Parousia) and watchful waiting for it, and firmness in faith in God and his Holy Spirit. Each sermon is preceded by a didactic use of narratives, events, and miracles leading up to them, many from the Marcan outline. Each of the five sections of narrative and discourse ends with a similar formula: "now when Jesus had finished these sayings. . . ." The style suggests a catechism for Christian behaviour based on the example of Jesus: a handbook for teaching and administration of the church. This presupposes a teaching and acting community, a church, in which the Gospel functions. The Greek word *ekklēsia*, ("church") is used in the Gospels only in Matthew (16:18 and 18:17).

The discourses are preceded by etiological (sources or origins) material of chapters 1–2, in which the birth narrative relates Jesus' descent (by adoption according to the will of God) through Joseph into the Davidic royal line. Though a virgin birth is mentioned, it is not capitalized upon theologically in Matthew. The story includes a flight into Egypt (recalling a Mosaic tradition). Some "Semitisms" add to the Jewish flavour, such as calling the Kingdom of God the Kingdom of the Heaven(s). The name Jesus (Saviour) is theologically meaningful to Matthew (1:21). Chapter 2 reflects on the geographical framework of the Messiah's birth and tells how the messianic baby born in Bethlehem came to dwell in Nazareth.

After the five narrative and discourse units, Matthew continues from chapter 26 on with the Passion narrative, burial, a Resurrection account, and the appearance of the risen Lord in Galilee, where he gives the final "great commission," with which Matthew ends.

The setting of Matthew

Matthew is not only an original Greek document, but its addressees are Greek-speaking Gentile Christians. By the time of the Gospel According to Matthew, there had been a relatively smooth and mild transition into a Gentile Christian milieu. The setting could be Syria, but hardly Antioch, where the Pauline mission had sharpened the theological issues far beyond what seems to be the case in Matthew. Matthew has no need to argue against the Law, or Torah, as divisive for the church (as had been the case earlier with Paul in Romans and Galatians, in which the Law was divisive among Gentile Christians and Jewish Christians), and, indeed, the Law is upheld in Matthew (5:17–19). For Matthew, there had already been a separation of Christianity from its Jewish matrix. When he speaks about the "scribes and the Pharisees," he thinks of the synagogue "across the street" from the now primarily Gentile church. Christianity is presented as superior to Judaism even in regard to the Law and its ethical demands.

The Matthean church is conscious of its Jewish origins but also of a great difference in that it is permeated with an eschatological perspective, seeing itself not only as participating in the suffering of Christ (as in Mark) but also as functioning even in the face of persecution while patiently—but eagerly—awaiting the Parousia. The questions of the mission of the church and the degree of the "coming" of the Kingdom with the person and coming of Jesus are handled by the Evangelist by a "timetable" device. The Gospel is arranged so that only after the Resurrection is the power of the Lord fully manifest as universal and continuing. Before the Resurrection the disciples are sent nowhere among the Gentiles but only to the lost sheep of the house of Israel; and the end time is expected before the mission will have gone through the towns of Israel. Even in his earthly ministry, however, Jesus proleptically, with a sort of holy impatience, heals the son of a believing Roman centurion and responds to the persistent faith of a Canaanite woman—whose heathen background is stressed even more than her geographical designation, Syro-Phoenician, given in the parallel in Mark—by healing her daughter. The Jewish origins of Jesus' teaching and the way the Evangelist presents them do not deny but push beyond them. The prophecies are fulfilled, the Law is kept, and the church's mission is finally universal, partly because the unbelief of the pious Jewish leaders left the gospel message to the poor, the sick, the sinner, the outcast, and the Gentile.

Jesus as teacher, Messiah, King, and Judge

In Matthew, because of the use of Q and Matthew's theological organization, there is stress on Jesus as teacher, his sharpening or radicalizing of the Law in an eschatological context; and Jesus is presented not in secret but as an openly proclaimed Messiah, King, and Judge. In the temptation narrative Jesus refuses Satan's temptations because they are of the devil, but he himself later in the Gospel does feed the multitude, and after the Resurrection he claims all authority in heaven and on earth. By overcoming Satan, Jesus gave example to his church to stand firm in persecution. Messianic titles are more used in Matthew than in Mark. In the exorcism of demoniacs, the demons cry out, calling him Son of God and rebuking him for having come "before the time" (8:29). Again, this shows that Jesus in his earthly ministry had power over demons, power belonging only to the Messiah and the age to come; and he pushed this timetable ahead. Yet, as in Mark, the miracles are not to be interpreted as signs. When asked for a sign, the Matthean account gives only the sign of Jonah, an Old Testament prophet—*i.e.*, the preaching of the gospel—which in later tradition took

on an added interpretation as presaging the Son of man (Jesus) being three days and nights in the tomb (12:40, a later addition to Matthew).

Even the antitheses in the Sermon on the Mount are not new but demonstrate a higher ethic—one that is sharpened, strict, more immediate because the end time is perceived as coming soon. People who took this intensification of the Law upon themselves dared to do it as an example of "messianic license"—*i.e.*, to use the ethics of the Kingdom in the present in a church still under historical ambiguity and in constant struggle with Satan.

At such points the peculiar nature of Matthew comes into focus. The sharpening of the Law and the messianic license for the disciples are clearly there. At the same time Matthew presents the maxims of Jesus as attractive to a wider audience with Hellenistic tastes: Jesus is the teacher of a superior ethic, beyond casuistry and particularism. Similarly, in chapter 15, he renders maxims about food laws as an example of enlightened attitudes, not as rules for actual behaviour.

According to Matthew, the "professionally" pious were blind and unhearing, and these traits led to their replacement by those who are called in Matthew the "little ones"; in Final Judgment the King-Messiah will judge according to their response to him who is himself represented as one of "the least of these." The depiction of Jesus as Lord, King, Judge, Saviour, Messiah, Son of man, and Son of God (all messianic titles) is made in a highly pitched eschatological tone. The Lord's Prayer is presented in this context, and, for example, the "temptation" (trial, test) of "Lead us not into temptation" is no ordinary sin but the ordeal before the end time, the coming of the Kingdom for which the Matthean church prays. Martyrdom, though not to be pursued, can be endured through the help of the Spirit and the example of Jesus.

The Passion narrative is forceful and direct. Pilate's part in sentencing Jesus to be crucified is somewhat modified, and the guilt of the Jews increased in comparison with the Marcan account. In Matthew the Resurrection is properly witnessed by more than one male witness so that there can be no ambiguity as to the meaning of the empty tomb. The risen Lord directs his disciples to go to Galilee, and the Gospel According to Matthew ends with a glorious epiphany there and with Jesus' commission to the disciples—the church—to go to the Gentiles, because the risen Jesus is Lord of heaven and earth for all time.

The Gospel According to Luke. Luke is the third in order of the canonical gospels, which, together with Acts, its continuation, is dedicated by Luke to the same patron, "most excellent" Theophilus. Theophilus may have been a Roman called by a title of high degree because he is an official or out of respect; or he may have been an exemplification of the Gentile Christian addressees of the Lucan Gospel. The account in Luke–Acts is for the purpose of instruction and for establishing reliability by going back to the apostolic age. The very style of this preface follows the pattern of Greek historiography, and thus Luke is called the "historical" Gospel. Historically reliable information cannot be expected, however, because Luke's sources were not historical; they rather were embedded in tradition and proclamation. Luke is, however, a historian in structuring his sources, especially in structuring his chronology into periods to show how God's plan of salvation was unfolded in world history. That he uses events and names is secondary to his intention, and their historical accuracy is of less importance than the schematization by which he shows Jesus to be the Saviour of the world and the church in its mission (Acts) to be part of an orderly progress according to God's plan.

The purpose of Luke

The sources of the Gospel are arranged in the service of its theological thrust with definite periodization of the narrative. Approximately one-third of Luke is from Mark (about 60 percent of Mark); 20 percent of Luke is derived from Q (sometimes arranged with parts of L). Almost 50 percent is from Luke's special source (L), especially the infancy narratives of John the Baptist and Jesus, and parables peculiar to Luke (*e.g.*, the prodigal son, the good Samaritan, the rich fool). L material is also interwoven into the Passion narrative. While Matthew structured similar teaching materials in his five discourses, Luke places them in an extensive travel account that takes Jesus from Galilee to Judaea via Jericho to Jerusalem. This is similar to the ways in which Acts is structured on the principle of bringing the word from Jerusalem to Rome (see below).

The author has been identified with Luke, "the beloved physician," Paul's companion on his journeys, presumably a Gentile (Col. 4:14 and 11; cf. II Tim. 4:11, Philem. 24). There is no Papias fragment concerning Luke, and only late-2nd-century traditions claim (somewhat ambiguously) that Paul was the guarantor of Luke's Gospel traditions. The Muratorian Canon refers to Luke, the physician, Paul's companion; Irenaeus depicts Luke as a follower of Paul's gospel. Eusebius has Luke as an Antiochene physician who was with Paul in order to give the Gospel apostolic authority. References are often made to Luke's medical language, but there is no evidence of such language beyond that to which any educated Greek might have been exposed. Of more import is the fact that in the writings of Luke specifically Pauline ideas are significantly missing; while Paul speaks of the death of Christ, Luke speaks rather of the suffering, and there are other differing and discrepant ideas on Law and eschatology. In short, the author of this gospel remains unknown.

The setting of Luke

Luke can be dated *c.* 80. There is no conjecture about its place of writing, except that it probably was outside of Palestine because the writer had no accurate idea of its geography. Luke uses a good literary style of the Hellenistic Age in terms of syntax. His language has a "biblical" ring already in its own time because of his use of the Septuagint style; he is a Greek familiar with the Septuagint, which was written for Greeks; he seldom uses loanwords and repeatedly improves Mark's wording. The hymns of chapters 1 and 2 (the Magnificat, beginning "My soul magnifies the Lord"; the Benedictus, beginning "Blessed be the Lord God of Israel"; the Nunc Dimittis, beginning "Now lettest thou thy servant depart in peace") and the birth narratives of John the Baptist and Jesus either came from some early oral tradition or were consciously modelled on the basis of the language of the Septuagint. These sections provide insight into the early Christian community, and the hymns in particular reflect the Old Testament psalms or the *Thanksgiving Psalms* from Qumrān. Though on the whole Matthew is the Gospel most used for the lectionaries, the Christmas story comes from Luke. The "old age" motif of the birth of John to Elizabeth also recalls the Old Testament birth of Samuel, the judge. All the material about John the Baptist, however, is deliberately placed prior to that of Jesus. When Mary, the mother of Jesus, visits Elizabeth, Jesus' superiority to John is already established. The Davidic royal tradition is thus depicted as superior to the priestly tradition.

Writing out of the cultural tradition of Hellenism and that of Jewish *'anawim* piety—*i.e.*, the piety of the poor and the humble entertaining messianic expectations—Luke has "humanized" the portrait of Jesus. Piety and prayer (his own and that of others) are stressed. Love and compassion for the poor and despised and hatred of the rich are emphasized, as is Jesus' attitude toward women, children, and sinners. In the Crucifixion scene, the discussion between the robbers and Jesus' assurance that one of them would be with him in Paradise, as well as the words, "Father, into thy hands I commit my spirit!"—which are in contrast to the cry of dereliction in Mark and Matthew—all point toward the paradigm of the truly pious man. Parables peculiar to Luke—among which are those of the good Samaritan, the importunate friend, the lost coin, and the prodigal son—have an element of warmth and tenderness. Thus, Luke "civilizes" the more stark eschatological emphasis of Mark (and Matthew), leading the way, perhaps, to a lessening of eschatological hopes in a time in which the imminent Parousia was not expected but pushed into the distant future.

The structure of the Gospel

The interplay between Luke and Acts reveals Luke's answer to the coming of the Kingdom. Once the church has the Holy Spirit, the delay of the Parousia has been answered for a time. Thus, Luke divides history into three periods: (1) the end of the prophetic era of Israel as a preparation for revelation, with John the Baptist as the end of the old dispensation; (2) the revelation of Jesus' ministry as the centre of time—with Satan having departed after the temptation and, until he once again appears, entering into Judas to betray Jesus; and (3) the beginning of the period of the church after Jesus' Passion and Resurrection.

Consistent with this schematization, John the Baptist's arrest occurs before Jesus' Baptism, though it is placed later in Mark and Matthew. From the beginning, the rule of the Spirit is a central theme, important in healing, the ministry, the message, and the promise of the continued guidance of the Spirit in the age of the church, pointing toward part two of Luke's work, the book of Acts of the Apostles, in which Pentecost (the receiving of the Holy Spirit by 120 disciples gathered together the 50th day after Easter) is a decisive event.

Just as Luke arranges his Gospel to show the divine plan of salvation in historical periodization, so he orders its structure in accordance with a geographical scheme. Chapter 1 (verse 8) of Acts provides the framework: after the coming of the Spirit, the church will witness in Jerusalem, in all Judaea and Samaria, and then to the end of the inhabited world. These places foreshadow the church's mission. The end of the old dispensation takes place in Jerusalem and its environs. The Resurrection appearances in Luke are placed in Jerusalem (Mark, Matthew, and John point toward Galilee). Jerusalem is also the place of the beginning of the church, and the old holy place thus becomes the centre of the new holy community. The necessity of suffering was made clear and interpreted as the fulfillment of prophecy. Rejection by people from his old home, Nazareth, and by Jewish religious leaders corresponds to the beginning of the ministry to the Gentiles—to the end of the earth.

Luke's interpretation of the Passion

Luke's account of the Crucifixion heightens the guilt of the Jews, adding a trial and mockery by Herod Antipas. The Crucifixion in Luke is interpreted as an anticipatory event: that the Christ must suffer by means of death before entering into glory. Jesus' death, therefore, is not interpreted in terms of an expiatory redemptive act. The centurion who saw the event praised God and called Jesus a righteous man, thus describing his fate as that of a martyr, but with no special meaning for salvation. The link between past salvation history and the period of the church is through the Spirit; salvation history continues in Acts.

THE FOURTH GOSPEL: THE GOSPEL ACCORDING TO JOHN

John is the last Gospel and, in many ways, different from the Synoptic Gospels. The question in the Synoptic Gospels concerns the extent to which the divine reality broke into history in Jesus' coming, and the answers are given in terms of the closeness of the new age. John, from the very beginning, presents Jesus in terms of glory: the Christ, the exalted Lord, mighty from the beginning and throughout his ministry, pointing to the Cross as his glorification and a revelation of the glory of the Father. The Resurrection, together with Jesus' promise to send the Paraclete (the Holy Spirit) as witness, spokesman, and helper for the church, is a continuation of the glorious revelation and manifestation (Greek *epiphaneia*).

Irenaeus calls John the beloved disciple who wrote the Gospel in Ephesus. Papias mentions John the son of Zebedee, the disciple, as well as another John, the presbyter, who might have been at Ephesus. From internal evidence the Gospel was written by a beloved disciple whose name is unknown. Because both external and internal evidence are doubtful, a working hypothesis is that John and the Johannine letters were written and edited somewhere in the East (perhaps Ephesus) as the product of a "school," or Johannine circle, at the end of the 1st century. The addressees were Gentile Christians, but there is accurate knowledge and much reference to Palestine, which might be a reflection of early Gospel tradition. The Jews are equated with the opponents of Jesus, and the separation of church and synagogue is complete, also pointing to a late-1st-century dating. The author of John knows part of the tradition behind the Synoptic Gospels, but it is unlikely that he knew them as literary sources. His use of common tradition is molded to his own style and theology, differing markedly with the Synoptics in many ways. Yet, John is a significant source of Jesus' life and ministry, and it does not stand as a "foreign body" among the Gospels. Confidence in some apostolic traditions behind John is an organic link with the apostolic witness, and, from beginning to end, the confidence is anchored in Jesus' words and the disciples' experience—although much has been changed in redaction. Traces of eyewitness accounts occur in John's unified Gospel narrative, but they are interpreted, as is also the case with the other Gospels. Clement of Alexandria, a late-2nd-century theologian, calls John the "spiritual gospel" that complements and supplements the Synoptics. Although the Greek of John is relatively simple, the power behind it (and its "poetic" translation especially in the King James Version) makes it a most beautiful writing.

Theories of various backgrounds of the Gospel According to John

Various backgrounds for John have been suggested: Greek philosophy (especially the Stoic concept of the *logos*, or "word," as immanent reason); the works of Philo of Alexandria, in which there is an impersonal *logos* concept that can not be the object of faith and love; Hermetic writings, comprising esoteric, magical works from Egypt (2nd–3rd centuries AD) that contain both Greek and Oriental speculations on monotheistic religion and the revelation of God; Gnosticism, a 2nd-century religious movement that emphasized salvation through knowledge and a metaphysical dualism; Mandaeanism, a form of Gnosticism based on Iranian, Babylonian, Egyptian, and Jewish sources; and Palestinian Judaism, from which both Hellenistic and Jewish ideas came. In the last source there is a Wisdom component and some ideas that possibly come from Qumrān, such as a dualism of good versus evil, truth versus falsehood, and light versus darkness. Of these backgrounds, perhaps, all have played a part, but the last appears to fit John best. In the thought world of Jewish Gnosticism, there is a mythological descending and ascending envoy of God. In the prologue of John, there is embedded what is proclaimed as an historical fact: The Logos (Word) took on new meaning in Christ. The Creator of the world entered anew with creative power. But history and interpretation are always so inextricably bound together that one cannot be separated from the other.

In John there is a mixture of long meditational discourses on definite themes and concrete events recalling the structure of Matthew (with events plus discourses); and, although the source problem is complex and research is still grappling with it, there can be little doubt that John depended on a distinct source for his seven miracles (the sign [or *sēmeia*] source): (1) turning water to wine at the marriage at Cana; (2) the healing of an official's son; (3) the healing of a paralytic at the pool at Bethzatha; (4) the feeding of the multitude; (5) Jesus walking on water; (6) the cure of one blind from birth, and (7) the raising of Lazarus from the dead. In chapter 20, verse 30, the purpose of the signs is stated: "Jesus did many other signs in the presence of the disciples, which are not written in this book; but these are written that you may believe that Jesus is the Christ, the Son of God, and that believing you may have life in his name."

The sign source for John

A major part of John is in the form of self-revelatory discourses by Jesus. Some would assign these to a distinct source, but they may rather be the work of the author.

Jesus' coming "hour"—the hour of his glorification—could not come about at any bidding but only according to a divine plan, and Jesus is obedient to it. The Paraclete is promised to come to the disciples, and it is necessary that Jesus go away in order that the Paraclete may come to the church. In John, Christ is depicted as belonging to a higher world, and his kingship is not of this world. He is said to have come into this world to his own people, and they rejected him, but this is but another example

of the church's mission having passed both historically and theologically to the Gentile milieu.

The Christology in John is heightened: though the Synoptics have Jesus speaking about the Kingdom, in John, Jesus speaks about himself. This heightened Christology can be seen in many of the "I am" sayings of Jesus (*e.g.*, "I am the bread of life") in the context of their discourses and accompanying signs. This type of discourse is a concentration in terms and titles of the way in which the Messiah openly reveals his identity by a striking phenomenon: in the Old Testament the association with "I am" is the revelation of the name of God in the theophany (manifestation of God) to Moses (Exodus), and this theophanic interpretation carries over in John.

The "I am" pronouncements

Jesus says "I am" with regard to his function as Messiah, as divine. These sayings are self-revelatory pronouncements: (1) bread of life, (2) light of the world, (3) door of the sheepfold, (4) good shepherd, (5) resurrection and life, (6) way, truth, and life, and (7) true vine. Such theophanic expressions are heightened in other sayings: "I and the Father are one"; "Before Abraham was, I am"; "He who has seen me has seen the Father"; and Thomas' cry after the Resurrection "My Lord and my God."

John 14 is a farewell speech, one of a series, before the Passion. In testament form, it is the bidding of farewell by one who is dying and giving comfort to those he loves. In John, however, the eons (ages) overlap. The significance of the farewell address, thus, is in the teaching that Jesus is God's representative. The fact that he must go to the Father means that the eschatological era already started in Jesus' presence as the Christ and will be intensified at his death and manifested further in the coming of the Spirit to the church. The times shift; the eschatology—here and still to come—also shifts but remains on the whole realized in John, although there is still a tension between the "already" and the "not yet."

John's allegorical thought is shown by his ending of the miracle of Jesus' walking on the sea. The frightened disciples took him into their boat, "and immediately the boat was at the land." This fits the pattern of John's Gospel, namely that, when Jesus is with his church, the new era has already arrived, and, where Jesus is, there is the Kingdom fulfilled. Similarly, the raising of Lazarus in chapter 11 is to demonstrate that the power of the Resurrection, of the fulfilled "eschaton" (last times), is already present in Jesus as Christ now, not only in some future time. Thus, there would appear to be a "realized eschatology" in John; *i.e.*, the last times are realized in the person and work of Jesus. The coming of the Spirit, the Paraclete, however, is still to come, so, even in this most eschatological Gospel, there is a building up, a crescendo, of glorification. In chapter 12, verse 32, Jesus is depicted as saying, "I, when I am lifted up . . . will draw all men to myself"—again an exaltation and glorification that points to the Cross. At the point of death on the Cross, Jesus' words "It is finished" are interpreted to mean that part of the "eschaton" is consummated, fulfilled. After the finding of the empty tomb, there is a Resurrection appearance to the disciples. This includes the "doubting Thomas" pericope, which teaches that those who have to depend on the witness of the Gospel are at no disadvantage.

In an appended chapter, 21, there is a touching story of the Apostle Peter, who, having denied his Lord thrice, is three times asked by Jesus if he loves him. Peter affirms his knowledge that Jesus knows what love is in his heart and is given the care of the church and a prediction that he himself will be persecuted and crucified.

Differences between John and the Synoptics

The numerous differences between the Synoptics and John can be summed up thus: in John eternal life is already present for the believer, while in the Synoptics there is a waiting for the Parousia for the fulfillment of eschatological expectations. This Johannine theology and piety has great similarities to the views that Paul criticizes in I Cor. 15 (see below). The contrast between Paul and John is even more striking if one accepts the most plausible theory that John as we have it includes passages (added later) by which the realized eschatology has been corrected so as to fit better into the more futuristic eschatology that was stressed in defense against the Gnostics. John 5:25–28 is such a striking correction.

The Johannine chronology also differs from the Synoptic. John starts the public ministry with the casting out of the money changers: the Synoptics have this as the last event of the earthly ministry leading to Jesus' apprehension. The public ministry in John occupies two or three years, but the Synoptics telescope it into one. In John Jesus is crucified on 14 Nisan, the same day that the Jewish Passover lamb is sacrificed; in the Synoptics Jesus is crucified on 15 Nisan. The difference in the chronologies of the Passion between John and the Synoptics may be because of the use of a solar calendar in John and a lunar calendar in the Synoptics. Nevertheless, the actual dating is of less importance than the fact that John places the Crucifixion at the time of the Passover sacrifice to emphasize Jesus as the Paschal lamb. There is no celebration of the Last Supper in John, but the feeding of the multitude in chapter 6 gives the opportunity for a eucharistic discourse. Because Jesus is regarded as the Christ from the very beginning of John, there is no baptism story—John the Baptist bears witness to Jesus as the Lamb of God—no temptation, and no demon exorcisms. Satan is vanquished in the presence of Christ. Each of the four Gospels presents a different facet of the picture, a different theology. Although in all the Gospels there is warning about persecution and the danger of discipleship, each has the retrospective comfort of having knowledge of the risen Lord who will send the Spirit. In John, however, there is a triumphant, glorious confidence: "In the world you have tribulation; but be of good cheer, I have overcome the world."

THE ACTS OF THE APOSTLES

The relationship between Luke and Acts

As indicated by both its introduction and its theological plan (see *The Gospel According to Luke*), Acts is the second of a two-volume work compiled by the author of Luke. Both volumes are dedicated to Theophilus (presumably an imperial official), and its contents are divided into periods. In the Gospel, Luke describes first the end of the old dispensation and then the earthly life of Jesus. Near the end of the Gospel, the stage is set for the next period: the "new dispensation" of the church as presented in Acts. After the Ascension of the risen Lord in Jerusalem (Acts 1), there is Pentecost, called Shavuot in Hebrew (*i.e.*, "the 50th day" after Passover). This Jewish festival of the revelation of the Law on Mt. Sinai becomes the day when the Spirit is poured out. For Acts this event marks the beginning of a new era (Acts 2): as in Luke, Jesus, endowed by the Spirit, was led from Nazareth to Jerusalem, so in Acts, the outpouring of the Spirit at Pentecost leads the church from Jerusalem to Rome.

The purpose and style of Acts. Although the title, Acts of the Apostles, suggests that the aim of Acts is to give an account of the deeds of the Apostles, the title actually was a later addition to the work (about the end of the 2nd century). Acts depicts the shift from Jewish Christianity to Gentile Christianity as relatively smooth and portrays the Roman government as regarding the Christian doctrine as harmless. This book is the earliest "church history," viewing the church as guided by the Spirit until a future Parousia (coming of the Lord).

Probably written shortly after Luke (*c.* 85) as a companion volume, in no manuscripts or canonical lists is Acts attached to the Gospel.

Luke edited his history as a series of accounts, and thus Acts is not history in the sense of accurate chronology or of continuity of events but in the ancient sense of rhetoric with an apologetic aim. The author weaves strands of varying traditions and sources into patterns loosely clustered around a nucleus of past events viewed from the vantage point of later development.

The structuring of the material by time and geography may account for the unique way in which both the Ascension of Christ to heaven (40 days after the Resurrection) and the outpouring of the Spirit at Pentecost (50

days after the Resurrection) became fixed and dated events.

The use of speeches and the "we-passages"

The redactor (editor) of Acts composed speeches with primary primitive material within them; about one-fifth of Acts is composed in this way. This manner of using speeches was part of the style and purpose of the work and was not unlike that of other ancient historians such as Josephus, Plutarch, and Tacitus.

In the latter part of Acts are several sections known as the "we-passages" (*e.g.*, 16:10, 20:5, 21:1,8, 27:1, 28:16) that appear to be extracts from a travel diary, or narrative. These do not, however, necessarily point to Luke as a companion of Paul—as has been commonly assumed—but are rather a stylistic device, such as that noted particularly in itinerary accounts in other ancient historical works (*e.g.*, Philostratus' *Life of Apollonius of Tyana*). Though the pronoun changes from "they" to "we," the style, subject matter, and theology do not differ. That an actual companion of Paul writing about his mission journeys could be in so much disagreement with Paul (whose theology is evidenced in his letters) about fundamental issues such as the Law, his apostleship, and his relationship to the Jerusalem church is hardly conceivable.

Acts was written in relatively good literary Greek (especially where it addresses the Gentiles), but it is not consistent, and the Koinē (vernacular) Greek of the 1st century was apparently more natural to the writer. There are some Semitisms, especially when stressing Jewish backgrounds; thus, Paul is called Saul in accounts of his conversion experience on Damascus road. In chapter 17, Paul's speech on the Areopagus, a hill in Athens that traditionally was the meeting place of the city's council, for an intellectual Athenian audience is in good Greek, assimilating Gentile thought patterns, but is expressed in Old Testament universalistic terms.

The content of Acts. The outline of Acts can be roughly divided into two parts: the mission under Peter, centred in Jerusalem (chapters 1–12); and the missions to the Gentiles all the way to Rome (*cf.* chapter 1, verse 8), under the leadership of Paul (chapters 13–28). The earlier sections deal with the Jerusalem church under Peter and the gradual spread of the gospel beyond Jewish limits (in chapters 10–11, for example, Peter is led by the Spirit to baptize the Roman centurion, Cornelius). References to Peter are abruptly ended in chapter 12; James, the brother of the Lord, has become the head of the Jerusalem church, and Philip, a Greek-speaking missionary, is commanded by the Spirit to baptize an Ethiopian eunuch.

Paul's missionary journeys and primitive traditions about Jesus

Paul's missionary journeys are traditionally separated into three: (1) 13:1–14:28; followed by the Council of Jerusalem *c.* AD 49 (15:1–35); (2) 15:36–18:22 with a stop at Antioch; and (3) 18:23–21:14. After that, Paul is imprisoned and sent to Rome where Acts leaves him witnessing openly and unhindered in the capital of the Empire. These journeys may be seen as a part of the writer's "theological geography," because they form one continuous circuit—with stops on the way—between the geographical poles of Jerusalem and Rome. After the Council of Jerusalem *c.* AD 49, the situation was changed, and Paul became the spokesman for the whole Christian mission.

The earliest chapters of Acts contain some primitive traditions important both for any study of the early church and its preaching and for the church's own development of its understanding of itself and of Jesus. After Peter healed a lame man, he made a speech, in chapter 3, in which Jesus is proclaimed as the one appointed but who is now in heaven and who will come as the Christ at the Parousia (Second Coming). In his Pentecost speech in chapter 2, Peter preached that God made Jesus Lord and Christ at his Resurrection.

The titles used for Jesus show both a preservation of primitive tradition and theology and a clear differentiation made by the writer between Jesus in his earthly life (in Luke) and reflection on him in Acts. Christ (Messiah) is consciously used as the title of Jesus; the title Son of man, used frequently in Luke, is used only once in Acts, at the death of the martyr Stephen, when he is granted a vision of the Lord in glory. Early titles, "servant" and "righteous one," reflect the Old Testament background of God's "suffering servant." The Hellenistic term saviour (*sōtēr*) is used in Acts in chapters 5 and 13. The more primitive Christologies and titles show not only a flexibility of traditions but also the functional nature of New Testament Christology.

Differences between views of Paul in Acts and in Paul's Letters

Acts presents a picture of Paul that differs from his own description of himself in many of his letters, both factually and theologically. In Acts, Paul, on his way to Damascus to persecute the church, is dramatically stopped by a visionary experience of Jesus and is later instructed. In his letters, however, Paul stated that he was called by direct revelation of the risen Lord and given a vocation for which he had been born (recalling the call of an Old Testament prophet, such as Jeremiah) and was instructed by no man.

The account of Paul's relation to Judaism in Acts also differs from that in his letters. In Acts, Paul is presented as having received from the Jerusalem apostolic council the authority for his mission to the Gentiles as well as their decision—the so-called apostolic decree (15:20; *cf.* 15:29)—as to the minimal basis upon which a Gentile could be accepted into fellowship with Jewish Christians. According to this decree, Gentile converts to Christianity were to abstain from pollutions of idols (pagan cults), unchastity, from what is strangled, and from blood (referring to the Jewish cultic food laws as showing continuity with the old Israel). Circumcision, however, was not required, an important concession on the part of the Jewish Christians.

In Acts Paul is not called an Apostle except in passing, and the impression is given, contrary to Paul's letters, that he is subordinate to and dependent upon the twelve Apostles. When Paul entered a new city, he went first to the synagogue. If his message of the gospel was rejected, he turned to the Gentiles. According to Paul's missionary practice and theology, the message had first to be spoken to the Jews as a reminder that Christianity is grounded in redemptive history; this prevents the connection with the old Israel from being forgotten. Because most Jews rejected Paul's message, the author proclaimed that salvation thus passed to the Gentiles.

Paul's relationship to the Romans and the dominance of the Holy Spirit

Roman authorities are depicted as treating Paul (and other Christians) in a just manner. The author repeatedly stressed that the Roman authorities did not find fault with the Christians but rather viewed Christian–Jewish antagonisms merely as one problem among Jewish factions. While in Corinth, during a conflict with the Jews, the Roman proconsul of Achaea in Greece, Gallio, refused to hear the charges brought against Paul because, according to Roman law, they were extralegal. On a later occasion in Ephesus, during a conflict with the silversmiths who derived their income from selling statuettes of the goddess Diana, Paul was protected from local antagonisms and a riot by Roman authorities. Toward the end of his career, after having been in the protective custody of the Judaean procurator Felix, Paul was heard by Felix's successor, Festus, and the Jewish king Agrippa II, and, had he not appealed to Caesar as a Roman citizen, he could have been set free. He thus had to go to Rome to be tried, and that is the last that is heard about him in Acts.

The doctrine of the Holy Spirit is a dominant theme in Acts, as it is in the Gospel According to Luke. Just as Jesus started his public ministry in Luke by reading from the Book of Isaiah: "The Spirit of the Lord is upon me . . ." so also in Acts the new age of the Spirit began at Pentecost, which is viewed as the fulfillment of the prophecy of Joel that in the new age the Spirit would be poured out on all men. That persons from many nations heard in their own tongues the mighty works of God has been viewed as a reversal of the Tower of Babel narrative, with languages no more confused and people no longer scattered.

Although Peter, Stephen, and Paul are central figures in Acts, the piety of the humbler members of the church also permeates the book. Church structure and organization, with apostles, disciples, elders, prophets, and teach-

ers, exhibits great fluidity. Paul, in bidding farewell at Miletus to the elders from Ephesus, exhorted them to "take heed . . . to all the flock in which the Holy Spirit made you guardians (bishops) to feed the church. . . ." Offices may be conveyed by prayer and laying on of hands but there is little stress on distinction of office or succession, thus indicating a very early period in the life of the church.

Because Peter "departs and goes to another place" and Paul is left under house arrest awaiting trial, the readers appear to be left in suspense concerning the fates of these two leaders. The readers, however, probably knew what had happened to them—*i.e.*, that these Apostles had eventually been martyred sometime in the 60s before Acts was written. What is more, the interest in Acts is not in the fates of Peter and Paul; the gospel has finally reached Rome, the center of the *oikoumenē* ("the inhabited world"), and thus the ending is suitable to the book—Paul is left "preaching the kingdom of God and teaching about the Lord Jesus Christ quite openly and unhindered."

THE PAULINE LETTERS

In the New Testament canon of 27 books, 21 are called "letters," and even the Revelation to John starts and ends in letter form. Of the 21, 13 belong to the Pauline corpus; the Letter to the Hebrews is included in the Pauline corpus in the East but not, however, in the West. Three letters of this corpus, the Pastoral Letters, are pseudonymous and thus are not considered here. Of the remaining 10, the Letters to the Colossians and Ephesians are from the hand of a later Pauline follower and II Thessalonians is spurious. How this Pauline corpus was collected and published remains obscure, but letters as part of Holy Scripture were an early established phenomenon of Christianity.

The church was poor and widespread, and, at least in the early stages, expected an imminent Parousia. A body of more formal sacred writings was thus superseded in importance by letters (*e.g.*, those of bishop Ignatius of Antioch) that answered practical questions of the early churches.

The letters of Paul, written only about 20–30 years after the crucifixion, were preserved, collected, and eventually distributed. In general, they answered questions of churches that he had founded. When all the Pauline Letters as a corpus were first known is difficult to determine. Because Pauline theology and some quotations and allusions were certainly known at the end of the 1st century, the Pauline Letters probably were collected and circulated for general church use by the end of the 1st century or soon thereafter. A disciple of Paul, possibly Onesimus, may have used Ephesians as a covering letter for the whole collection.

The significance of the Pauline Letters to the addressees

The letters Galatians and Romans both contain an extensive discussion about the Law (Torah) and justification (in language not found in the other letters) to solve the problem of the relation of Christianity to Judaism and of the relationship of Jewish Christians with Gentile Christians. Galatians is older and differs from Romans in that it deals with Judaizers—*i.e.*, Gentile Christians who were infatuated with Jewish ways and championed Jewish ceremonial law for Gentile Christians. On the other hand, Romans speaks to the question of the Jews and the Christian faith and church in God's plan of salvation.

In I and II Corinthians (which may include fragments of much Corinthian correspondence preserved in a somewhat haphazard order), there is no preoccupation with either Jews or Judaizing practices. They deal with a church of Gentile Christians and are therefore the best evidence of how Paul operated on Gentile territory.

The earliest book in the New Testament is I Thessalonians, which is concerned with the problem of eschatology. Though II Thessalonians is obvious in its imitation of the style of I Thessalonians, it reflects a later time, elaborates on I Thessalonians, and is thus not viewed as genuine.

Philippians may be a composite letter in which various themes of Pauline teaching are held together by a testament form. Thus, it is a compendium without too specific a focus on the Philippian situation. Philemon, although addressed to a house church, is uniquely concerned with the fate of a slave being returned to his master, with the hope that he will be forgiven and be sent back to help Paul in prison, an example of manumission in Paul's name.

Ephesians appears to be dependent on Colossians, and both, although using the Pauline style, reflect a time and imagery sometimes different from and later than Paul's genuine letters. Ephesians covers the content of Colossians in more compact form and may be a covering letter for the entire Pauline corpus by a disciple or other later Paulinist.

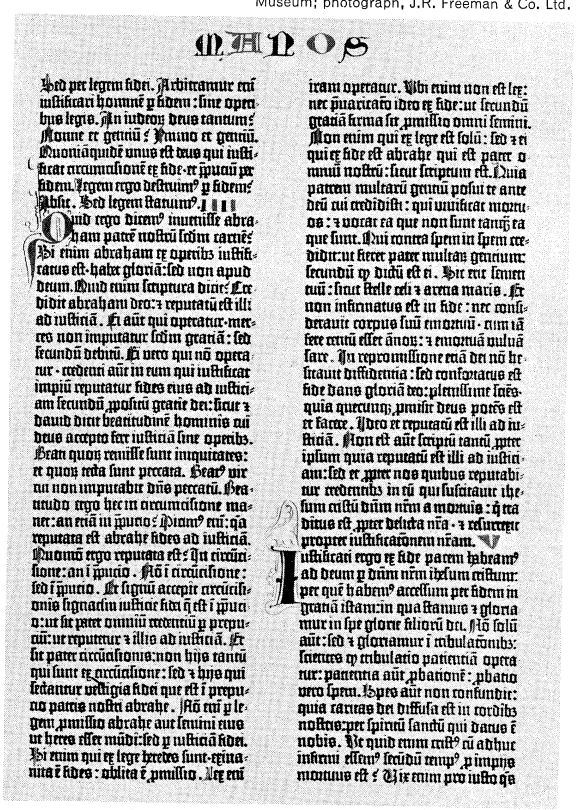
The style and form of Paul's letters

The style of Paul's letters is an admixture of Greek and Jewish form, combining Paul's personal concern with his official status as Apostle. After his own name, Paul names the addressees or congregation being addressed and adds "grace and peace." This is often followed by thanksgivings and intercession that are significantly adapted to the content and purpose of the letter. Doctrinal material usually precedes advice or exhortation (*parenesis*), and the letters conclude with personal news or admonition and a blessing: "The grace of our Lord Jesus Christ be with you." Paul's letters were probably dictated to an amanuensis (who might be named, for example, Sosthenes, I Cor. 1:2), and some greetings were written at the end of the letters in his own hand. They were obviously meant to be read aloud in the church, however, and thus their style is different from that of purely personal letters.

The Letter of Paul to the Romans. Romans differs from all the other Pauline letters in that it was written to a congregation over which Paul did not claim apostolic authority. He stressed that he was merely going to Rome in transit, because it was his principle not to evangelize where others had worked. Because his apostolic ministry appeared to be completed in Asia Minor and Greece, Paul planned to go to Spain via Rome, a city that he had never visited. Before going westward, however, he first had to go to Jerusalem to deliver to the church there a collection of money.

The reason for Paul's Letter to the Romans

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Chapter 4 and the opening of chapter 5 of the Letter of Paul to the Romans from a facsimile of Gutenberg's 42-line Bible. In the British Museum.

Because Paul was going to a church he had not founded, his writing to the Roman Christians offered him an opportunity to present his theological views in a systematic way, which he had not done in other letters. Paul reflected on how his special mission fitted into God's plan for the salvation of mankind, of both Jews and Gentiles —a theme that reached its climax in chapters 9–11. Chapters 1–8 unfold with great specificity how the coming of Jesus the Messiah has made it possible for the Gentiles to become heirs to God's promises. His argument is at first negative, stating that neither Gentile nor Jew could effect his own salvation. He then shows a new way in which eventually both can be delivered from the bondage of sin by being justified—*i.e.*, made "right with God"—not through acceptance of the Law but by faith in the crucified Lord.

The theological section (chapters 1–11) is followed (as is often the case in Pauline letters) by ethical instructions. There is little doubt about the integrity of Romans 1–15; chapter 16, however, seems to be a later addition. The letter was written from Corinth *c.* 56. It contains numerous salutations to individuals (which is unusual in that Paul had never been to Rome) and an antinomian (antilegalistic) tone that would be more appropriate to the situation in Asia Minor. The doxology (16:25–27) is rhetorical and its vocabulary is not in keeping with that of Paul's usual thought. Because the doxology occurs in different manuscripts in varying positions in the course of textual transmission, it is probably secondary. Chapter 16 may thus preserve portions of a letter or letters from some other time or to some place other than Rome, possibly Ephesus.

The main theme of the Letter to the Romans

In chapter 1, verses 1–17, there are greetings and thanksgivings leading to the main theme of the letter: the gospel is

> the power of God for salvation to every one who has faith (*i.e.*, that Jesus is the Messiah), to the Jew first and also to the Greek. For in it the righteousness of God is revealed through faith for faith; as it is written, "The righteous shall live by faith."

Paul took this sentence from the Old Testament Book of Habakkuk, chapter 2, verse 4, not as a principle but as a prophecy now fulfilled. Thus, the translation should read "will live" rather than "shall live." This does not refer to God's faithfulness but rather to the believer's trust. Justification by faith is not, however, the answer to the question of man, plagued by conscience, about his salvation nor is it deep theology. It is rather an argument totally grounded in the problem of the relationship of Jews and Gentiles—*i.e.*, how it will be possible for the Gentiles to be fellow heirs with Jews and how both Jews and Gentiles can be members of the church. In chapters 2–3 both Gentiles and Jews are demonstrated to have fallen short of the glory of God and to be under condemnation. A turning point, however, is emphasized in chapter 3: "But now the righteousness of God has been manifested apart from law. . . ." Justification is a gift through Jesus Christ and his expiating death for the salvation and vindication of all who believe in him. Because all this is through Christ and not by works of the Law, salvation is equally available to the Gentiles as well as to the Jews. For both, the means is the same: faith in Jesus the Christ.

The central problem after chapter 8, which describes the glory of the new dispensation in Christ and the Spirit (presented in chapters 9–11), centres on the mystery revealed to Paul, namely, that the Gentiles should be incorporated and be fellow heirs with the Jews. This is what Paul yearned for with respect to his fellow Jews. What makes it equally possible for Jew or Gentile to come to Christ is justification by faith, with the Law viewed as obsolete because Christ is the end of the Law (chapter 10, verse 4). Thus, there are, in effect, no distinctions between Gentile and Jew. Paul viewed his ministry as having made possible the inclusion of the Gentiles; as an apostle to the Gentiles he never urged them to carry on a mission to the Jews. He envisaged the Jewish acceptance of Christ as a mystery beyond human planning and effort, a divine event that will be the climax of history.

The ethical section (12:1–15:13) has no special reference to a situation in Rome. A close analysis shows that Paul here repeats thoughts and admonitions that are more specific in other letters. A metaphor of the church as a body (12:5), for example, is stylized and compressed as compared with the fuller use of the same in I Corinthians, chapter 12 and the pattern of weakness and strength in matters of food is best understood in the light of the fuller exposition in I Corinthians, chapters 8 and 10.

The First Letter of Paul to the Corinthians. This letter is part of Paul's correspondence with the Corinthian congregation founded by him and composed of Gentile Christians. The problems of Galatians and Romans, written to Christians with Jewish and Roman legal concepts, are different from those of I Corinthians, and, thus, the justification language is absent.

Except for the brief communication with Philemon (see below), I Corinthians is the most specifically practical, situation-oriented of Paul's letters. No other Pauline letter is so directly devoted to the consideration of practical and theological problems, many of them apparently communicated by the congregation through correspondence or by delegations. The letter, therefore, does not tend to stand as a unit and it is not uniform in its treatment of the varying situations.

Emphasis on practical and theological problems

Literary criticism—or redaction—has traditionally split the letter into several fragments with a presumed historical development within a relatively short period in the Corinthian church. Paul's reference to a previous letter of his in chapter 5, verse 9, has been the object of scholarly efforts to restore the earlier letter. The fragmentary and not-too-uniform nature of both I and II Corinthians, however, precludes much probability of success in such searches.

Writing from Ephesus *c.* 53 or 54 upon hearing from a certain Chloe's people that the church was rent by party factions, Paul tried to bring unity to the congregation. Whether these factions actually represented outside interference (*e.g.*, Cephas [Peter], Apollos, or others) or were factions of the congregation under the influence of a widespread heresy of the time is a question perhaps best answered by the fact that the factions do not come up again after I Corinthians, chapter 1 and that I Corinthians, chapter 3 reduces the factions to Apollos and Paul, who claims he is head of no party. The Christ "party"—*i.e.*, those who claim no party at all—(1:12; *cf.* 3:23) may be the only "party" Paul advocated because Christ is not divided. Paul warned that Christians should not fashion themselves into parties under various leaders, because all these leaders are servants of Christ and stewards of the mysteries of God through whom Christians come to belief. The church is not a society with competitive philosophical schools.

The cause of the difficulties in Corinth: proto-Gnosticism

The letter is a response to difficulties caused or increased by a relatively strong group in Corinth that may be described as "enthusiasts." This group of enthusiasts may have been proto-Gnostics (early religious dualists not yet organized into definite sects). The Corinthian enthusiasts did, however, have some characteristics that would later be found in 2nd–3rd-century Gnosticism: a belief in salvation through spiritual knowledge or wisdom communicated by a revealer (not a redeemer); an otherworldliness that could lead either to licentiousness (scorn) or asceticism (withdrawal); and a basically dualist and deliberately syncretistic system of beliefs using the mythical speculations and magical ideas of their time.

The Corinthian problems might well be traced to such enthusiasts. Their *gnōsis* ("esoteric knowledge") was a religious knowledge that gave them the feeling of superiority over more pedestrian Christians. This *gnōsis* Paul identified as false wisdom. In chapter 14 Paul describes the views and related practices of those maintaining that they have spiritual gifts of inspiration, especially speaking in tongues (glossolalia) and *gnōsis*. Such enthusiasts prized eloquent or secret wisdom; they sought a revealer who had come into the world hidden from the evil powers and known only to those, the *pneumatikoi*, or the spiritual elite, who recognize him; and they tolerated gross immorality by claiming anything to be lawful for them (es-

pecially their slogan quoted by Paul: "for me all things are lawful"). These enthusiasts also rejected marriage because it furthered the propagation of the present evil world; they claimed to possess knowledge that made them indifferent to the world; and they believed that their salvation was guaranteed by ritual and rites. Though they prized spiritual gifts, they scorned the ordinary Christian services for the community; and they did not believe in a future resurrection of the dead, which in their system had no place or was nonsense.

The concept of love and views on the resurrection

The main Pauline answer (*e.g.*, as emphasized in chapter 13) was that love, namely concern for the building up of the community, surpasses all knowledge or spiritual gifts and that love is a corrective because it demands service, edification (*i.e.*, building up) of the church, and involves Christians with one another. Those Corinthians whom Paul viewed as opponents emphasized *gnōsis* over against love. The discussion of the resurrection in chapter 15 sheds further light on this. The opponents did not deny the Resurrection of Jesus Christ about which there was common agreement, but rather they debated about the future resurrection of Christians from the dead. Their view was perhaps similar to that reported as heresy in II Timothy, chapter 2, verse 18—*i.e.*, the believer already had eternal life and that a future resurrection of the body was meaningless. In holding such a view, Paul's opponents claimed they were faithful to the received kerygma (proclamation).

Another indication that some Corinthians had no disagreement with tradition but interpreted it too enthusiastically is found in I Corinthians, chapter 11. The liturgical formula pertaining to the Lord's Supper is sound:

> The Lord Jesus on the night when he was betrayed took bread, and when he had given thanks, he broke it, and said "This is my body which is for you. Do this in remembrance of me." In the same way also the cup, after supper, saying, "This cup is the new covenant in my blood. Do this, as often as you drink it, in remembrance of me." (11:23–25.)

In a discussion of the sacraments in chapter 10, however, the enthusiasts probably believed in a rather magical efficacy of Baptism and the Eucharist, though Paul qualified such an interpretation and took exception to it. The misunderstanding of the enthusiasts points to a special reinterpretation of Scripture and tradition (which resembles that of the 1st-century Jewish philosopher Philo and also the later Gnostics)—taking Scripture, tradition, and liturgical practices as effectively bringing about an otherworldly, spiritual reality immediately for those who really understand (*i.e.*, those who have *gnōsis*). Paul also criticized these spiritualists for their disregard of the poor members of the congregation, who found no food left when they came from their work.

The question of freedom and unity

Discussions about Christian and apostolic freedom (in chapters 5, 6, 7, 9, and 11) and also a discussion about being free to eat meat that had been sacrificed to idols and leftovers of pagan sacrifices sold in the marketplace were caused by conflicts with the enthusiasts who paraded their spiritual freedom, strength, and superiority at the expense of their weaker brothers in the faith, who were not ready for this freedom. A shift in the discussion in chapter 12 (the body and its members are equal in Christ) —from a very speculative idea of the body of Christ to a more metaphorical one reminiscent of Stoic philosophical ideas about society as an organism—can best be understood if it is assumed that the enthusiasts actually pressed for a mythical understanding of Christianity, in which one became literally incorporated into Christ, otherworldly, and divine. Paul added some qualifications that brought the church into concrete everyday life and even provided a source of political reality. A somewhat drastic understanding of spiritual gifts presupposed and criticized by Paul in chapters 12–14 fits well into such a pattern.

Permeating all the discussion of individual topics in I Corinthians is the theme of Christian unity and edification, a topic introduced and underscored in the preface and thanksgiving of this letter and in its introduction. Such unity is defended as being very inclusive, real, and concrete—as over against the enthusiastic attempt to speak in terms of spiritual reality and achievement, in which the true life of the spirit is only for the few (*i.e.*, the Gnostic elitists).

Paul viewed the necessity of unity in the wisdom of God as it is evinced in the scandal of the cross. In order to deflate the exalted and to make foolish the destructive (speculative) wisdom established by men, God showed his wisdom in the "foolishness" of Jesus' crucifixion. Here, although hidden, is God's true wisdom. The opponents hailed their ideal teachers as bringers of hidden wisdom. To this Paul said that it is Christ who is the Wisdom.

Ethical questions and views on marriage

In chapters 5 and 6 Paul dealt with certain ethical scandals and difficulties in the congregation: incest and fornication; the use of pagan courts for settling disputes among Christians; traffic with prostitutes—all for the demonstration of Christian "freedom." These wrongs might have been the direct or indirect consequences of the spiritual "powers" of the enthusiasts. According to Paul, however, such immorality was impossible for the Christian because of the concreteness of his allegiance to Christ and of inspiration (with the idea of the body as the temple of the Holy Spirit).

Because Paul expected an imminent Parousia (Second Coming of Christ), he suggested (chapter 7) the unmarried state as the preferable one, but conceded that marriage can prevent fornication. Paul even advised against breaking up mixed marriages between baptized Christians (both Jews and Gentiles) and unbaptized Gentiles. He advocated the practice of ascetics living together as "virgins," male and female, although he took this as a strain that is hard to bear and thus suggested marriage in unbearable cases. Not only the imminence of the Parousia but also radical change ("the form of this world is passing away") caused Paul, on the whole, to affirm the social status quo—whether it concern circumcision, slavery, or other matters. Everybody is advised to remain—for the short time ahead—in the state in which he finds himself. Such eschatological fervour caused Paul to argue against any worldly anxiety, fear, or worries stemming from them. This is reflected in the ethical criterion of possessing things as though one did not have them.

In chapter 9, Paul used his own conduct, in contrast to that of the enthusiasts who flaunted their freedom in such a way that it often had destructive influences, as a paradigm for an understanding of responsible freedom. Here he showed by various examples from his own life style that he had never made use of his rightful privileges to the fullest, that he has, rather, been guided by what serves the weaker brothers and sisters. It is in this sense that he subdued his body and that he urged the spiritual "snobs" to imitate him.

Discussions on corporate worship

In chapters 11–14, Paul turned to problems of corporate worship. Paul did not question the right and ability of prophetically gifted women to make inspired statements in Christian worship, but he pointed out that women need protection. Arguments about a veil or long hair for a woman are in the context of the church's worship before God himself, in which the congregation worships in the presence of the angels. Paul stressed the subordination of women in chapters 11 and 14; they are forbidden to speak in worship. In chapter 14 Paul stated (perhaps) a general principle that would allow for exceptions in cases of clear prophetic inspiration of women (*cf.* however, Galatians, chapter 3, verse 28).

In discussion of proper restraint and mutual regard in celebrating the Lord's Supper, Paul seemed to presuppose a prior common meal (possibly an agape meal) as part of the eucharistic celebration. This common meal, however, had apparently been devalued because of the interest of the enthusiasts in the sacrament itself. As a result, the communal aspect showed up social differences in the community; and some brought ample food, whereas others, of lower station, had nothing. In view of this, Paul again used the criterion of love and suggested that people eat their meal at home and then come together, being sensitive to each other's needs. The Lord's Supper would then be what it is, a proclamation of the death of Christ in anticipation of his return; mutual and corporate concern and responsibility thus become a part of the Eucharist.

Similarly, mutual edification and love are linked in chapter 13 as the appropriate centre of the discussion of spiritual gifts, manifested particularly in public worship (chapter 14).

The emphasis on the communal aspect of the church is continued in chapter 15. Paul did not dwell on his own vision of Christ nor on his role in founding the church at Corinth but rather argued for the resurrection of all as a future experience, not as though each person had already had this experience. Paul viewed the resurrection as a collective phenomenon in the expectation of an end-time resurrection from the dead, with Christ as the first fruits of those who have died.

That love is to extend beyond the immediate community and be shared with all the saints (members of the church) is demonstrated in chapter 16, the closing chapter, by the collection for the Jerusalem church. The keynote might be: "Let all that you do be done in love." The final passage—including the cry: "Our Lord, come!"—may reflect or repeat a eucharistic formula or setting.

The Second Letter of Paul to the Corinthians. This letter, as is I Corinthians, is composed of a collection of fragments of Paul's correspondence with the Corinthians about a year later (*i.e.*, *c.* 55) from Macedonia. The diversity of I Corinthians was caused by the variety of problems discussed, but the diversity of II Corinthians was the result of a reflection of the underlying, rather turbulent history of Paul and his congregation. A pattern of fragments that make up II Corinthians can be understood in terms of a development that can be reconstructed. Gaps and editorial seams in this pattern are more recognizable and abrupt than those in I Corinthians, and a more original order for II Corinthians can be restored by fitting together blocks of material that obviously belong with one another in terms of context and unity of thought.

Though historical settings can be reconstructed with a high degree of validity to account for the fragments of II Corinthians, later editorial processes account for the order in which the fragments appear in the letter as it is now written. Based both on internal and external evidence, II Corinthians probably was later than I Corinthians, which was written after Paul's first trip to Corinth. Not long before the composition of II Corinthians, Paul was in mortal danger in Asia and travelled to Macedonia, where he remained.

New problems within the Corinthian congregation

New apostles and heresies had apparently invaded the Corinthian congregation and Paul sent his companion Timothy to try to bring them back to the true gospel as Paul had preached it. This mission was apparently unsuccessful, and Paul, in chapters 2 to 7, wrote to the church with a defense of his apostolic office, still counting on the loyalty of the Corinthians. His letter apparently did not change things, and there is some dispute as to whether Paul himself made an intermediate second visit to Corinth that was abruptly cut short by conflict with a member of the Corinthian church who violently opposed him. He considered such a second visit, but, according to chapter 2, verse 4, and chapters 10 to 13, he sent Titus to Corinth with a strongly polemical "letter of tears" and anxiously awaited his return, going from Troas to Macedonia to meet him.

Paul had almost been in despair over the Corinthians, but Titus and the letter seemed to have restored the Corinthian church to order. Titus and some companions were then sent to take up the collection for the church at Jerusalem, a sign of Christian mutual love and unity. He took with him Paul's "letter of reconciliation" written from Macedonia, which can be noted in chapter 1, verse 1, to chapter 2, verse 3; chapter 7, verses 5 and 6; and chapter 8. In chapter 8 the Macedonians are held up as an example of generosity. A similar section regarding the collection is in chapter 9, and the Achaeans (and probably their capital city, Corinth) were cited as an example to the Macedonians for generous giving. This was probably sent shortly before Paul's third (and last) visit to Corinth. From Corinth Paul wrote to the Roman church a letter that shows no sign of difficulties with the Corinthians and that presumed the conveying of the collection to Jerusalem.

If the Corinthian controversy had been smoothed out, a question is raised as to why II Corinthians ends in the "letter of tears" rather than in the "letter of reconciliation." This may be understood if the literary order of the several sections was arranged by a redactor who collected the fragments probably in the last decade of the 1st century. The redactor may have used a "form" amply illustrated in Christian writings of the late 1st and early 2nd century; one of the end-time expectations was that "false prophets would show signs and wonders to lead the elect astray," and chapters 10–13 deal with "false prophets" and "servants of Satan." Such warnings were placed at the end of writings of that time.

The problem of the "letter of tears" and the "letter of reconciliation"

Several abrupt editorial seams that resulted from an arrangement of a letter of reconciliation, an apology on the nature of Paul's apostolic authority, a polemic against opponents, two letters concerning the collection, and a possible non-Pauline insertion (in chapter 6, verse 14, to chapter 7, verse 1) can thus be understood. The reconciliation of chapters 1 and 7 is hardly in agreement with Paul's elaborate defense of his ministry in chapter 2. Even more jarring to such a reconciliation is the polemic of chapters 10–13. These latter chapters are viewed as a substantial fragment of Paul's "letter of tears," after which the Corinthians disengaged themselves from outside agitators and caused them to leave. Such opponents, who are mentioned in chapter 11, verse 4, and who tried to attract the congregation away from Paul's ideas, were probably Hellenized Jewish Christians from Palestine.

The problem of outside agitators

The outside agitators (who provoked the response of chapters 10–13) probably were Christians who imitated the Hellenistic-Jewish missionaries and had developed an elaborate propagandizing missionary theology and practices analogous to the missionary movements in the pagan world. Their goal was to prove the spiritual power of their own religion in conscious and aggressive competition with other religions, thus hoping to attract others and convert them to Christianity.

The major criteria for successful competition were affinity or identity with the ancient Mosaic traditions and objective manifestations of the current power of that tradition in the form of miraculous demonstrations. The link between the ancient traditions and the current careers of the itinerant missionaries was the record of Jesus as understood from the miracle stories of the Gospels—a demonstrated epiphany of the powers of the Spirit. These missionaries were seen as "divine men," as were the heroes of old. Their miracles were to be imitated. Such traditions about Jesus as a wonder-worker might have been used by Paul's opponents, with over-emphasis on such works as criteria of power.

Paul's attacks on the "super-apostles"

That which Paul attacks as "bragging" or "boasting," particularly the preaching of the so-called "super-apostles," in chapter 11, verse 5, was probably understood by his opponents as no more than faithful testimony to, and a demonstration of, the spiritual powers of tradition as they perceived it in their own experiences. To them faithfulness to Jesus was primarily the acknowledgment of Jesus' being the most powerful "divine man" and, secondarily, their establishment and maintenance of relationship to him through imitation in their powerful demonstrations and wondrous acts.

Paul (who in I Corinthians, chapter 1, had advocated the dialectic of the cross) would thus be discredited by miracle-working men like the opponents in II Corinthians. Paul's credibility and validity as an Apostle came into question along with his Christology, which was a "theology of the cross." Confronted with the challenge of the powerful "super-apostles," Paul's message could be distorted as hiding his own inability or incapacity—an apostle who dared not take money because, being an ineffective speaker and a weak person, he had nothing for which to ask payment. His defense was Paul's first attempt to deal with these new problems caused by invading opponents who had undercut his authority.

Paul's defense of his own apostleship

Paul centred his defense around the issue most debated; true apostleship and his own sufficiency. Because he derived his ministry from God himself as a servant preaching not himself but Jesus Christ as Lord, no "peddler of

God's word" or selling or recommendation is called for, but only the living record—*i.e.*, the people brought to believe in Christ. Paul quickly alluded to his own weakness and "carrying in the body the death of Jesus, so that the life of Jesus may also be manifested. . . ." (chapter 4, verse 10). Paul found his weakness one of the things that made him one with the Lord and that made his ministry a true ministry of Jesus Christ, who was crucified through weakness but lives by the power of God—as does his true apostle. This weakness seems to refer to a physical handicap of Paul's (epilepsy?), the "thorn in the flesh" that interfered with his travel plans.

Paul placed his own apparent weakness, in which he proclaimed that God had manifested himself, against the boastings of the "super-apostles." Unlike them, he strikes a non-heroic note. It is confidence in the power of Jesus' Resurrection that produces glory for the Gospel message and final (eschatological) reward and recognition for the Apostle.

Though Paul may himself sound "enthusiastic," his statements are made with a realistic assessment of the world, as demonstrated not least in the sufferings of Paul himself. Emphasis on God's act of grace, however, makes Paul urge the Corinthians to accept him and to reach out to the promise of God's salvation even in the present.

Paul's defense of his apostleship and a following visit did not succeed. Agitation from outside opponents apparently increased and solidified. The "letter of tears" reflects this situation. Paul revealed himself personally, coming close to autobiographical statements. Paul spoke of himself only with theological purpose and as part of his tactical argument with his opponents concerning attitudes and conduct. His point was that a style of life is a reflection of an underlying theology. He demonstrated to his opponents that his work for the church is constructive, and that though he boasted of his ministry, he boasted only "of the Lord," of the work Christ had done through him.

Paul's use of the technique of the "fool's speech"

In his so-called fool's speech, in which he blatantly asked the Corinthians to "bear with me in a little foolishness," Paul adopted the technique of the mime of the street theatres of his times, consciously drawing on the laughter and mockery of his audience, but then he successfully reversed the scene and made his audience realize that in laughing at him they mocked themselves, thus revealing the perversion of their criteria of superiority. Paul used metaphorical images, identifying the congregation with the bride, Jesus as the bridegroom, himself as the best man, and Satan (the opponents) as the adulterer. The plot assumed a successful seduction, and the best man who recommended the bride stands disproven. Paul then pretended to try to shift this balance by bragging about himself and scolding both seducers and the seduced. He accepted no inferiority to the opponents—the seducers ("super-apostles")—and claimed that they preached another Christ than the true Christ and brought another spirit and that he would accept no support from the church that was led astray.

In chapter 11, Paul continued to boast "as a fool," claiming to have all the qualifications of his opponents, but that he was more truly a representative of Christ. This he explained ever more intensely in an ironic and almost sarcastic trend in the dialectic of the so-called fool's speech. He boasted not of strength but of weakness—though he could boast of ecstatic experience as his opponents had—and that he had learned through bitter experience (possibly a chronic illness) that he must not exalt himself, but rather that he has been told through a word of Christ that his power is made perfect in weakness. In the enumeration of his qualifications, Paul has jested "as a fool" concerning his suffering, visions, miraculous heavenly travels, and oracles. Yet, it is clear that through Christ these modes of experience and communication have been transformed. Thus, Paul establishes that he is a true apostle and not inferior to the "super-apostles."

Paul expressed his intention of visiting the congregation and told them that he desired to come not as a judge but as a father. Neither he nor Titus had or would deceive or take advantage of them. At this, the end of the "letter of tears," Paul announced his possible third visit and revealed a definite fear that he might be forced to act as a judge of the congregation, which was increasingly falling away from the apostolic gospel. Paul, however, still hoped that reconciliation might be accomplished, that truth would prevail, and that his authority could be used for building up rather than destruction. He exhorted the community to keep peace and blessed them.

Paul's intention to visit the congregation again

The "letter of reconciliation," found in chapters 1, 2, and 7, assumed that Titus had returned with good news of the Corinthians, their eagerness to prove that they had amended their ways. Paul responded with a report of the consolation this had brought him and of the grave danger he had escaped (in prison in Ephesus). He exhorted the church at Corinth to remember the Christian message in love—of Paul for them and of the congregation for him. The shadow between Paul and the Corinthians had been dispersed, and Paul reaffirmed his constant and continuous concern for them and God's love in Christ manifest in Baptism and the gift of the Spirit. Paul interceded for a man who had offended him and forgave him. Paul then told the Corinthians of his eagerness for Titus' news of them that occasioned his special trip to Macedonia. This news brought joy and consolation; therefore, Paul urged the Corinthians again to forgive the man who had offended him.

Fragments of two letters concerning the collection for Jerusalem, a sign of unity of the church (chapter 8 especially being close to the "letter of reconciliation" and chapter 9, a fragment probably later than chapter 8), are signs that Paul's relation to the Corinthians again became close and joyful. The collection was a bond of mutual and reciprocal relationship that reached its climax in thanksgiving and praise of God. For the whole church he exclaimed: "Thanks be to God for His inexpressible gift!"

The question of the relationships between Jewish and Gentile Christians

The Letter of Paul to the Galatians. Paul's Letter to the Galatians is a forceful and passionate letter dealing with a very specific question: the relation of Jewish Christians and Gentile Christians in the church, the problem of justification through faith not works of the Law, and freedom in Christ. Paul probably wrote from Ephesus *c.* 53–54 to a church he had founded in the territory of Galatia in Asia Minor.

This congregation had been "unsettled" since his last visit to Galatia. Gentile Christians, Judaizers who were fascinated with Jewish customs and festivals and who asserted that Gentiles must adhere to the Law, the Torah, had attempted to undermine Paul's message and effectiveness. The Judaizers believed that Gentile Christians should be circumcised and keep the Jewish food laws. There were probably some Jewish Christians in this church, but the majority were Gentile Christians. Paul attacked the Judaizers vigorously by defending his own call and the independence of the revelations of his personal apostolate. This is supported by reports of agreement between him and the Jerusalem church and by argument from Scripture. In these, he proved that the Law was given only a limited role in the total history of salvation. The letter ends with Paul pointing out that through the Spirit the Christian in faith is admonished to good behaviour and brotherly love. He admonishes faith in the cross of Christ, wishes peace upon his followers, and prays for mercy on Israel.

This Pauline letter is the only one without either kindly ingression, thanksgiving, or personal greetings appended to the final blessing. It is very specific in dealing with the problems concerned. In chapter 1, an account of Paul's call, he defended his apostolic office, having received it directly from God in the revelation of Christ. He provided autobiographical data concerning his former persecution of the church and zeal in his Jewish tradition. He referred to his call on the model of that of the Old Testament prophets called by God in order that they may serve him and said that his mission had been revealed to him to be the apostle to the Gentiles. Paul viewed himself as being chosen to be an instrument to take the message of God and Christ to the Gentiles,

a call rather than a "conversion experience." Hand-picked as God's servant (slave), he received a revelation—not from men but by secret knowledge from God—that the Gentiles will come to the Christian faith without the Law, the Torah of the Jews. He himself could bear the Law, but he was told that the Gentiles do not need the Law in order to be accounted righteous. The conviction that the Gentiles stand equal before God was reinforced by his visit to James, Cephas (Peter), and John in Jerusalem, who confirmed his mission, enjoining him only to remember the poor (probably reference to the Jerusalem collection). Faith in Christ has thus superseded righteousness of works, and the Law is no longer needed.

The freedom of the gospel

The freedom of the gospel is the theme developed in chapters 3–4 in a series of allegorical-typological interpretations based on the Law. Paul first recalled the covenant promise to Abraham: that he "believed God and it was reckoned to him as righteousness" and that through Abraham all nations would be blessed.

In chapter 3 there is a complex line of thought: Christ has redeemed men from the curse of the Law by becoming a "curse" for men; Christ has taken away this curse by accepting it himself in order that all men by faith might receive the Spirit that was promised. But the promise had already been made to Abraham and his seed (singular), the Messiah, Christ; the Law had come only 430 years later, a sign that it is not eternal. In this chapter, Paul constructed arguments against the Law. First, the Law was added because of transgressions committed first by the people who caused Moses to shatter the first tablets of the Law and was thus not ultimate but rather time-bound, limited, and tainted by the evil reality it had to counteract; secondly, the Law was given only for a restricted time, from Moses "till the offspring should come to whom the promise had been made" (*i.e.*, Christ); thirdly, the Law came "ordained by angels through an intermediary," who is not God and thus is neither something glorious in itself nor the absolute manifestation of the salvation of God. Paul expanded on the Law in the image of a *paidagōgos* (instructor or custodian). Such a custodian is now not needed and served only as a restraint so that in God's timetable of salvation the Gentiles could be delivered after the Law has been "outgrown." Paul then showed the reasoning behind his statement that the Law was obsolete: in Christ (*i.e.*, in the church) there are no divisions between Greek and Jew, slave or free, male or female—all divisions or partitions are broken down.

Freedom from the elemental powers of the universe

Paul's arguments are bold. He even claimed that, as heirs through Christ, men were no longer bound under the elemental powers of the universe, which were apprehended as negative, as was the Law, in Paul's mind. In chapter 4 the Judaizers are said to keep themselves, like many Greeks, under astrological powers—not unlike the Jewish calendar of feasts—which kept man, according to Paul, enslaved by cosmic order. But to those free from the Law and possessing the Spirit, sonship and inheritance can come by adoption. Thus, Paul was negative in Galatians concerning the Law, and taught that freedom from it brings unity and the fruits of the Spirit.

In chapters 5–6 Paul listed catalogs of virtues and vices, fruits of the Spirit or the flesh, and stressed mutual forgiveness in the church. This is an exhortatory section that leads to the closing of the letter in Paul's own hand and to his stress on seeing his only glory in the cross of Christ.

The Letter of Paul to the Ephesians. The authenticity of Ephesians as a genuinely Pauline epistle has been doubted since the time of the Dutch Humanist Erasmus in the 16th century. It is most reasonable to consider it as "deutero-Pauline"—*i.e.*, in the tradition of Paul but not written by him. The problem of Ephesians cannot be solved apart from that of Colossians, because many similarities are noted in the style and development of Pauline thought into cosmic imagery; yet they treat different problems. In both, the heritage of Paul is preserved by a "Paulinist," and it is on this basis that Ephesians and Colossians were accepted into the canon. Both are "captivity epistles," ostensibly written by Paul from prison. Of the 155 verses in Ephesians, 73 have verbal parallels with Colossians; and when parallels to genuine Pauline letters are added, 85 percent of Ephesians is duplicated elsewhere. It would appear that Ephesians is dependent on an earlier more specifically oriented Colossians, and it may be that Ephesians uses, combines, and condenses the material of Colossians for its own needs.

Though Colossians is directed explicitly and strongly against a particular Judaizing proto-Gnostic heresy—*i.e.*, an incipient form of a religious dualistic system that emerged as a very attractive heretical movement in the 2nd century—Ephesians is not polemically oriented and is not clearly connected to a particular congregation, its problem, or its individuals. Though Ephesians uses a letter style with an introduction, greeting, and closing benediction, the only person mentioned in it is Tychicus, already mentioned in the same context in Colossians. The doctrinal section shows that the whole world—not only the Jews—is in a cosmic sense subjected to Christ, and Jew and Gentile are reconciled and united through him. This is the mystery of God's plan revealed to the church through Paul but expanded in scope. All are saved and reconciled through Christ, who has made both Jew and Gentile one and has "broken down the dividing wall of hostility," bringing peace and unity. The author of Ephesians continues Pauline language and makes it more Pauline than Paul himself.

The universal appeal of Ephesians

After the address—which, according to the best manuscripts, lacks a reference to Ephesus—there is a hymn of praise to God in terms of a cosmic plan of redemption. Through the ascended Christ, salvation is for all, and he is the head of the body, his church. Because the address and thanksgiving are to the church in general (the place name, Ephesus, being an early gloss), it is possible that Ephesians was meant as an encyclical, to be distributed, perhaps, as a covering letter for the whole Pauline collection. The "mystery of God's will" (chapter 1, verse 9) is spelled out in chapter 2 as the reconciling act of Christ for both Gentile and Jew. In chapter 3 Paul's role in giving knowledge of this mystery in his ministry leads to a doxology. After this semi-epistolary form, the general admonitions follow in terms of gifts of grace with stress on unity: one hope, one Lord, one faith, one baptism, one God for all. A warning against a heathen way of life is given in contrast with the Christian's old nature as opposed to his new being in Christ. In chapter 6, verses 10–20, the Christian is enjoined "to put on the whole armor of God" as defense against evil and Ephesians ends as a letter, with a blessing.

The Christology and ecclesiology imply a background of a Christianized, mythological proto-Gnosticism, or a strongly Hellenized Judaism. Perhaps one of the best clues to the lateness and pseudonymity of Ephesians in comparison with the genuine Pauline letters, however, is the phrase "revealed to his (Christ's) *holy* apostles and prophets by the Spirit." Such an expression is certainly later than Paul and looks back on the apostolic age as a time in the past.

The date and style of Ephesians

A possible date is shortly after Colossians, in the early 2nd century. Because there are so many similarities to Colossians, Asia Minor might be the place of composition, but this is merely conjecture. The non-Pauline use of the term mystery to denote that Gentiles are fellow heirs with Jews, the uniting of all in Christ, and an analogy between marriage and Christ's relation to the church, all point to a different and later time than that of Paul. The style of Ephesians builds up long, almost unmanageable, unpunctuated, excited, and abundant sentences, even longer than those of Paul when he is most provoked or, perhaps, absent-minded and does not finish sentences that he begins. A comparison of the table of duties of Colossians 3 and Ephesians 5 and 6 also shows a strong development in the direction of making the relationship of Christ and his church the basis for all other relationships.

The eschatology of Ephesians is attenuated, if not far in the background, and a continuation of the church is implied. In chapter 1, verse 13, the writer sees the Spirit as the guarantee (down payment) of the Christian's in-

heritance—a present indication through the Spirit that the Christian can live in faith in the world looking for the Kingdom but already sure he can draw on the powers thereof without an imminent expectation of the end-time. Ephesians gives hope for universal salvation, grace as a gift of God, strength in patience, and an example of unity for the church as well as freedom in the Spirit to attain maturity as a Christian.

The Letter of Paul to the Philippians. In its present canonical form Philippians is, according to several scholars, a later collection of fragments of the correspondence of Paul with the congregation in Philippi that was founded by Paul himself. The first of the two major difficulties leading to this conclusion concerning redaction of the letter is created by a discrepancy between chapters 2 and 3—*i.e.*, an entirely unexpected polemic in chapter 3 after a calm second chapter. Another major difficulty is the relationship of chapter 4, verses 10 and following with Paul's joyful acceptance of his suffering and the remainder of the present letter that deals with the collection the Philippians had made and sent to Paul in prison. The place of the expression of Paul's gratitude at the end of the letter is odd, particularly because Epaphroditus, the Philippian delegate conveying the gift, is thanked as though he had just arrived; yet he has already been described as ill when he was with Paul (who apologized in chapter 2 for not having told about Epaphroditus' illness sooner and the delay in sending him back). Yet, Epaphroditus is obviously back and the sequence of events is, indeed, confusing.

The fragmentary condition of the letter in its present form

The following rearrangement of the parts of the letter is probably acceptable. Chapter 4, verses 10–20 shows Paul reacting to the gift of the Philippians and the arrival of its bearer, Epaphroditus, and seems to be the earliest fragment, written probably during Paul's imprisonment (*c.* 53–54). The portions of the letter that treat of the theme of mutual joy (1:1–3, 4:4–7, and probably 4:21–23 that refers back to chapter 1) are best taken together as fragments of a second and somewhat later letter. The third section is 3:2–4:3 and possibly 4:8–9, which addresses the danger caused by outsiders and opponents who had started to penetrate the Philippian congregation with a theology Paul considered heretical and against which he aimed his polemic. Because this is an entirely new situation, it is probably a third letter, of which only the preface is missing. This arrangement also attempts properly to account for the fact that chapter 4 actually comprises endings of several letters. Thus, chapter 3, verse 1, which is itself a summation and ending, fits in.

The reference to frequent visits between Paul and the Philippians referred to in the correspondence makes its origin in Rome unlikely and points rather toward Ephesus as the place of imprisonment. Paul's reaction to the gift of the Philippians is almost rude (although he accepted gifts from no other congregation but preferred to support himself during his apostolic mission). He actually avoided expressing direct gratitude and attempted to divert the significance of the gift from its material side to its spiritual meaning. He emphasized the sympathy proven by the Philippians, the importance of the value of the gift for them as a spiritual sacrifice for God.

The "letter of joy"

The "letter of joy" section describes Paul's enthusiasm in his mission efforts—and their success—and his joy in the energy and growth of the mission in Philippi, which Paul shared with his congregation. Paul's address to "bishops and deacons," terms unique in Paul's letters except here, are, perhaps, circumlocutions for missionaries active in Philippi, a congregation that had become a strong and stable Christian community. Paul had traditionally remained there about one week and, in chapters 1 and 2, encouraged and praised the Philippians for continuing in their faith in his absence. This is part of the thanksgiving in Philippians—an emphasis on the participation, cooperation, collaboration, and empathy of the Philippians with respect to the preaching of the gospel. Thus, the terms bishop and deacon may belong to the language of a self-supporting mission church with its own overseers (bishops) and workers (deacons) and does not carry the connotations of later ecclesiastical structures. Paul expressed his confidence in the fine beginning of this young church that sought "to become pure and blameless for the day of Christ," the final judgment.

Paul then turned to his own experience of imprisonment, which he viewed as advancing the gospel. Though he considered that not all preachers of Christ preach on the basis of selfless motives, the fact that Christ is proclaimed is a most important cause for rejoicing. Paul then exhorted the Philippians to work hard for the sake of the gospel, not minding any opposition, and to do this in a sense of unity and mutual support.

This exhortation toward a strong and active sense of community was reinforced by quoting an early Christian hymn that described the humiliation (*kenōsis*) and exaltation of Jesus who is made the Lord of the universe and confessed by all cosmic powers. A part of Jesus' humiliation, his death on the cross, can be taken as part of his manifest glorification. The verses following the hymn make clear that the incorporation of the hymn with its triumphal ending also has a missionary purpose, because Paul emphasized again the need to responsibly act out one's own calling even before non-Christians. Thus, active responsibility continuously exercised in the perspective of the approaching Parousia merges with Paul's own readiness to sacrifice himself.

The missionary emphasis

In chapters 3–4 the situation may be totally different. Paul reacted to the threat of the appearance of Jewish-Christian missionaries who are rather close in theology to the Galatian Judaizers. Paul's polemic indicates that in addition to Jewish tradition, they must have emphasized the Law in particular. Reference is made to circumcision, and Paul emphatically claimed that he could compete with heretics boasting of their Jewish tradition and, in elaborating on that, emphasized his former pious righteousness under the Law, in which he was blameless. He then stressed categorically that for him the experience of Christ has terminated his former piety completely and that he has left it behind as of no value. Such a polemic implies that for his opponents such was not the case. Paul also argued against libertinistic tendencies, which indicates that his opponents were not legalists in an ordinary sense but combined faithfulness to the Law with a strong and fanatical enthusiasm that could lead toward "mysticism" and easily be misinterpreted as libertinism. Paul's emphasis on true Christian experience as not being completed but rather still being in the state of expectation might be a further polemic against overenthusiasm. In chapter 4, verse 8, Paul reaffirms his own example, making it, in imitation of the teaching of popular philosophy, the epitome of all positive ethical values and virtues, and thus the pattern to be imitated. This tendency toward the paradigmatic, together with warnings and autobiographical material in chapter 3, verse 2, to chapter 4, verse 3, can be seen as a "testament" of Paul, consciously written with an awareness of impending death or martyrdom. Thus Paul presents himself—his life, ideas, admonitions, and an eschatological section—as his heritage and as an incorporation of the message he preached and its value.

The Letter of Paul to the Colossians. Colossians presents the problem of having, on the one hand, numerous (though superficial) affinities with the circumstances of the Letter of Paul to Philemon while, on the other hand, being addressed mainly to a different situation. In this new situation he uses ideas and expressions that seem to be rather a development of Pauline ideas about the cosmic realm than genuinely Pauline argumentation. In this latter aspect, Colossians and Ephesians share the heritage of Paul, but a later "Paulinist" changed details to meet different situations.

The purpose of Colossians

Colossians was written ostensibly by Paul from prison (in Ephesus) to a predominantly Gentile Christian congregation founded by his co-worker, Epaphras, at Colossae. The Colossian congregation was endangered by a heresy involving a "philosophy" that was connected with the elemental spirits of the universe to which men seemed to be bound, with circumcision, feast days and food laws, visions, and an asceticism that was not only false in its piety but foreign to the Christian faith.

To combat these proto-Gnostic, syncretistic, and Judaizing tendencies, the Paulinist appealed to the authority of Paul's apostolate and his thought but accented his theology in a new way, enlarging Paul's theological dimensions, so that they included the whole universe, the fate of the entire cosmos. This whole world is depicted as subject to Christ and has its meaning, aim, and goal in the church, which is Christ's body and over which he is the head. This transformation of Paul's theology would appear to be somewhat later than Paul, yet not so much later than Philemon, and its import has been forgotten. Colossians cannot be dated or placed with certainty, but the end of the 1st century or the beginning of the 2nd century has been suggested.

In a first edition, before the Paulinist changed or added to it, Colossians seems close to the situation of Philemon. In both letters Paul is in prison. Onesimus appears in Colossians, chapter 4, and the readers of Colossians are asked to transmit a special injunction through the church of the Laodiceans to Archippus—possibly that the former slave, Onesimus, now referred to as a "beloved brother," be freed for service of the gospel. The same five names appear in Philemon and Colossians (Col. 4:10 ff.; *cf.* Philem. 23), which is unusual because the church at Colossae is strange to Paul. The lost letter to the Laodiceans may possibly be the Letter to Philemon, and the request to the slave owner would, by being read aloud in a neighbouring large church (Colossae), reinforce Paul's request that the slave be freed.

Later substantial redaction has obviously taken place, however, and it is the heresy at Colossae rather than the situation of Philemon that is mainly addressed in Colossians. Though Paul asserted that he did not preach and exhort where another has founded a church, here the Paulinist, using and amplifying Pauline theology, taught, gave thanks, and interceded for a church that he did not found and that was in danger of accepting heretical Judaizing teachings, thus falling away from Christ. The doctrinal section of Colossians sets forth in a hymn Christ's pre-eminence over the whole cosmos, all principalities and powers, to bring redemption through the cross and to be the head of the body, the church.

From this cosmological beginning, the style and imagery differ from the authentic Pauline letters. Colossians is wider and broader in scope, with long, almost breathless sentences. There is a hierarchy in Christ being head of the body, his church, which differs from the Pauline expression of equality of all the members, although with differing functions (cf. I Corinthians, chapter 12, and Romans, chapter 12).

Christological and ethical emphases

The Christology is applied to the situation of the church and Paul's role in behalf of the church—his suffering with Christ and knowledge of God's mystery, Christ—is used to bolster his defense against heresy. This polemic is based first on tradition and then proceeds to specific warnings against false teaching, cult, or practice. An admonition "to set your minds on the things that are above," because in Baptism the Christian has died and been raised with Christ, is followed by the conclusion that the Christian's conduct should be ruled by love and be thus free from all wrongdoing.

Another difference from the genuine Pauline letters can be noted in this latter section. When Paul referred to the resurrection of Christians he used the future tense in most cases, but Colossians, chapter 2, verse 12, and chapter 3, verse 1, presuppose that because the Christian is risen with Christ, ethical demands can be made.

In Colossae, such Christian ethics apparently were lacking, thus the inclusion of a table of duties—*i.e.*, a list of household duties and of relations between members of a household. General exhortations to prayer and right conduct are followed by the conclusion of the letter with its list of greetings. There are some similarities in Colossians to Paul's polemic against Judaizers in Galatians, but Colossians seems to reflect a later time and a more developed "cosmic" theology of a later deutero-Pauline writer.

The First Letter of Paul to the Thessalonians. In all probability I Thessalonians is the earliest of Paul's letters, particularly because the memory of the events that led to the founding of that congregation are still fresh in the mind of the Apostle. The letter was written from Corinth. According to I Thessalonians, chapter 3, verse 2, Paul had sent Timothy to Thessalonica from Athens during his brief stay there, had just experienced the delegate's return, and had received reports about the congregation to which he is reacting in this letter. I Thessalonians gives expression to Paul's surprise over the rapid growth of the Christian mission at Thessalonica, which was achieved despite immediate persecutions from pagan contemporaries. Paul acknowledged that the successful development had been wrought in the Thessalonians by their own acceptance, fully recognizing the human frailty of the Apostle, their founder (2:1–12), and not by a mistaken understanding that he himself was divine.

The rapid growth of the church at Thessalonica and Paul's response to eschatological expectations

Paul's surprise results, therefore, in overwhelming gratitude, and the customary Pauline thanksgivings here exceed the usual limits. A second reason for this unusually long thanksgiving—which actually makes thanksgiving the theme of the letter—is Paul's intent to undergird the encouragement he gives in 4:13–5:11. After having dwelt so extensively on his being moved by the change in the Thessalonians, Paul continues to state that therefore they have no reason for giving up faith in the face of the death of some fellow Christians, who had died between their conversion and the expected imminent Parousia of Christ. Apparently, they had expected the Parousia and final salvation as the promise of the Christian message. Paul encouraged his congregation that he had a "word of the Lord" that the dead and the living in Christ will rise together. "Word of the Lord" could refer to a word of Jesus known to Paul but could instead be a direct revelation to Paul.

In chapter 5 there is further thanksgiving, emphasizing the present gift and power of Christian faith and corporate Christian life. This emphasis is linked with ethical applications, with stress on brotherhood, diligence in keeping the faith, and religious industriousness. The difficulties of balancing the expectation of the Christian with God's timetable is outweighed by the hope and joy in what has already been experienced and what is hoped for. Paul's real emphasis is more on the actual description of Christian life in the face of coming salvation and vindication than on the preceding discussion of the fate of those who had died or on the actual circumstances of Christ's appearance from heaven.

The encouragement of the Thessalonians was introduced in chapter 4 by a genuinely ethical exhortation to proceed properly on the way to holiness and sanctification already begun. The brevity of this rather traditional exhortation is most unusual in Paul's letters and supports the observation that it was written in joy and confidence for a new congregation well begun in order to support it against attacks and doubts as it matured in the faith.

The Second Letter of Paul to the Thessalonians. A feature of II Thessalonians that resembles the otherwise most unusual feature of I Thessalonians is its excessively long thanksgiving. Within this thanksgiving there is an excursus dealing with the timing of the Parousia, but in II Thessalonians Paul aggressively argues against any expectation of an imminent coming of Christ that might be expected from the things he wrote in I Thessalonians. II Thessalonians perhaps presupposes I Thessalonians and intimates that believers had a false understanding of that communication of Paul. In II Thessalonians, much to the surprise of the reader of both letters, the statement is made that a letter "purporting to be from us" is "to the effect that the day of the Lord has come." II Thessalonians then presents a problem as to whether it was a self-correction of Paul or directed to the situation of a later time and thus the writing of a later author in a "Pauline" tradition. II Thessalonians does have more apocalyptically catastrophic language than I Thessalonians. Such a description not only underestimates the positive work of God and Christ for the believer but also says little about the Parousia. II Thessalonians claims that not all the events preceding the Parousia

Corrections about apocalyptic expectations

have yet occurred. The "mystery of lawlessness," opposed to the "mystery of godliness," is still at work in the world, and the full activity of Satan has not yet unfolded itself. Emphasis in II Thessalonians is on steadfastness as God's gift and promise in the days of tribulation, which makes the apostle ask for support in prayer. Criticism of people leading disorderly and idle lives follows. The perhaps casual admonition to work is thus elaborated into a major point.

Salvation seems to be sought almost exclusively in futuristic terms. Incipient or actual Gnosticism in the church could account both for the assertion that the fulfillment has already come and for the depiction of disorderly lives (because in "proto-Gnostic" terms the world is evil and provokes a response either of total renunciation or libertinism). II Thessalonians may thus reflect these problems and fit into the late 1st century. Verbal agreements between the two letters may be evidence of deliberate spurious writing, as also the suggestion in II Thessalonians that false letters may be circulating. A later author saw Paul's heritage threatened by too enthusiastic an understanding of Paul in Thessalonians and composed this letter to preserve Paul's meaning.

THE PASTORAL LETTERS: I AND II TIMOTHY AND TITUS

The Pastoral Letters as a unit. The First and Second Letters of Paul to Timothy and the Letter of Paul to Titus, three small epistles traditionally part of the Pauline corpus, are written not to churches nor to an individual concerning a special problem but to two individual addressees in their capacity as pastors, or leaders of their local churches. The purpose of the letters is to instruct, admonish, and direct the recipients in their pastoral office. Since the 18th century they have been referred to as a unit, the Pastoral Letters, and they contain common injunctions to guard the faith, to appoint qualified officials, to conduct worship, and to maintain discipline both personally and in the churches. Their similar peculiarities of style and vocabulary as well as the similarity of the heresies and other problems they faced place them in a common time and allow them to be dealt with as a unit. Their content presents a picture of the post-apostolic church when pastoral offices and tradition came to the fore and the formerly high apocalyptic tension appears attenuated.

Reasons for accepting a non-Pauline authorship

The Muratorian Canon (a list of biblical books from *c.* 180) includes references to the Pastoral Letters and notes that they were written "for the sake of affection and love." They have a place in the canon because "they have been sanctified by an ordination of the ecclesiastical discipline." These letters, however, do not appear among the Pauline letters in P 46, an early-3rd-century manuscript, and there is no clear external attestation in the primitive church concerning them until the end of the 2nd century. Not until the 19th century were doubts expressed about the Pastorals as being authentically Pauline, when German scholars and others noted discrepancies in style and vocabulary, church organization, heresies, biographical and historical situations, and theology from those found in the Pauline letters. The problems of authorship, authenticity, and dating almost paralyze investigation of the Pastorals unless discussion of these problems is seen as connected also with the literary character of the material.

Attempts have been made to apply the tools of statistical analysis in comparing these disputed letters to the rest of the New Testament (particularly to the Pauline corpus) for the purpose of establishing authorship. The studies, utilizing computer technology, point toward non-Pauline authorship with affinities to language and style of a later, possibly 2nd-century, date. More refined and complex analyses, however, are still needed.

Linguistic facts—such as short connectives, particles, and other syntactical peculiarities; use of different words for the same things; and repeated unusual phrases otherwise not used in Paul—offer fairly conclusive evidence against Pauline authorship and authenticity.

Content and problems. Church offices are more developed in the Pastoral Letters than in Paul's time. There are presbyters and bishops, but these are sometimes used interchangeably and the monarchical episcopate is not yet depicted, although church offices appear to be heading in that direction. Requirements for office are strict and leaders are chosen and ordained by laying on of hands. Such leaders must be able to teach true and sound doctrine and guard what has been entrusted to them, the *parathēkē*—*i.e.*, the deposit of teaching or the message to be carried on. They must also be able to stand firm and argue against heresy. Such offices and aims suggest an expectation of future generations of faithful witnesses to carry on the traditions, perhaps particularly necessary as some may be killed for the witness they make.

Internal organizational and theological developments

The heresies referred to appear to be Gnostic and the arguments are rather mild and reasonable, unlike Paul's urgency in combatting heresy with strenuous argumentation. The heresies taught by false teachers are an early partly Encratitic (abstaining) Gnosticism, with "higher knowledge" that emphasizes "godless and silly myth," or are statements that the resurrection has already taken place, which is a denial of future resurrection and a glorification and spiritualizing of resurrection as a rebirth, as, for example, in Baptism.

Biographical notes about Paul's journeys and situations contradict his own letters as well as the accounts in Acts. The Pauline sense of living in a time close to the end of the age is missing in these descriptions of churches; they are viewed as settling down with a succession of tradition with Hellenized expressions of salvation and a replacement of enthusiasm with bourgeois ethics. This indicates a period of de-emphasized eschatology and an expectation of a long community life in which people must live out their lives in Christian responsibility and moral behaviour.

I Timothy and Titus are more similar to each other than they are to II Timothy, but all three mark exhortations to personal lives of exemplary conduct and give rules of conduct for church order and discipline for the group as a whole and for individual parts of it—sometimes in terms of catalogs of virtues and vices recalling the Jewish two-way orders: the way of life being good, the way of death including a list of sins. Each concludes with a final blessing or salutation. They are all pseudonymous, using Paul as an epistolary model and using pseudonymous devices, such as naming individuals known to be Paul's co-workers. The authority of Paul is invoked to lend authority to the teachings contained in the letters: the avoidance of heresy, holding to sound doctrine, and piety of life. The author is anonymous, the place of writing and the addressees are unknown, but they probably are later spiritual children of Pauline teaching. The date of the letters is about the turn of the 2nd century.

II Timothy uses the background of Pauline imagery most fully. It is cast at least in part in the testament form to Timothy as his spiritual heir because Paul is depicted as suffering, fettered in prison, and awaiting the martyr's crown. He exhorts Timothy and through him the church to share in these sufferings as they will eventually share in glory. II Timothy, chapter 2, verses 1–13, is an exhortation to martyrdom with a faith that Christ, triumphant over death, will save his faithful witnesses. Recollection of the creed is followed by a direct application to bearing suffering and its meaning in God's plan of salvation. The words "faithful is the word" occur in 2:11. This "word," unlike Paul or any Christian, cannot be bound. It both confirms salvation described in the preceding verses and introduces a hymn that may represent liturgical usage in that it is poetic and balanced.

Early liturgical Christological hymns

> Faithful is the word:
> If we have died with him, we shall also live with him;
> if we endure, we shall also reign with him;
> if we deny him, he also will deny us;
> if we are faithless, he remains faithful—for he cannot deny himself
>
> (II Tim. 2:11–13)

The hymn preserves within itself a reflection of sayings of Jesus that those who endure and persevere will reign with the Lord and that even to those who deny him (as

did Peter) God will remain faithful because Christ cannot deny his own faithfulness. Even in this hymn there is allusion to a "testament" form, with Paul already martyred, as a pseudonymous device to spur the Christian on to endurance and faithfulness as a member of the redeemed community.

Another small poetic hymnic section serves to demonstrate that the church of the Pastorals, albeit somewhat de-eschatologized, retains the "mystery" in God's household, the church—*i.e.*, the gospel and creed alive in the liturgy in the mystery of piety and worship.

> Great indeed, we confess, is the mystery of our religion:
> He who was manifested in the flesh,
> vindicated in the Spirit,
> seen by angels;
> who was proclaimed among the nations,
> believed in throughout the world,
> glorified in high heaven
>
> (I Tim. 3:16)

Here in miniature are creed and gospel somewhat reminiscent of the Gospel According to Matthew.

The Letter of Paul to Philemon. From Ephesus, where he was imprisoned (*c.* 53–54), Paul wrote his shortest and most personal letter to a Phrygian Christian (probably from Colossae or nearby Laodicea) whose slave Onesimus had run away, after possibly having stolen money from his master. The slave apparently had met Paul in prison, was converted, and was being returned to his master with a letter from Paul appealing not on the basis of his apostolic authority but according to the accepted practices within the system of slavery and the right of an owner over a slave. He requested that Onesimus be accepted "as a beloved brother" and that he be released voluntarily by his master to return and serve Paul and help in Christian work. Paul appealed to the owner that Onesimus (whose name in Greek means "useful") is no longer useless because of his conversion and claimed that the owner owed Paul a debt (as he probably was also instrumental in his conversion) and that any debt or penalty incurred by the slave would be paid by Paul. Such manumission is part of Paul's concept of being an ambassador to further the mission of Christianity, rather than a judgment on the social framework of slavery, because in the Lord such social order is transcended.

Paul's views of Christian brotherhood transcending slavery

Philemon, however, is not a purely personal letter, because it is addressed to a house church (a small Christian community that usually met in a room of a person's home), and it ends with salutations and a benediction in the plural form of address. The body of the letter, however, uses "you" (singular) and is addressed to the slave's owner, a man whom Paul himself has not met. Philemon, the first name in the address, is called a "beloved fellow worker," which implies that he knew Paul, and it has been convincingly argued that the slave's owner was Archippus (see above *Colossians*), perhaps Philemon's son, who was called a "fellow soldier," a term usual in business accounts and suitable for a document on the manumission of a slave. The thanksgiving contains the main theme of the whole letter: sharing of faith for the work of promoting knowledge of Christ.

The letter was written from prison, and Paul apparently expected a release in the near future, because he requested a guest room, a suggestion that he was not very far from Colossae or Laodicea, which would be true of Ephesus. Colossae would be reached from Ephesus via Laodicea, and the letter could be addressed to a house church there.

The role of Onesimus in the early 2nd century

In a letter to the Ephesians (*c.* 112) by Ignatius, bishop of Antioch, the language is very reminiscent of Philemon, and the name of the bishop of Ephesus (*c.* 107–117) was Onesimus. It has been suggested that the slave was released to help Paul, that in his later years he might have become bishop of Ephesus, and that his "ministry" or "service" was the collection of the Pauline corpus. This is based not simply on the identity of name, but on similarities to Philemon found in Ignatius' letter to the Ephesians, as well as two possible plays on words in chapter 2, verse 2 (*cf.* Philemon, verse 20), and chapter 4, verse 2 (*cf.* Philemon 11) relating to the bishop and unity of the church. Such a prominent position and role for one of Paul's followers might shed further light on why Philemon, apparently a very personal plea, became a part of the canon and Pauline corpus. Even if this suggestion cannot be proved, Philemon still shows Paul in his apostolic ministry, furthering the message of Christ and seeing beyond the limitations of the social order of his day, in which both slaves and freemen are servants of God.

THE LETTER TO THE HEBREWS

The writing called the Letter to the Hebrews, which was known and accepted in the Eastern church by the 2nd century, was included also by the Western church as the 14th Pauline epistle when the canon of East and West was assimilated and fixed in 367. Hebrews has no salutation giving the name of either the writer or the addressees, although it does have a doxology and greeting at the end, which suggest that at some point the writing was sent as a letter to a community known to the author. There are also numerous admonitions in the text that appear to be directed to a definite circle of addressees and some admonitions to the church at large. In chapter 6, verses 4–8, is a severe warning against the sin of apostasy, for which there is no second repentance. Even so, Hebrews is essentially more a theological treatise than a letter. It is homiletical in style and calls itself a *paraklēsis*, which has many meanings: consolation, exhortation, sermon, advocacy, and even intercession.

The question of authorship and peculiarities of style and content

The thoughts, metaphors, and ideas of Hebrews are distinct from the rest of the New Testament, with closest affinities to Stephen's speech in Acts, chapter 7. It attempts to prove the superiority and ultimacy of the revelation in Christ and the perfection of his offering of himself once and for all supersedes and makes obsolete any other revelation. Hebrews gives strength to its readers through the example of Christ and the hope and promise of free access to God and to eternal rest, an access in which Christ is High Priest and mediator forever. Such promise, on the basis of Christological developments and new covenant hopes, enables endurance in persecution, but its vocabulary is that of the sacrificial language of the Old Testament. Another theme is a typological analogy with the wilderness wanderings of Israel in which, despite their murmurings of unbelief and the hardening of their hearts in their trials, they persevered. Thus, the church, as the pilgrim people of God, travels toward the future place of Sabbath rest with Christ as their pioneer and perfector of faith.

A "word of consolation" is needed to strengthen faith in time of trouble. Actual persecution leading to martyrdom is seen as not yet come, but the church is sharply warned against apostasy, the sin of all sins. Hope during persecution and trial is expressed in the image of Christ as the perfect everlasting high priest, one of whose functions is to stand as intercessor and protector.

Suggested authors of Hebrews

Hebrews was considered a Pauline letter in the early Eastern church. Clement of Alexandria, a theologian of the late 2nd and early 3rd centuries, held that Paul had written it in Hebrew for the Hebrews and that Luke had translated it into Greek. Origen, Clement's successor as leader in the catechetical school at Alexandria, commented that its thoughts reflected Paul but that it was written at a later time with a totally different style and phraseology, and he stated "who wrote the epistle, God knows." Paul, for example, uses the term mediator only once and in a negative sense, in Galatians, chapter 3, verse 19, but Hebrews uses it several times of Christ as mediator of the new covenant. In the West, Tertullian, a North African theologian of the late 2nd and early 3rd centuries, suggested Barnabas as the author, because Hebrews, called a "word of consolation," might have been written by Barnabas, whose name is translated by Luke as "son of consolation" in Acts, chapter 4, verse 36. After Hebrews' acceptance into the canon in the mid-4th century, it was considered Pauline, but doubts persisted; and because of basically different content and style in contradiction to Paul, various authors have been suggested for Hebrews—*e.g.*, Apollos (a Jewish Christian

Alexandrian), or a follower of Stephen and the Hellenists, who had come into conflict with those not sharing his universalistic ideas. Hebrews, however, remains anonymous. The title "To the Hebrews" is secondary and may reflect either an idea as to its addressees or that it was influenced by its extensive Old Testament material.

According to internal evidence, Hebrews was written in a second or later generation of Christians. Persecution references suggest a time after Nero's persecution and about the time of the emperor Domitian but early enough to be quoted or alluded to in the First Letter of Clement (*c.* 96), thus suggesting a date of *c.* 80–90.

The place of the addressees may be Italy, because 13:24 is understood as a greeting sent home from one writing from abroad, but this is not certain. The addressees were probably Gentile Christians who needed instruction in "the elementary doctrines of Christ" and concerning faith in God.

Allegorical or typological interpretive techniques

Hebrews constitutes the first Christian example of a thoroughly allegorical, typological exegesis (critical interpretation) of the Old Testament. There were precursors of such a methodology in Jewish Alexandrian biblical exegesis (*e.g.*, Philo), and Platonic tendencies found in Hebrews can also be found in Jewish-Alexandrian methods of interpretation of the Old Testament. The language of Hebrews is extremely polished, elegant, and cultured Greek, the best in the New Testament. Linguistically and stylistically, it shows only a slight influence of the Koine (common Greek). The Attic style is broken only in passages in which Hebrews quotes the Septuagint. Plays on words and synonyms with similar beginnings for emphasis show the author's literary craftsmanship.

There are more Old Testament citations in Hebrews than in any other New Testament book. They are drawn mainly from the Pentateuch and some psalms.

The church is viewed as being in danger of discouragement in the face of persecution and possible apostasy. If faithless, church members risk total loss, for no second repentance is possible. Through his special Christology, the author seeks to help the readers by showing that Christ is the saviour superior to any other and that as Saviour, Son of God, High Priest, pioneer, guide, and forerunner, he who has already suffered and been glorified will lead the wandering people of God to their eternal Sabbath rest, an eschatological future state of peace and renewal.

Christological and eschatological motifs

This high type of Christology is combined with much stress on Jesus' humanity. He partook of man's nature and overcame death to destroy the power of the devil in order to deliver man. Thus, having been made like his brethren he has become a faithful High Priest to make expiation for the sins of the people. Because he himself suffered and was tested, he can help those who are tested and tempted. Through suffering, tears, and obedience Jesus was made perfect and thus the source of help and salvation, being designated by God a High Priest after the order of Melchizedek.

Christ and his once for all (*ephapax*) sacrifice has superseded and made all Old Testament sacrifices and cultic practices obsolete. Christ is superior to the prophets because he is a son, superior to the angels because they worship him, and (in the light of his cosmic role as apostle and High Priest) superior to Moses, who brought God's Law to Israel, because Moses was a servant in God's house and Christ a son. Christ is also superior to Moses' successor Joshua, because Joshua did not bring the wandering people into a perfect rest; superior to the Old Testament priesthood of Aaron, because Christ, the true High Priest, has sacrificed himself once for all and is without sin; and superior to the patriarch Abraham, because Abraham paid tithes to the priest of Salem, Melchizedek, who as the prototype of Christ had no human antecedents. Christ, High Priest forever by obedient suffering and perfection in that he lives up to the demand, has become the source of salvation. He is High Priest in the heavenly tabernacle and mediator for the new covenant. On the basis of this Christology and ecclesiology, the rest of Hebrews is composed of injunctions to faithful life in all situations, spiritual or temporal. In chapter 11, verse 1, Hebrews gives a programmatic statement that should be translated: "Faith is the Reality [rather than "assurance," as in the usual translation] of what is hoped for and the Proof concerning what is invisible." In Hebrews, Jesus is that Reality and that Proof, and everything else is unreal or at best an earthly copy or a shadow. The heroes and martyrs of old were looking toward his coming (chapter 11) and those now under persecution look toward him and find strength (chapter 12) as they leave the ultimately unreal structures of this world, seeking the "coming city" and going out to him who was executed outside the walls of the city made with hands. Thus, the message of Hebrews is: Reality versus sham and shadow, Christ's sacrifice (priest and victim in one) versus the cult of temples, and the real heavenly rest and heavenly city versus the sabbath and Jerusalem.

THE CATHOLIC LETTERS

As the history of the New Testament canon shows, the seven so-called Catholic Letters (*i.e.*, James, I and II Peter, I, II, and III John, and Jude) were among the last of the literature to be settled on before the agreement of East and West in 367. During the 2nd and 3rd centuries, only I John and I Peter were universally recognized and, even after acceptance of all seven, their varying positions in Greek manuscripts and early versions revealed some conflict concerning their inclusion. The designation Catholic Letters was already known and used by the church historian Eusebius in the 4th century for a group of seven letters, among which he especially mentions James and Jude. The word catholic meant general—*i.e.*, addressed to the whole, universal church as distinguished, for example, from Pauline letters addressed to particular communities or individuals. The earliest known occurrence of the adjective "catholic" referring to a letter is in the account of an anti-Montanist, Apollonius (*c.* 197) in his rebuke of a Montanist writer who "dared, in imitation of the Apostle [probably John] to compose a catholic epistle" for general instruction. In the time of Origen (*c.* 230), the term catholic was also applied to the *Letter of Barnabas* as well as to I John, I Peter, and Jude.

The meaning of "Catholic Epistles"

In the West, however, "catholic" took on the meaning in Christian usage as implying a value judgment as to orthodoxy or general acceptance. Thus, the West used it for all the New Testament letters that were in the canon along with the four gospels and Acts. All letters considered authoritative and of equal standing with those of Paul were therefore termed canonical in the West. Not until the Middle Ages did both East and West designate the seven as "catholic epistles" in the sense of being addressed to the whole Christian Church, in order to distinguish them from letters with more particular addresses. Had not the main tradition placed Hebrews in the Pauline corpus, it would perhaps rather have been counted among the Catholic Letters. Hebrews, however, looked "Pauline" rather than "Catholic" in that it presented an extensive theological argument to which the parenesis (advice or counsel) was applied at the end.

These seven letters are grouped together despite their disparate authorship and dates because of a number of characteristics common to all of them. Though the three Johannine letters, and especially I John, are distinctly Johannine in character, the four other Catholic Letters are of special interest precisely because they lack strong personal or peculiar traits both in their theological and their ethical statements. This characteristic makes them a good source for understanding the piety and life style of the majority of early Christians. These letters differ from the Pauline letters in that they seem to have been written for general circulation throughout the church, rather than for specific congregations. Though Paul wrote as a missionary responsible for his recent Gentile converts, these letters address established congregations in more general terms. It is interesting to note, for example, that in I Pet. 2:12 the word Gentiles refers to "non-Christians" without any awareness of its older and Pauline meaning of "non-Jews."

The purpose of the Catholic Letters is to meet ordinary

The purpose of the Catholic Letters

problems encountered by the whole church: refuting false doctrines, strengthening the ethical implications of the Gospel message, sharing in the common catechetical and moral materials, and giving encouragement in the face of the delay of the Parousia and strength in the face of possible martyrdom under Roman persecution. They guide the ordinary Christian in his day-to-day life in the church.

The Catholic Letters preserve a considerable common legacy of ethical themes and quotations. Such themes and quotations (from the Old Testament) were handed down traditionally, though the writers interpreted them independently for their situations. For example, Proverbs, chapter 3, verse 34, showing God's scorn to scorners and favour to the humble, is used in James, chapter 4, verse 6, as a warning against involvement in the world and an exhortation to submission and humility, but in I Peter, chapter 5, verse 5, it exhorts Christians to humility and submission in relation to one another in the church and brotherhood. Because the Catholic Letters represent a common pool of Christian teaching, there are overlapping points, but these come from shared tradition rather than literary dependency. The virtues extolled in the early church are not particularly Christian but often coincide with those cultivated in Hellenistic culture, sometimes with a Jewish Hellenistic emphasis. An act of mercy and virtue valued both in Jewish and Hellenistic tradition is epitomized in hospitality (*e.g.*, I Peter 4:9). Similarly, Hellenistic lists of virtues and vices occur as needed from the general body of early Gentile Hellenistic tradition applied to the Christian communities. In these epistles, theological and credal statements are woven in and used for immediate ethical application. Thus, they differ from the Pauline style of extensive theological sections coupled with ethical applications that follow at the end of the epistle.

In the Catholic Letters, to be a Christian was to be in opposition to the world, a member of a minority church and thus at any time liable to be called as witness to the faith and perhaps to suffer and die for it. Eschatological trials are coming (*e.g.*, I Pet. 1:6f., 4:12–19; II Pet. 3:2–10; I John 2:18 ff., 4:1–4; Jude 17 ff.), and the Christian views false prophecy and heresy as well as hostile encounter with the world as part of the trials. The theme of joy in persecution, suffering, and the final trial or ultimate "testing" is based on Christ's victory over these events and the sense of being a member of his community. Thus, the Christian should show submission, nonretaliation, humility and patience, good conduct, and obedience to authorities, because his witness must be blameless when his faith is tested in the world, in the courtroom, and in martyrdom.

The Letter of James. The Letter of James, though often criticized as having nothing specifically Christian in its content apart from its use of the phrase the "Lord Jesus Christ" and its salutation to a general audience depicted as the twelve tribes in the dispersion (the Diaspora), is actually a letter most representative of early Christian piety. It depicts the teachings of the early church not in a missionary vein but to a church living dispersed in the world knowing the essentials of the faith but needing instruction in everyday ethical and communal matters with traditional critiques on wealth and status. In matters of church discipline and the practice of healing, there is stress on prayer, anointing, and confession of sin in order that the healing of the sick may be effected. Steadfastness, even joy, in persecution is based on pure religion with strong ethical demands, as noted in chapter 1, verses 2–4, and 19–27.

A debate as to how James' statement that "faith apart from works is dead" compares with Paul's "justification by faith without works" in Romans has a long history. The debate, central to the history of Christianity, has usually overlooked the simple fact that Paul speaks about "works of the Law" and does so with reference to those "works" that divide Jews and Gentiles—*e.g.*, circumcision and food laws. James, on the other hand, refers to works of mercy. Thus, the two statements are not only reconcilable but address themselves to quite distinct and different issues. Even Paul referred to mutual support of the brethren by the glorious phrase "the law of Christ" (Gal. 6:2) and this is the same as James' "royal law" (James 2:8). The Pauline language presumably was not in James' mind. In James, chapter 2, the example of Abraham's faith is used to show justification by works. It is to be noted that Paul also used Abraham as the paradigm of righteousness to demonstrate justification by faith in Romans, chapter 4, again showing the difference in purpose and setting of the two epistles.

In view of the post-apostolic situation depicted, James, the son of Zebedee, who died as a martyr before AD 44, could not have been the author. From the content, neither could James, a brother of the Lord and the leader of the Jerusalem church; his martyrdom is reported as *c.* AD 62. Thus, James is pseudepigraphical, with the purpose of gaining apostolic authority for its needed message. The date of writing is probably at the turn of the 1st century, and its addressees are the whole church.

James' stress on ethical imperatives

Of James' 108 verses, 54 contain imperatives—an obvious proof that advice is stressed. Such admonitions are expressed in the form of general ethical wisdom sayings, Hellenistic Jewish lists of virtues and vices, and Christian as well as pagan aphorisms sometimes related to popular preaching of the Stoic Cynic style.

In chapter 5 the community is enjoined to patience, steadfastness, and good behaviour. The Old Testament prophets, who spoke in the name of the Lord, are used as examples of suffering and endurance as they awaited the Judge. Thus, reference to the Parousia of Christ may have been conflated by the Christian writer to the coming of the Lord in judgment, an interpretation with "the day of the Lord" in mind. "Behold, the Judge is standing at the doors" is accompanied by the admonition, "You also be patient. Establish your hearts, for the coming of the Lord is at hand," (chapter 5, verses 8 and 9).

The purpose of I Peter

The First Letter of Peter. The purpose of the First Letter of Peter is exhortation directed to "the exiles of the Dispersion" in Asia Minor in order that they "stand fast" in God's grace in the face of persecution. On the one hand, such persecution is viewed as part of the trials of the end-time that the community must undergo before the coming of the new age. On the other, persecution is viewed as a simple fact of Christian community life in the world. In imitation of Christ, tribulations and testing can be a basis for joy.

In the address, the author calls himself "Peter, an apostle of Jesus Christ," and in chapter 5, verse 1, a "fellow-elder and witness of the suffering of Christ." Any Christian, not just a fellow eyewitness, however, might be such a witness and hope to partake in the future "glory that is to be revealed." The writer or the redactor of I Peter used Pauline and gospel theology and terminology both in quotations and allusions and, if literary dependency cannot always be demonstrated, there is dependence on the catechetical traditions known in the post-apostolic church.

The milieu of the letter seems to reflect the time and temper of the correspondence of the emperor Trajan with Pliny the Younger, governor of Bithynia (*c.* 117). Pliny requested clarification as to the punishment of Christians "for the name itself" or for crimes supposedly associated with being a Christian. I Peter, chapter 4, verse 15, appears to reflect this situation: that a Christian be blameless of all crime and, if punished, be persecuted only "as a Christian." Pliny continued that denounced Christians are executed if they persevere in their belief but that whatever their creed "contumacy and inflexible obstinacy deserved punishment"; Trajan's response was that those denounced as Christians be punished. The warning in I Peter, chapter 3, on a Christian's manner of defense and submissiveness to authorities points to a date in the first quarter of the 2nd century. Such a date does not preclude reflection on earlier persecutions, such as those under Domitian.

The Greek style is hardly in keeping with a Galilean Peter—described as illiterate or uneducated in Acts chapter 4, verse 13. The Greek is fluid, and the Old Testament citations are from the Septuagint. The addressees

appear to be Gentile Christians portrayed as the new Israel dispersed among the (heathen) Gentiles, based on the analogy of the old Israel, a diaspora among the nations.

The work is thus pseudonymous, attributed to Peter through Silvanus, whose name constitutes a part of the pseudepigraphic device that strengthens the authority of the epistle. I Peter is an excellent example of the testament form modelled on the traditions of an Apostle and the message of his martyrdom. Peter, whose death and traditions concerning him were known to the readers of the time of I Peter, gives weight and authority to the letter that is formed in many ways as a farewell and admonition to those who follow, in order that they may stand firm.

The letter in the form of a testament

Warnings are given from the Apostle's own example along with counter-virtues for vices. Such testament forms have a mixture of wisdom material, advice, exhortation, hymns for ethical admonition, and apocalyptic elements with accounts of trials to come. This mixture is found in strange arrangements, but is perhaps solved if read as a testament form. Peter had denied that Christ must suffer and in I Peter suffering is the way of discipleship and even of joy. In Luke, chapter 22, Peter's denial was prophesied, and Jesus interceded for him in order that he might repent and strengthen his brethren (*cf.* I Peter, chapter 5, verses 10 and 12). In Mark and Matthew the defection of the Apostles was foretold in terms of the scattering of the sheep when the shepherd was stricken, and Peter does deny his Lord. In John, chapter 21, the risen Lord paralleled Peter's threefold denial with a threefold question as to Peter's love. At each affirmation the Lord responds with the forgiving command to feed the sheep—to care for the community. This is a central motif in I Peter. Immediately following the charge to Peter in John is the prediction of his own martyr death, and in I Peter the church is urgently admonished to accept trials as nothing strange, because they are a sharing in the sufferings of Christ. In the Garden of Gethsemane, Peter in particular was rebuked because he did not watch, and in I Peter the church is admonished to watch and be vigilant against the Devil. Prayer against temptation is also stressed.

In the Matthean account, Peter is delegated to build the church, and in I Peter it is the chief Apostle (Peter) who points to Christ as Shepherd and Bishop, who through his suffering collected the wandering sheep to himself. In like manner—on the model of Christ or perhaps Peter—the elders are exhorted to feed their flocks humbly and faithfully. Thus, there is a typical testament form: Peter has failed and repented; and the church is warned, admonished, and strengthened as by the Apostle, who, on the analogy of Jesus' Passion and death in innocence, exhorts the church to share in the vocation of innocent suffering and to do good in innocence. Finally, I Peter, viewed as a "testament," is in itself an apocalyptic "witness," and with its admixture of advice, example, and general address to the faithful living in the Diaspora as sojourners, with the authority of its martyred "author," it constitutes authority and strength for the church that faces the persecution of the world. References in chapter 5 to Rome (called Babylon) and to Mark are then also part of the pseudepigraphic testament form, as they presuppose the common tradition of Peter's martyrdom in Rome and his connection with Mark.

Christological elements

There are three Christological hymnic fragments in I Peter: 1:18–21, ransom by Christ; 2:21–25, with reference to the Book of Isaiah, chapter 53, used as ethical admonition; and 3:18–20, Christ's descent into hell. The last is in the context of Christ's going and preaching to the spirits in prison (a reference to the apocryphal *First Book of Enoch* with Satan chained under the earth but his descendants at work in the world until the end-time) in order to show that Christ, through his descent, has overcome the powers that underlie and engender persecution of the Christians. This is reaffirmed in chapter 5 by encouraging Christians in their fight against the Devil, for, though suffering will be a part of this resistance, there will be victory at the end. Imitation of Christ is a basis for joy even in suffering. The end is viewed as near, and final salvation can thus be anticipated.

The Second Letter of Peter. The Second Letter of Peter was written as a letter to the whole church purporting to be similar in testament form to that of I Peter. It deals with the problems of the delay of the Parousia and accounts for it in terms of God's time being different from that of man and God's patience in waiting for all men to be better ethically. This letter, the latest of the New Testament, shows how Christendom dealt with the delay of the Parousia, discarded older Jewish apocalyptic ideas by substituting those with Hellenistic emphases, and is clearly in its content and exposition a methodically worked out artistic product, fictionalizing the older beliefs, in order to bring them into some agreement with traditional Christian terminology.

Problems dealt with in II Peter

II Peter names Simon Peter as its author and declares his position by setting down rules for true faith as he sees it. His work is different in meaning and interpretation from the earlier tradition and understanding of the church. He regards the Transfiguration of Jesus on the mountain as the first Parousia and urges patient waiting for the final coming of the Lord. Although he refers to his letter as a second letter of Peter, his Hellenistic concepts and rhetoric could hardly be attributed even to the author of I Peter. II Peter speaks of "partakers of divine nature," a term from the mystery religions, and mixes proverbs with familiar quotations from Hellenistic tradition. Thus, not only is this letter pseudepigraphic, but it is an even later fiction, probably nearer to AD 150 than the end of the 1st century.

Almost all of Jude is used in II Peter, but II Peter drops out a quotation from *I Enoch* in Jude 14 ff., possibly demonstrating some fear of using apocryphal writings.

Attacks on heresies and eschatological views

Heresies are attacked by criticism of their interpretation of scripture and misuse of set tradition, another evidence of the late date of II Peter. Reference is made to "all the epistles of Paul which contain things hard to understand" and to "other scriptures," evidence of a New Testament canon well on its way to being delineated over against the Old Testament. Though skillfully composed, II Peter cannot hide the Gnosticism included in its view and much misinterpretation of the traditional body of faith of the early church. Thus, II Peter is an example of the church at a relatively late period, de-eschatologized for the most part and brought near to early institutionalized religion with a ministry but depending on ideas and a theology so changed that it is almost unrecognizable.

The eschatology of II Peter awaits a new heaven and a new earth after the dissolution by fire of the old, evil earth with its unrepentant people. The Parousia no longer is Christological in nature but anthropologically oriented, with a vindication of the good and a punishment of the wicked. II Peter presents a picture of the church at the latest point in the canon and illustrates the necessity to re-evaluate and recall more normative Christian traditions.

The Johannine Letters: I, II, and III John. The three epistles gathered under the name of John were written to guide and strengthen the post-apostolic church as it faced both attacks from heresies and an ever increasing need for community solidarity—along with the concomitant love and ethics necessary to such unity.

The purpose of the Johannine Letters

I John, though lacking any formal epistolary salutation or ending, directs itself to a circle of readers with whom the writer is acquainted. Taking the form of an anonymous "homily" for admonition against heresy and instruction in faith and love, it was directed to a wide audience or was to be circulated beyond a particular congregation. II and III John are brief letters from an author described only as "the elder," implying a position of some authority. II John, chapter 1, is addressed to an "elect lady and her children," probably a designation of a church with difficulties similar to those found in I John. III John is the most personal, being addressed by the elder "to the beloved Gaius," who has been praised particularly for his hospitality (probably to missionaries)

and his brotherly love. The presbyter (elder), probably the author of II and III John, apparently was a man who was authoritative enough to influence and direct mission activities. All three letters, despite their differences of address, appear to have been accepted among the Catholic Letters as having been circulated for the church at large.

I, II, and III John share much common terminology, style, and general situation. They are all called Johannine because they are loosely related to the Gospel According to John in style and terminology and could be the outcome of its theology.

The early church attributed I, II, and III John to John, the Apostle, the son of Zebedee. Although II and III John may have been written by the same presbyter, this "elder" is not necessarily the author of I John, though it is commonly accepted that the three Johannine letters came from a "Johannine" circle. The earliest reference to the Johannine letters is in the *Letter to the Philippians* by Polycarp of Smyrna (7:1). Papias, a 2nd-century bishop of Hierapolis, mentions I John and quotes it several times but distinguishes between John, the Apostle, and John, the presbyter. Polycarp, Papias, and internal evidence point to the region of Asia Minor as sources of the Johannine literature. These references, the organization of the churches indicated in the letters, and the lack of signs of persecution suggest a date for the letters around the beginning of the 2nd century.

The First Letter of John. I John assumes a knowledge of the Johannine Gospel (the author of I John may be the ecclesiastical redactor of the Gospel According to John) and adds ethical admonition and instruction regarding the well-being of the church as it confronts heresy and stresses the lack of moral concern that springs from it. There is strong defense against the threat of a type of Gnosticism called Docetism that denied the reality of Jesus' earthly life and thus the meaning of the cross. Possessing special spiritual knowledge, the Docetic Gnostics had no need of the earthly Jesus and the humanity of Christ. This Docetic heresy led them to reject the Lord's Supper, but not Baptism. Their special possession of the Spirit had led them erroneously to consider themselves sinless and to deny the fellowship that has the cleansing of sins. Because the heresy may have led to libertinism, the ethics of Christians must accord with their faith and find expression in the love of the brethren in the church. "He who hears my word and . . . believes has passed from death to life" (John 5:24) is continued in I John 3:14, "We have passed out of death into life, because we love the brethren." The Gnostics separated themselves from the church in schism and have thereby committed the "sin unto death." They are false prophets and deceivers described by the term Antichrist. The true Christians, the "children of God," hold the true faith evidenced by their loyalty to the church and their charity toward its members.

Apologetics against Docetic Gnosticism

A constant theme in I John is that of God's love, which makes Christians the children of God. As children of God they keep the new commandment of love, which is of light—that of brotherly love—and resist the world, evil, and false teaching. Because Christ gave his life for man, the Christian's response is also to be self-giving. Through obedience and faith, God forgives even when man's heart condemns him, "for God is greater than his heart." It is of interest to note that in I John 2:1–2, Jesus is referred to as paraclete (advocate), but in the Gospel According to John, such references are to the Spirit. John 14:16, however, refers to "another Counselor." This discrepancy can be resolved by interpreting Jesus with his disciples as their advocate with another to come (the Spirit), and, in I John 2:1–2, the risen Lord becomes the advocate for the expiation of all sin. Righteousness and faith are emphasized in chapters 4–5, and again these characteristics are those of the children of God, who will finally in the end-time be like him who gave the promise, the commandment, and the joy of love.

The emphasis on love in I John

The Second Letter of John. II John warns a specific church (or perhaps churches), designated as "the elect lady and her children," against the influence of the Docetic heresy combatted in I John, whose proponents lured Christians from "following the truth, just as we have been commanded by the Father." In II John, as in the Gospel According to John and I John, the light–darkness images are similar to those of the Dead Sea Scrolls. To "walk in the truth" in II John is to reject heresy and follow the doctrine of Christ.

The Third Letter of John. III John, addressed to Gaius, shows that the writer is concerned about and has responsibility as presbyter for the missionaries of the church. It is somewhat of a short note concerned with church discipline, encouraging hospitality to true missionaries, and thus not unconnected with true doctrine and the command of love.

The Letter of Jude. The Letter of Jude, after a salutation that attributes it to Jude, the brother of James, and addresses itself to the church as a whole, develops the theme of the short letter—a polemic against heretics who have abandoned the transmitted traditional faith and who will thus be judged by the Lord. They deny Christ, and punishment similar to that of Sodom and Gomorrah in the Old Testament for such a denial is threatened. Heretical beliefs led to various sins and libertinism, and the judgment that will come upon them is cited from *Enoch* 1:9, demonstrating that this short letter reflects the post-biblical Jewish apocalyptic train of thought in the early Christian era.

The purpose and authorship of Jude

"Jude, a servant of Jesus Christ and brother of James" is probably meant pseudepigraphically to relate this Jude to James the brother of the Lord so that this Jude is also a brother of the Lord. This, however, is impossible because the letter reflects a later time. Verse 17 refers to "the predictions of the apostles of our Lord Jesus Christ" concerning mockers and sinners. Thus, the author is recalling a former time that was prophesied regarding the heresies and trials of the end-time. Such a bearer of apostolic tradition is violently attacking heresy in the interest of transmitted traditional faith. Again, it would appear that the letter is pseudepigraphic and may have originated in Syria or Asia Minor.

The author struggles forcefully against heretics who deny God and Christ and attempts to strengthen his readers in their fight against such heresy that leads to wickedness and disorder. Libertinism is a characteristic of such heresy and the punishment of the heretics will be similar to that which befell the unfaithful in the Old Testament patriarchal times. Only steadfastness in faith, true doctrine, and prayer can lead to mercy, forgiveness, restoration, and final salvation. An attempt to bring the erring to repentance may save them. The letter concludes with a typical doxology.

The form is less a catholic letter than a declared position that lays down general rules. The date is probably near the end of the 1st century and before II Peter, which draws upon it.

THE REVELATION TO JOHN

The Revelation (*i.e.*, Apocalypse) to John is an answer in apocalyptic terms to the needs of the church in time of persecution, as it awaits the end-time expected in the near future. The purpose of the book is to encourage and admonish the church to be steadfast and endure. The form of an apocalypse shows affinities with contemporary Jewish, Oriental, and Hellenistic writings in which problems of the end of the world and of history are linked both with prophecy of an eschatological nature and with "sealed" secret mysteries. Such revelations are traditionally received in trances, characterized by strange symbols, numbers, images, and parables or allegories that represent people and historical situations. Apocalypticism is essentially dualistic, presenting the present eon as evil and the future as good, with an ultimate battle between the divine and the demonic to be won only after one or more cosmic catastrophes. The aim of apocalyptic literature is to depict in the age of present tribulation a knowledge of a future glorious victory and vindication, thus giving hope and assurance.

Apocalyptic themes

In Revelation it is God who gives the revelation to

Jesus Christ to be shown by Christ through an angel to his servant John, in exile on the island of Patmos, in order that John become his seer and prophet to the church. John is to write down what he has seen, what is, and what is to come. In contradistinction to most Jewish apocalyptic works, Revelation is not pseudonymous and John is to give finally unsealed, clear prophecy related to the present and to the end-time.

As in the rest of the New Testament, the starting point of eschatological hope is the saving act of God in Jesus, a historical centre pointing toward historical developments that will bring about the establishment of God's kingdom and vindication of his people, ransomed by the blood of Christ, the Lamb who was slain. It provides certainty and encouragement with the example of the faithfulness of those who have already witnessed unto death (martyrs) and their reward—special inheritance in the eternal kingdom.

The content of Revelation

After the introduction, Revelation continues first as a series of seven letters to seven churches in the province of Asia, thence to the whole church with an epistolary introduction and, after the apocalypse proper, an epistolary blessing as the last verse. The letters sent from the heavenly Christ through John (chapters 2 and 3) exhort, comfort, or censure the churches according to their condition under persecution or danger of heresy. From chapters 4–22 there are series of visions in three main cycles, each recapitulating but expanding the former in greater and clearer detail with groups of seven symbols predominating in each (seals, chapters 6–7; trumpets, chapters 8–10; and bowls, chapters 15–16). This material is interspersed with visions of God in His heavenly council, various visions of catastrophe and of Satan, the destroyer, the appearance of two witnesses and other martyr examples to spur the church to endurance, the victory of the archangel Michael over the dragon (Satan) by the blood of the Lamb (Christ), and the representation of the powers of emperor cult and false prophecy as beasts who bring destruction to the unfaithful in God's judgment. A heavenly woman who bears a messianic son is threatened by a dragon. Her child is carried up to heaven by God, and she escapes by hiding in a place prepared for her by God. The beasts who appear persecute the Christians and the "number" signifying the second beast is that of a man, "666" (or, in a variant reading "616") probably indicating the emperor Nero. God's triumph in history is depicted in his judgment on the harlot Babylon (Rome), and the final consummation portrays the victory of Christ over the Antichrist and his followers. In chapter 20 the thousand-year reign of Christ with those who witnessed unto death is depicted. Satan, again loosed, is vanquished by fire from heaven with the beasts (empire power and false prophet), and the last judgment leads to a new heaven and a new earth, the new Jerusalem. This writing is, thus, a prophetic-apocalyptic work.

In summary, the seer reminds the reader that the words, because they are of God, are trustworthy and true. The motif that the Lord is coming soon is again repeated. This reflection of the early Christian watchword suggests a sacred liturgical style. The last verse is the closing benediction—perhaps not only of the letters in the beginning of Revelation but of the whole of Revelation, which was to be read aloud in a worship setting.

Authorship and style

After AD 70 (the fall of Jerusalem), apocalypticism was introduced into Asia Minor and *c.* 80–90 a prophetic circle was formed near Ephesus. Its leader was John, a prophet, who might well have been the author of Revelation, which is deeply steeped in apocalyptic traditions. The "Johannine circle" bearing the tradition of John, the Apostle of the Lord, and from which emerged the Gospel and letters bearing his name, might have been a continuation of the prophetic conventicle of Ephesus in which John was prominent. The various writings do not have to be consistent except in their basic faith in Jesus Christ; and, as the situations to which they addressed themselves were different, different styles and content were required. The seer was probably involved in an actual historical situation in the late 80s under Domitian, a time when there was open conflict between the church and the Roman state. There is a tradition supported by Irenaeus, a 2nd-century bishop of Lyons, that in this persecution punishment was death or banishment. John's prominence might have led to banishment to Patmos, an isle off the coast of Asia Minor, from his homeland in or around Ephesus. From Patmos he wrote a circular letter to the churches in Asia.

Though the style of Revelation is certainly eclectic in form and content, containing elements of a heavenly epistle and with more than three-fourths of the rest made up of prophetic-apocalyptic forms from varied sources, it reflects a systematic and careful plan. Even the apocalyptic, however, is "anti-apocalyptic" in that the seer's message is open and the mysteries serve not to conceal but to heighten what is seen and to be expected. Apocalyptic schemata and motifs are, however, used toward this purpose, and allegorical incorporation of sources is more a demonstration of the true, ultimate message than a literary device. Blurred images (*e.g.*, God, Christ, and angels; chiliastic [1,000-year] eras and temporal duplications; as well as interpretations) are part of the apocalyptic style, but a current concrete historical situation is the foundation. Revelation is written in fantastic imagery, blending Jewish apocalyptic, Babylonian mythology, and astrological speculation. It is pictorial, dramatic, and poetic.

Revelation contains long sections characterized by Greek that is grammatically and stylistically crude, strangely Hebraized to give a unique almost Oriental colour. This may have been deliberate. Although Revelation is replete with Old Testament allusions, there are no direct quotations, and this may reflect the seer's conviction that the work is a direct revelation from God. In other sections the poetry of Revelation might stem from the seer's experience in the heavenly throne room of God, from hearing the hymns of the angelic host, or from his recollection on Patmos of the liturgical practice of the church. The image of the Bride and wedding feast together with the "Come, Lord Jesus!" have associations with the eucharistic liturgy of the early church.

The recapitulation of images and the use of numbers

The recapitulations of the seven seals, trumpets, and bowls may be deliberate schematization. The purpose of such repetition and increasing revelation can be a way of heightening enthusiasm to encourage the church.

Mysterious numbers and divisions (such as 7, 3, 12) recur and are part of the theme of assurance, because God has numbers in their order as a sign of his plan of salvation, turning chaos to orderly cosmos. The mysterious name of the second beast, 666 in 13:18, can be calculated by "gematria," assigning their numerical values to letters of the word and summing them up. The most adequate solution is Nero (the numerical value of the Hebrew letters for *Caesar Neron* equals 666), a demonic Nero *redivivus* (revived), who returns from the dead as Antichrist. Astronomy and astrology have also been applied to Revelation in terms of the signs of the zodiac or a calendar of feasts and seasons as keys to understanding its structure, because it is God who orders the times and seasons.

Martyrological themes

Two witnesses described in chapter 11 have been assumed to be Elijah and Moses, Peter and Paul, or simply two examples of martyrs through whom God shows His punishment of the wicked and vindication of the righteous to his glory. There are strong martyrological themes throughout Revelation, and it seems to stand on the border line of the point at which the word witness (*martys*) became a technical term for a witness unto death, or martyr. The cosmic battle in heaven is fought by those willing to give their lives, who mix their blood with the blood of the Lamb, whose blood "ransomed men for God." The writer of Revelation based his hope for the church on perseverance, on endurance even to death, and on what the future will bring when the church will live with the glorified Christ, slain as a lamb. The harlot of Babylon will be destroyed and the church will endure; Babylon falls and the new Jerusalem, the city of God that is to come, is depicted in all its glory. These are the hopes to strengthen the persecuted church, assurance

that God will soon triumph. With trumpet call and heavenly voices there is the joyful promise that "The kingdom of the world has become the kingdom of our Lord and of his Christ, and he shall reign for ever and ever."
(K.St./E.T.Sa.)

IX. New Testament Apocrypha

NATURE AND SIGNIFICANCE

The title New Testament Apocrypha may suggest that the books thus classified have or had a status comparable to that of the Old Testament Apocrypha and have been recognized as canonical. In a few instances such has been the case, but generally these books were accepted only by individual Christian writers or by minority heretical groups. The word apocryphal (secret) is applied to Gnostic traditions and writings both by Gnostics (adherents to a heretical dualistic religious system) and by their critics; from the 2nd century, for example, come the *Apocryphon* (secret book) *of John*. In the 4th century the word referred to books not publicly read in churches. It meant apocryphal in the modern sense (*i.e.*, fictitious) only by implication, as when the church historian Eusebius speaks of some of "the so-called secret books" as forgeries composed by heretics.

Pseudepigraphical gospels, acts, letters and apocalypses

Like the New Testament books themselves, the New Testament apocryphal books consist of gospels, acts, letters, and apocalypses. The apocryphal writings, however, are almost exclusively pseudepigraphical—*i.e.*, written in the name of the apostles or disciples or concerning individual apostles. In general, they were created after and in imitation of the New Testament books but before the time when a relatively restricted canon, or list, of approved books was being formulated. They arose chiefly during the 2nd century, when the lines between orthodoxy and heresy were not absolutely fixed and when popular piety seems to have been rather freely expressed. What these works tell about Jesus and his disciples resembles the imaginative Midrashic (didactic commentarial) retelling of Old Testament stories among Jewish teachers.

As the New Testament canon was gradually given definite shape, these apocryphal books came to be excluded, first from public reading in churches, then from private reading as well. With the development of creeds and of systematic theologies based on the nascent canon, the apocryphal books were neglected and suppressed. Most of them have survived only in fragments, although a few have been found in Greek and Coptic papyri from Egypt. They are valuable to the historian primarily because of the light they cast on popular semi-orthodox beliefs and on Gnostic revisions of Christianity; occasionally, they may contain fairly early traditions about Jesus and his disciples. In the 3rd century, Neoplatonists (followers of the philosopher Plotinus, who advocated a system of levels of reality) joined Christians in attacking such books as "spurious," "modern," and "forged."

The difficulties the New Testament apocryphal books caused at the end of the 2nd century are well illustrated in a letter by Serapion, bishop of Antioch. He stated that he accepts Peter and the other apostles "as Christ" but rejects what is falsely written in their name. When some Christians showed him the *Gospel of Peter*, he allowed them to read it, but after further investigation he discovered that its teaching about Christ was false, and he had to withdraw his permission.

Categories of authenticity and spuriousness

In the early 4th century Eusebius himself found it difficult to create categories for the various books then in circulation or used by earlier authors. He seems to have concluded that the books could be called "acknowledged," "disputed," "spurious," and absolutely rejected. Thus, the *Acts of Paul*, the *Apocalypse of Peter*, and the *Gospel According to the Hebrews* were rather well attested, and he called them spurious but disputed. He definitely rejected books used by heretics but not by church writers: the gospels ascribed to Peter, Thomas, and Matthias, and the *Acts of Andrew*, John, and other apostles. About a century earlier, the North African theologian Tertullian had written about how a presbyter who wrote the *Acts of Paul* had been deposed.

Without reference to the standards of canonicity and orthodoxy gradually being worked out by the churches of the 2nd through 4th centuries, it is evident that many of these books reflect the kinds of rather incoherent Christian thought that church leaders were trying to prune and shape from the 1st century onward. Often such works represented what was later viewed as inadequate orthodoxy because the views presented had become obsolete. All the apocrypha taken together show the variety of expression from which the canon was a critical selection.

THE NEW TESTAMENT APOCRYPHAL WRITINGS

This article will classify these documents in relation to their literary forms: gospels, acts, letters, and apocalypses.

Gospels. A few papyrus fragments come from gospels not known by name (*e.g.*, Egerton Papyrus 2, Oxyrhynchus Papyrus 840, Strasbourg Papyrus 5–6). There are also the *Gospel* produced in the 2nd century by Marcion (a "semi-Gnostic" heretic from Asia Minor), who removed what he regarded as interpolations from the Gospel According to Luke; the lost Gnostic *Gospel of Perfection;* and the *Gospel of Truth*, published in 1956 and perhaps identical with the book that Irenaeus (*c.* 185), bishop of Lyon, said was used by the followers of Valentinus, a mid-2nd-century Gnostic teacher. The *Gospel of Truth* is a mystical–homiletical treatise that is Jewish–Christian and, possibly, Gnostic in origin. In addition, there were gospels ascribed to the Twelve (Apostles) and to individual apostles, including the *Protevangelium of James*, with legends about the birth and infancy of Jesus; the lost Gnostic *Gospel of Judas* (Iscariot); the *Gospel of Peter*, with a legendary account of the resurrection; the *Gospel of Philip*, a Valentinian Gnostic treatise; the *Gospel of Thomas*, published in 1959 and containing "the secret sayings of Jesus" (Greek fragments in Oxyrhynchus papyri 1, 654, and 655); and an "infancy gospel" also ascribed to Thomas. Beyond these lie gospels ascribed to famous women, namely Eve and Mary (Magdalene), or named after the groups that used them: Ebionites (a Jewish Christian sect), Egyptians, Hebrews, and Nazarenes (an Ebionite sect).

Popular stories of travel and adventure

Acts. The various acts, close in form and content to the contemporary Hellenistic romances, turned the apostolic drama into melodrama and satisfied the popular taste for stories of travel and adventure, as well as for a kind of asceticism generally rejected by Christian leaders. They dealt with the adventures of apostles and other early Christian personages: Andrew (including the *Acts of Andrew and Matthias Among the Cannibals*), Barnabas (a companion of St. Paul), Bartholomew, John (with semi-Gnostic traits), Paul (including the *Acts of Paul and Thecla*, with a Christian version of the story of Androcles and the lion), Peter—with the apostle's question to the risen Lord, "Lord, where are you going?" ("Domine, quo vadis?") and Peter's crucifixion upside down, Philip, Thaddaeus (his conversion of a king of Edessa), and Thomas (with the Gnostic "Hymn of the Pearl").

Letters. Among the apocryphal letters are: a 2nd-century *Epistula Apostolorum* ("Epistle of the Apostles"; actually apocalyptic and antiheretical), the *Letter of Barnabas*, a lost *Letter of Paul to the Alexandrians* (said to have been forged by followers of Marcion), the late-2nd-century letter called "III Corinthians" (part of the *Acts of Paul* and composed largely out of the genuine letters of Paul), along with a letter from the Corinthians to Paul, and a Coptic version of a letter from Peter to Philip. There is also a famous forgery purporting to have been written by Jesus to Abgar, king of Edessa (noted in Eusebius, *Church History* I. 13).

Apocalypses. Other than the Revelation to John, which some early Christian writers rejected, there are apocalypses ascribed to two Jameses, the Virgin Mary, Paul, Peter, Philip, Stephen, and Thomas. Only the *Apocalypse of Peter* won any significant acceptance and is important for its vivid description of the punishment of the wicked.

In addition, it should be noted that there were apocryphal books with titles not so closely related to the New

Testament. Among these are: the *Didachē*, or *Teaching of the Twelve Apostles* (and its later revisions, such as the *Didascalia Apostolorum*, or the "Teaching of the Apostles," and the *Apostolic Constitutions*), and the *Kerygma of Peter*, a favourite at Alexandria, as well as various Gnostic works, such as *The Dialogue of the Redeemer*, *Pistis Sophia* ("Faith-Wisdom"), and the *Sophia Jesu Christi* ("Wisdom of Jesus Christ"). From the 5th century there is even a *Testamentum Domini* ("Testament of the Lord"), an expansion of the 2nd–3rd-century Roman Church leader and theologian Hippolytus' *Apostolic Tradition.* (R.M.G.)

X. Biblical literature in liturgy

BIBLICAL LITERATURE IN THE LITURGY OF JUDAISM

The liturgy of Judaism is that of the synagogue, which arose during and after the Babylonian Exile of 586–538 BCE and gradually replaced the Temple cult as the spiritual centre of Jewish life. The Hebrew biblical canon and the liturgy of the synagogue, to a great extent, grew up together.

The synagogue liturgy

Because the synagogue arose in a land separated from the Jerusalem Temple with its sacrificial emphasis and its priestly class, worship in the synagogue differed from what went before it in several respects. A local congregation worshipped together on a certain day of the week in a place set apart for that purpose, rather than primarily on special festival days and periods. The people worshipped without priest or cultic sacrifice, yet consciously as a community within a larger covenant fellowship and in response to a divine word that was written down in a holy scripture. Bible reading and interpretation, the singing of psalms, and prayers, both corporate and individual, were the staple content of the liturgy. The ancient synagogue liturgy has come down to the present in two books: the *Siddur*, or daily prayer book, and the *Mahzor*, or festival prayer book.

The biblically prescribed rhythm of days, weeks, months, and years gave order to the lives of the people. The Bible became familiar to old and young by being read aloud in the synagogue, and no part of worship was esteemed more highly than the reading of scripture. The Torah, the first five books of the Bible, is handwritten on a scroll. Viewed as the holiest object in the synagogue, it is kept in a sacred cabinet called the ark. Special prayers and ceremonies accompany its being taken out and replaced in the ark, and during the course of the year it is read in its entirety at the sabbath services. Torah portions are also read on the religious holidays.

A reading from the Prophets, called the Haftarah, follows each Torah reading. One of the five Megillot (Scrolls) is read on certain holidays: the Song of Solomon at Pesaḥ (Passover), the Book of Ruth at Shavuot (Weeks), Lamentations of Jeremiah at Tisha be-Av (Av 9), Ecclesiastes at Sukkot (Tabernacles), and the Book of Esther at Purim (Lots). The Book of Jonah is read on the afternoon of Yom Kippur (Day of Atonement). Psalms are said or sung in every service. From the chanting of biblical texts, especially the Psalms, the music of the synagogue's cantor has developed into an incomparable art form (see also JEWISH RELIGIOUS YEAR).

BIBLICAL LITERATURE IN THE LITURGY OF CHRISTIANITY

Eastern Orthodoxy. The first Christians were Jews, and they worshipped along with other Jews in the synagogue. The earliest Gentile converts also attended the synagogue. When Christians met outside the synagogue, they still used its liturgy, read its Bible, and preserved the main characteristics of synagogue worship. Every historic liturgy is divided into (1) a Christian revision of the sabbath service in the synagogue and (2) a celebration of Jesus' Last Supper with his disciples as a fulfillment of the Passover and a new covenant with a newly redeemed people of God. Thus, the church was never without traditional forms of worship.

For over 100 years Christians had no authorized New Testament, the Old Testament being read, as had been done previously, in the worship service. By the middle of the 2nd century, however, Christian writings also were read in the Sunday service. The Old Testament, generally in its Greek translation (the Septuagint), was the Bible from which the Gospel was preached. Its reading preceded that of the Christian writings and was far more extensive than it is in modern Christian churches.

The use of the Old and New Testament texts

As the liturgies grew longer and more elaborate, the biblical readings were reduced, and the New Testament gradually displaced the Old Testament. No Old Testament lesson remained in the Greek or Russian liturgy or in the Roman mass, though it has been reintroduced in the 20th century in most liturgies. All liturgies have at least two readings from the New Testament: one from a letter or other (non-Gospel) New Testament writing, and one from a Gospel, in that order. The Eastern liturgies all honour the Gospel with a procession called the Little Entrance. This action is accompanied by hymns and prayers that interpret the Gospel as the coming of Christ to redeem the world.

The Eastern liturgies, especially after the great theological controversies of the first four centuries, have favoured composed texts of prayers, hymns, and choral anthems that summarize the thought of many biblical passages, thus becoming short sermons or confessions of faith. The Nicene Creed (4th century) itself is one such text, in contrast with the Shema ("Hear, O Israel"—a type of creed) in Judaism, which consists of verbatim passages from Deuteronomy and Numbers.

The Divine Liturgy of the Eastern Orthodox churches contains many such composed texts, such as prayers that proclaim Orthodox theology (*e.g.*, the "Only begotten Son and Word of God" following the second antiphon). Isaiah, chapter 6, verse 3, ("Holy, holy, holy is the Lord of hosts; the whole earth is full of his glory"), used in the Jewish Kedusha (Glorification of God), generates two separate texts in the Eastern liturgy: the Trisagion (a solemn threefold acclamation to God) at the Little Entrance and the Greek original of the "Holy, holy, holy" in the eucharistic liturgy.

Psalms are sung extensively at the daily hours of prayer in the East as in the West. At the beginning of the Sunday service, entire psalms or more than one psalm are sometimes sung. More often, however, a psalm verse or two are combined with other material into a composite text of a hymn or anthem. A mosaic of selected psalm verses may be used either as a text for music or a spoken prayer. Most characteristic of all, especially in the Greek Church's tradition, however, is the freely composed and imaginative hymn text, based on a biblical incident or person, or an extended paraphrase of a passage of scripture. In addition to such biblically based psalms and other hymns, there are the famous Cherubic Hymn of the Greek and Russian liturgies and the original texts of hymns that have become well known in the Western churches—*e.g.*, "O gladsome light of the Father immortal," and "Let all mortal flesh keep silent."

Roman Catholicism. Liturgical worship in both Judaism and Christianity is an action that moves within the framework of biblical ideas and explains itself in biblical language. Preoccupied with really different views from opposite windows, Jews and Christians have often overlooked the house they share. This has likewise been true of differences between Eastern and Western Christians.

Development of the Latin rite

At Rome, the liturgy was sung and said in Greek till the 4th century and was probably more like the liturgy of Syria at that time than that of Rome after the 16th century. The Latin rite developed many distinctive features, but what happened in Rome happened also to some extent in the East. The biblical readings at mass were reduced to two: the first reading, formally called the Epistle, was usually from an apostolic letter but sometimes from the Acts of the Apostles or even the Old Testament, and the second was a Gospel passage selected as appropriate for that particular day in the Church Year. The West, like the East, retained the Jewish week and developed a yearly cycle of Easter–Pentecost and Christmas–Epiphany celebrations with appropriate biblical selections. The development of the Church Year became so elaborate in the West, however, that the Roman calendar provided for every day in the year.

In the West as in the East, monastic and other religious communities observed the daily hours of prayer, in which there was little Bible reading as such but a great deal of corporate praying as well as the reading or singing of psalms. The Roman canonical hours were further enriched with homilies and legends from many sources, with Latin metrical hymns, and with biblical canticles, including a daily singing of the early Christian songs that are quoted in the Gospel According to Luke: the "Benedictus" ("Song of Zechariah") in chapter 1, verses 68–79, at Lauds (morning prayer), the "Magnificat" ("Song of Mary") in chapter 1, verses 46–55, at Vespers (evening prayer), and the "Nunc Dimittis" ("Song of Simeon") in chapter 2, verses 29–32, at Compline (prayer at the end of the day). The great anonymous canticle called the "Te Deum," a vast array of biblical images ascribing praise and glory to God, is sung every day at Matins (an early morning prayer).

The mass is an abbreviation of a much longer liturgy. Many items are mere vestiges of more elaborate actions or texts. The psalms once sung at the entrance, for example, have been reduced to a traditional form of a sung text: an antiphon of one or two verses from a psalm, the first verse of the psalm, the "Glory be to the Father," and the antiphon repeated. The same has occurred in other parts of the mass. Psalms were once interspersed among the readings of scripture. The traditional gradual was a formalized text sung between the Epistle and Gospel, but in the reformed mass it becomes a responsorial psalm between the first and second readings. The short texts at the Offertory (offering of the bread and wine) and Communion are fragments in biblical language, but they are also masterpieces of the Latin genius for brevity, clarity, and order—as are the inimitable Latin collects (prayers), each basing its definite petition on an equally definite biblical revelation.

For centuries the mass was heard only in Latin and repeated the same readings on the same days every year, with the result that only a limited number of unconnected passages were heard in church. The second Vatican Council (1962–65) approved the plan of having a three-year cycle of biblical readings, providing an Old Testament lesson for every mass, a more nearly continuous reading from one of the Gospels each year, and a reading from one of the letters or other New Testament books over a period of weeks.

Protestantism. The term Protestant covers so wide a variety of theological views and religious and cultural groups and so many different ways of worshipping and using the Bible in worship that it is virtually impossible to say anything about the liturgy or the Bible's place in worship that would be true of all Protestants. Among Anglicans, what was said of the Bible in the Roman Catholic liturgy would generally apply. It would also apply to most Lutherans in the 20th century, but not to all Lutherans. On the other hand, there have been and are Protestants who claim or tacitly assume that nothing but the Bible should be used in worship. The use of the Bible in Protestant liturgy lies between these extremes.

Variations in Protestant liturgies

In the 16th century, the New Testament was appealed to as a guide for reforming the worship as well as the doctrine of the time. Because the worship reflected in the New Testament is synagogue worship, Protestant worship of the less liturgical kind became, in many respects, a return to synagogue worship. Protestants separated the two services (instructional and Eucharistic) that had been joined together in the historic liturgy of Christendom. The Protestant Sunday service is the Liturgy of the Learners, a new revision of the synagogue liturgy. It centres in the biblical word read and preached. The congregation worships in anticipation of and response to the scriptural word. Praise becomes corporate only in hymns sung by the congregation, and prayer voices human need and misery as revealed in the Bible and claims the promises heard there.

The absence of a developed liturgy generally limits the amount and variety of scripture read in the course of a year, as well as the forms of congregational participation. On the one hand, it limits worship to the resources and skill of local ministers, but, on the other hand, it also leaves a freedom to choose what is useful from any source—this has become an increasing practice in almost every Protestant church in the 20th century. Such freedom has been welcomed by many in the latter part of the 20th century—when all Protestant and Catholic liturgies seem likely to change without much advance notice (see also CHURCH YEAR). (H.G.D.)

BIBLIOGRAPHY

Nature and significance: General articles and notes in *The Oxford Annotated Bible* (1962), *The Jerusalem Bible* (1966), and the *Genesis* volume of *The Anchor Bible*, by E.A. SPEISER (1964); E. H. GOMBRICH, *The Story of Art*, 12th rev. ed. (1972); ABRAHAM J. HESCHEL, *Man Is Not Alone* (1951), a classic statement of modern Judaism; ERICH AUERBACH, *Mimesis: Dargestellte Wirklichkeit in der abendländischen Literatur* (1946; Eng. trans., *Mimesis: The Representation of Reality in Western Literature*, 1953), a classic work; WILLIAM R. MUELLER, *The Prophetic Voice in Modern Fiction* (1959), religious themes interpreting Joyce, Camus, Kafka, Faulkner, Greene, and Silone; NATHAN A. SCOTT, JR. (ed.), *The Tragic Vision and the Christian Faith* (1957), essays by 12 writers on faith and the tragic dimension of existence.

Old Testament canon, texts, and versions: OTTO EISSFELDT, *Einleitung in das Alte Testament*, 3rd ed. (1964; Eng. trans., *The Old Testament: An Introduction*, 1965); *The Cambridge History of the Bible* (CHB), 3 vol. (1963–70). (*The Canon*): FRANTS BUHL, *Kanon und Text des Alten Testaments* (1891; Eng. trans., *Canon and Text of the Old Testament*, 1892); MAX L. MARGOLIS, *The Hebrew Scriptures in the Making* (1922); HERBERT E. RYLE, *The Canon of the Old Testament*, 2nd ed. (1895); SOLOMON ZEITLIN, "An Historical Study of the Canonization of the Hebrew Scriptures," *Proceedings of the American Academy for Jewish Research*, pp. 121–158 (1932). (*Textual criticism, texts and manuscripts, and early versions*): FRANK MOORE CROSS, *The Ancient Library of Qumrân and Modern Biblical Studies*, 2nd ed. (1961); "The History of the Biblical Text in the Light of Discoveries in the Judaean Desert," *Harvard Theological Review*, 57:281–299 (1964); and "The Contribution of the Qumrân Discoveries to the Study of the Biblical Text," *Israel Exploration Journal*, 16:81–95 (1966); CHRISTIAN D. GINSBURG, *Introduction to the Massoretico: Critical Edition of the Hebrew Bible* (1897, reprinted 1966); MOSHE H. GOSHEN-GOTTSTEIN, *Linguistic Structure and Tradition in the Qumran Documents* (1958); "Theory and Practice of Textual Criticism," *Textus*, 3:130–158 (1963); and *The Book of Isaiah: Sample Edition with Introduction* (1965); MOSHE GREENBERG, "The Stabilization of the Text of the Hebrew Bible," *Journal of the American Oriental Society*, 76:157–167 (1956); PAUL KAHLE, *The Cairo Genizah*, 2nd ed. (1959); FREDERICK G. KENYON, *The Bible and the Ancient Manuscripts*, 5th ed. rev. (1958); HARRY M. ORLINSKY, "The Textual Criticism of the Old Testament," in GEORGE E. WRIGHT (ed.), *The Bible and the Ancient Near East*, pp. 113–132 (1961); BLEDDYN J. ROBERTS, *The Old Testament Text and Versions* (1951); and "The Old Testament: Manuscripts, Text and Versions," *CHB*, vol. 2, pp. 1–26 (1969); P.W. SKEHAN, "Qumran and the Present State of Old Testament Text Studies," *Journal of Biblical Literature*, 78:21–25 (1959); S. TALMON, "Aspects of the Textual Transmission of the Bible in the Light of Qumran Manuscripts," *Textus*, 4:95–132 (1964); ERNST WURTHWEIN, *Der Text des Alten Testaments* (1952; Eng. trans., *The Text of the Old Testament*, 1957). (*Later and modern versions–English versions*): DAVID DAICHES, *The King James Version of the English Bible* (1941, reprinted 1968); MARGARET DEANESLY, *The Lollard Bible and Other Medieval Biblical Versions* (1920, reprinted 1966); HERMAN HAILPERIN, *Rashi and the Christian Scholars* (1963); WILLIAM F. MOULTON, *The History of the English Bible*, 5th ed. (1911); ALFRED W. POLLARD, *Records of the English Bible* (1911); and with G.R. REDGRAVE, *A Short-Title Catalogue of Books Printed in England, Scotland, and Ireland and of English Books Printed Abroad 1475–1640* (1926, reprinted 1969); B.F. WESTCOTT, *A General View of the History of the English Bible*, 3rd ed. rev. by W.A. WRIGHT (1905). (*Continental versions and non-European versions*): THOMAS H. DARLOW and HORACE F. MOULE, *Historical Catalogue of the Printed Editions of the Holy Scripture in the Library of the British and Foreign Bible Society*, 2 vol. (1903–11); JOSEF SCHMID (ed.), "Moderne Bibelübersetzungen," *Zeitschrift für katholische Theologie*, 82:290–332 (1960).

Old Testament history: Two current histories of Israel exhibit the full range of historiographical problems and methods relating to the subject: JOHN BRIGHT, *A History of Israel* (1959); and MARTIN NOTH, *Geschichte Israels*, 3rd ed. (1956; Eng. trans., *The History of Israel*, 1958). They differ mainly in where they begin; Bright begins with Abraham, Noth with

the federation of tribes that calls itself Israel in the land of Canaan. They disagree about the demonstrability of such a community in the pre-Canaanite times because of their respective assessment of the character of the Pentateuch. Bright assumes that it was intended as a history concerned to record the early past, while Noth assumes that its thematic traditions were intended to define and celebrate the identity of the later Israel and hence do not constitute a usable historical resource about its earliest beginnings. This whole methodological problem in Israelite historiography is lucidly discussed and illustrated in a little book by JOHN BRIGHT—*Early Israel in Recent History Writing: A Study in Method* (1956). For the use of archaeology, geography, and history of religion in the study of the history of Israel, see GEORGE ERNEST WRIGHT, *Biblical Archaeology*, rev. ed. (1962); LUC H. GROLLENBERG, *Atlas van de Bijbel*, 3rd ed. (1954; Eng. trans., *Atlas of the Bible*, 1956); YEHEZKEL KAUFMANN, *The Religion of Israel, from Its Beginnings to the Babylonian Exile* (1960); and HELMER RINGGREN, *Israelitische Religion* (1963; Eng. trans., 1966).

Old Testament literature: For various modern critical methods of studying the formation of the Old Testament, see the "Old Testament Series" of *Guides to Biblical Scholarship:* NORMAN C. HABEL, *Literary Criticism of the Old Testament*, GENE M. TUCKER, *Form Criticism of the Old Testament*, and WALTER E. RAST, *Tradition History and the Old Testament* (1971–72). Among general introductions, the most exhaustive is OTTO EISSFELDT (*op. cit.*), based mainly on literary criticism. The other methods are reflected to a somewhat greater extent in AAGE BENTZEN, *Introduction to the Old Testament*, 3rd ed. (1957); and in the briefer, less original but very readable work of ARTUR WEISER, *Einleitung in das Alte Testament*, 4th ed. (1957; Eng. trans., *The Old Testament: Its Formation and Development*, 1961). For pioneering research in tradition analysis of the Pentateuch and the Former Prophets, see MARTIN NOTH, *Überlieferungsgeschichte des Pentateuch*, 3rd ed. (1966; Eng. trans., *A History of Pentateuchal Traditions*, 1972), and *Überlieferungsgeschichtliche Studien* (1957); the latter deals with what its author calls "The Deuteronomic History," an envisioned work containing the books of Deuteronomy, Joshua, Judges, Samuel, and Kings. The contribution of form criticism to the understanding of the history of the Book of Psalms may best be approached through HERMANN GUNKEL, *The Psalms: A Form-Critical Introduction* (1967), a translation of his article in *Die Religion in Geschichte und Gegenwart* (2nd ed.) summarizing his seminal work in *Die Psalmen* (1926) and *Einleitung in die Psalmen* (1928). ELMER A. LESLIE, *The Psalms, Translated and Interpreted in the Light of Hebrew Life and Worship* (1949), is heavily dependent on Gunkel and illustrates his use of form criticism. The celebrated work of SIGMUND MOWINCKEL on the Psalter, culminating in his masterful *Offersang ob Sangoffer* (1951; Eng. trans., *The Psalms in Israel's Worship*, 2 vol., 1962), combines the methods of Gunkel with those of the comparative historian of religion and locates the setting for the production of most of the psalms in the cult of the Solomonic temple. The application of the newer methods to the study of the Latter Prophets is evident in the essays in HAROLD H. ROWLEY (ed.), *Studies in Old Testament Prophecy* (1950). The new approaches were deeply under the impact of HENRIK S. NYBERG, *Studien zum Hoseabuche* (1935). Other books that amplify the implications of his assumptions include: JOHANNES LINDBLOM, *Prophecy in Ancient Israel* (1962); CURT KUHL, *Israels Propheten* (1956; Eng. trans., *The Prophets of Israel*, 1960); and SIGMUND MOWINCKEL, *Prophecy and Tradition: The Prophetic Books in the Light of the Study of the Growth and History of the Tradition* (1946). ABRAHAM J. HESCHEL, *The Prophets* (1962), though of independent origin, belongs with those new interpretations of the prophetic materials. An old classic in a new edition, OLIVER S. RANKIN, *Israel's Wisdom Literature: Its Bearing on Theology and the History of Religion* (1936, reprinted 1969), presents Israel's wisdom literature in relation both to its extra-Israelite cultural connections and to the rest of Israel's heritage in the Old Testament. Two new approaches to the legacy of wisdom literature—through literary form and through theology—are presented, respectively, in R.B.Y. SCOTT, *The Way of Wisdom in the Old Testament* (1971); and GERHARD VON RAD, *Weisheit in Israel* (1970).

Intertestamental literature: Standard translations of the Jewish intertestamental literature are ROBERT H. CHARLES (ed.), *The Apocrypha and Pseudepigrapha of the Old Testament in English* (1913); and EMIL KAUTZSCH (ed.), *Die Apocryphen und Pseudepigraphen des Alten Testaments* (1900). PAUL RIESSLER, *Altjüdisches Schrifttum ausserhalb der Bibel* (1928), is indispensable because it contains translations of the fullest number of writings. The best translations of the Dead Sea Scrolls are GEZA VERMES, *The Dead Sea Scrolls in English* (1962); JOHANN MAIER, *Die Texte vom Toten Meer* (1960); and ANDRE DUPONT-SOMMER, *Les Écrits esséniens découverts près de la Mer Morte*, 3rd ed. (1964). ALBERT-MARIE DENIS, *Introduction aux Pseudépigraphes grecs d'Ancien Testament* (1970), does not treat the Apocrypha and is important mainly for its bibliography. Basic books dealing with intertestamental literature are R.H. PFEIFFER, *History of New Testament Times, with an Introduction to the Apocrypha* (1949); EMIL SCHURER, *Geschichte des jüdischen Volkes im Zeitalter Jesu Christi*, 3rd–4th ed., 3 vol. (1898–1901; Eng. trans., *A History of the Jewish People in the Time of Jesus Christ*, 2nd and rev. ed., 5 vol., 1885–91); and ROBERT H. CHARLES, *Religious Development Between the Old and the New Testaments* (1914). Still interesting is ROBERT TRAVERS HERFORD, *Talmud and Apocrypha* (1933, reprinted 1971). Information about the library of the Dead Sea Scrolls is in two books: JOZEF T. MILIK, *Dix Ans de découvertes dans le désert de Juda* (1957; Eng. trans., *Ten Years of Discovery in the Wilderness of Judaea*, 1959); and FRANK MOORE CROSS, *The Ancient Library of Qumrân and Modern Biblical Studies*, 2nd ed. (1961). A fragment of Ben Sira from antiquity was published by YIGAEL YADIN, *The Ben Sira Scroll from Masada, with Introduction, Emendations and Commentary* (1965). The best book about Jewish eschatology is PAUL VOLZ, *Die Eschatologie der jüdischen Gemeinde im neutestamentlichen Zeitalter* (1934). On Apocalyptic and Messianism, see HAROLD H. ROWLEY, *The Relevance of Apocalyptic*, 3rd ed. (1963); DAVID S. RUSSELL, *The Method and Message of Jewish Apocalyptic, 200 BC–AD 100* (1964); SIGMUND MOWINCKEL, *Han som kommer* (1951; Eng. trans., *He That Cometh*, 1954); ERIK SJOBERG, *Der Menschensohn im äthiopischen Henochbuch* (1946); and A.S. VAN DER WOUDE, *Die messianischen Vorstellungen der Gemeinde von Qumrân* (1957).

New Testament canon, texts, and versions: (*Canon*): For the relevant primary texts on the history of the canon, see DANIEL J. THERON (ed.), *Evidence of Tradition* (1957), with selected source material in Greek or Latin with English translation; and JAMES STEVENSON (ed.), *A New Eusebius* (1957). For introductions, see ALEXANDER SOUTER and C.S.C. WILLIAMS, *The Text and Canon of the New Testament*, 2nd ed. rev. (1954); and ROBERT M. GRANT, *The Formation of the New Testament* (1965). For a phenomenological approach, see GERARDUS VAN DER LEEUW, *Phänomenologie der Religion*, 2nd ed., 2 vol. (1956; Eng. trans., *Religion in Essence and Manifestation*, 2nd ed., 2 vol., 1963), ch. 64. (*Texts*): The major text for further study is BRUCE H. METZGER, *The Text of the New Testament* (1964). (*Translations*): On translation in general, see REUBEN A. BROWER (ed.), *On Translation* (1959). For translation of the Bible into English, see FREDERICK F. BRUCE, *The English Bible: A History of Translations from the Earliest English Versions to the New English Bible*, 2nd ed. (1970).

New Testament history: (*Jewish culture and history*): Standard works are R.H. PFEIFFER (*op. cit.*); and GEORGE F. MOORE, *Judaism in the First Centuries of the Christian Era*, 3 vol. (1927–30). (*Qumrān, the Dead Sea Scrolls*): FRANK MOORE CROSS (*op. cit.*); on Qumrān and New Testament problems, see KRISTER STENDAHL (ed.), *The Scrolls and the New Testament* (1958). (*Graeco-Roman culture and history*): WILLIAM W. TARN, *Hellenistic Civilisation*, 3rd ed. rev. (1952). For a broad cultural comparison, see ERIC R. DODDS, *Pagan and Christian in an Age of Anxiety* (1965). (*Pauline chronology*): The debate can be best assessed by comparing JOHN KNOX, *Chapters in a Life of Paul* (1950) with DIETER GEORGI, *Die Geschichte der Kollekte des Paulus für Jerusalem* (1965).

New Testament literature: The following works are useful for commentary, survey articles, and bibliographic material: GEORGE A. BUTTRICK (ed.), *The Interpreter's Bible*, especially vol. 1, 7, and 12 (1952–57); MATTHEW BLACK (ed.), *Peake's Commentary on the Bible*, 2nd ed. (1962); and RAYMOND E. BROWN, JOSEPH A. FITZMYER, and ROLAND E. MURPHY (eds.), *The Jerome Biblical Commentary* (1968). WERNER G. KUEMMEL, *The New Testament: The History of the Investigations of Its Problems* (1972), covers the whole history of New Testament studies with ample excerpts from the major scholars since the 18th century. For a rich introduction to the 27 books of the New Testament with full and balanced reporting on all major issues of contemporary discussion and extensive bibliographies, see PAUL FEINE, JOHANNES BEHM, and WERNER G. KUEMMEL, *Einleitung in das Neue Testament*, 14th rev. ed. (1965; Eng. trans., *Introduction to the New Testament*, 1966). For a general dictionary to the Bible, see GEORGE A. BUTTRICK (ed.), *The Interpreter's Dictionary of the Bible*, 4 vol. (1962). The most extensive tool for the study of New Testament theological terms is GERHARD KITTEL (ed.), *Theological Dictionary of the New Testament*, vol. 1–8 (1964–72, in progress). For New Testament theologies, see RUDOLF BULTMANN, *Theologie des Neuen Testaments*, 3rd ed. (1958; Eng. trans., *Theology of the New Testament*, 2 vol., 1951–55); HANS CONZELMANN, *Grundriss der Theologie des Neuen Testa-*

ments, 2nd ed. (1967; Eng. trans., *An Outline of the Theology of the New Testament*, 1969). For general commentary, see *The International Critical Commentary on the Holy Scriptures of the Old and New Testaments*, 41 vol. (1895–1920); for the major German commentary, see *Kritisch exegetischer Kommentar über das Neuen Testament* ("Meyer Series," frequently updated); *Handbuch zum Neuen Testament* (Lietzmann-Bornkamm); and *Das Neue Testament Deutsch* (*Göttinger Bibelwerk*). For a major French Protestant commentary, see *Commentaire du Nouveau Testament* and for major French and German Roman Catholic commentaries, see *Études bibliques* and *Das Neue Testament übersetzt und erklärt* (the *Regensburger New Testament*). (*Gospels—texts*): KURT ALAND (ed.), *Synopsis Quattuor Evangeliorum* (1964), a Greek synopsis, includes the Gospel of John and an English translation of the Coptic *Gospel of Thomas*. For a synopsis, see BURTON H. THROCKMORTON, JR. (ed.), *Gospel Parallels: A Synopsis of the First Three Gospels*, 3rd ed. (1967). For a general study of the Gospels and the Synoptic problem, see FREDERICK C. GRANT, *The Gospels: Their Origin and Their Growth* (1957). For arguments against the priority of Mark, see WILLIAM R. FARMER, *The Synoptic Problem* (1964). Significant new approaches to gospel study are found in JAMES M. ROBINSON and HELMUT KOESTER, *Trajectories Through Early Christianity* (1971). For form criticism, see RUDOLF BULTMANN, *Die Geschichte der synoptischen Tradition*, 3rd ed. (1958; Eng. trans., *The History of the Synoptic Tradition*, 1963). AMOS N. WILDER, *Early Christian Rhetoric* (1971), goes beyond form-criticism by fuller attention to modern literary criticism. (*Mark*): ROBERT H. LIGHTFOOT, *The Gospel Message of St. Mark* (1950); and WILLI MARXSEN, *Der Evangelist Markus* (1959; Eng. trans., *Mark the Evangelist*, 1969), are two outstanding works representing different periods and methods of scholarship. (*Matthew*): For discussion of the arrangement, Old Testament citations, and theology of Matthew, see GUENTHER BORNKAMM, GERHARD BARTH, and HEINZ J. HELD, *Auslegung im Mattäus-evangelium* (1960; Eng. trans., *Tradition and Interpretation in Matthew*, 1963); KRISTER STENDAHL, "Prayer and Forgiveness," in *Svensk Exegetisk Årsbok*, 22–23:75–86 (1957–58), in English; and *The School of St. Matthew and Its Use of the Old Testament*, 2nd ed. (1968). (*Luke*): HENRY J. CADBURY, *The Making of Luke-Acts*, 2nd ed. (1958); and HANS CONZELMANN, *Die Mitte der Zeit: Studien zur Theologie des Lukas*, 3rd ed. (1960; Eng. trans., *The Theology of St. Luke*, 1960), represent a classic treatment of Luke-Acts. (*John*): Among the most important recent studies on John are CHARLES H. DODD, *Historical Tradition in the Fourth Gospel* (1963) and *The Interpretation of the Fourth Gospel* (1953); ERNST KAESEMANN, *Jesu letzter Wille nach Johannes 17.* (1966; Eng. trans., *The Testament of Jesus: A Study of the Gospel of John in the Light of Chapter 17*, 1968); and JAMES L. MARTYN, *History and Theology in the Fourth Gospel* (1968). (*Acts*): See also *Luke* above. For Acts viewed in its own time, see HENRY J. CADBURY, *The Book of Acts in History* (1955). Literary style and methods of composition are discussed in MARTIN DIBELIUS, *Aufsätze zur Apostelgeschichte* (1951; Eng. trans., *Studies in the Acts of the Apostles*, 1956). The scope and purpose of Acts are treated in P.M. MENOUD, "Le Plan des Actes des Apôtres," *New Testament Studies*, 1:44–50 (1954–55); and W.C. VAN UNNIK, "The 'Book of Acts' the Confirmation of the Gospel," *Novum Testamentum*, 4:26–59 (1960). (*Paul*): For general works on Paul and the epistles, see GUENTHER BORNKAMM, *Early Christian Experience* (1970), and *Paulus* (1969; Eng. trans., 1971); WILLIAM D. DAVIES, *Paul and Rabbinic Judaism*, 2nd ed. (1955); MARTIN DIBELIUS and WERNER G. KUEMMEL, *Paulus* (1951; Eng. trans., 1953); JOHANNES MUNCK, *Paulus und die Heilsgeschichte* (1954; Eng. trans., *Paul and the Salvation of Mankind*, 1959); ARTHUR D. NOCK, *St. Paul* (1938); and HANS J. SCHOEPS, *Paulus: Die Theologie des Apostels . . .* (1959; Eng. trans., *Paul: The Theology of the Apostle . . .*, 1961). See also KRISTER STENDAHL, "The Apostle Paul and the Introspective Conscience of the West," *Harvard Theological Review*, 51: 199–215 (1963). For a survey of Pauline studies, see EDWARD E. ELLIS, *Paul and His Recent Interpreters* (1961); WAYNE A. MEEKS (ed.), *The Writings of St. Paul* (1972); and ERNST KAESEMANN, *Paulinische Perspektiven* (1969; Eng. trans., *Perspectives on Paul*, 1971). (*Romans*): JOHN KNOX, "A Note on the Text of Romans," *New Testament Studies*, 2: 191–192 (1955–56); KRISTER STENDAHL, "Hate, Non-Retaliation, and Love: 1QS x, 17–20 and Romans 12:19–21," *Harvard Theological Review*, 50:343–355 (1962). (*I Corinthians*): For a discussion of the heresies met in I Corinthians, see WALTER SCHMITHALS, *Die Gnosis in Korinth*, 3rd ed. (1969; Eng. trans., *Gnosticism in Corinth*, 1971). (*II Corinthians*): For the arrangement of the fragments of II Corinthians and their redaction, see GUENTHER BORNKAMM, "The History of the Origin of the So-Called 2nd Letter to the Corinthians," *New Testament Studies*, 8:258–264 (1961–62). For a discussion of Paul's opponents in II Corinthians, see DIETER GEORGI, *Die Gegner des Paulus im 2. Korintherbrief: Studien zur religiösen Propaganda in der Spätantike* (1964); and his shorter article on this subject, "Forms of Religious Propaganda," in HANS J. SCHULTZ (ed.), *Die Zeit Jesu* (1966; Eng. trans., *Jesus in His Time*, 1971). (*Galatians*): For a discussion of the heretics in Galatia, see WALTER SCHMITHALS, "Die Häretiker in Galatien," *Zeitschrift für die Neutestamentliche Wissenschaft und die Kunde der Aelteren Kirche* (ZNW), pp. 25–67 (1956). (*Ephesians*): For the meaning and goal of Ephesians, see EDGAR J. GOODSPEED, *The Meaning of Ephesians* (1933) and *The Key to Ephesians* (1956). See also C. LESLIE MITTON, *The Epistle to the Ephesians* (1951). (*Philippians*): For the place of Philippians in the Pauline collection and the meaning of its various sections, see HELMUT KOESTER, "The Purpose of the Polemic of a Pauline Fragment (Philippians III)," *New Testament Studies*, 8:317–332 (1961–62). For the concept of Philippians as a testament, see DIETER GEORGI, "Ein Testament des Paulus (Phil. 3, 2ff.)," *ZNW* (1972). (*Philemon*): JOHN KNOX, *Philemon Among the Letters of Paul*, 2nd ed. (1959). (*Pastoral Epistles*): For evidence against Pauline authorship, see PERCY N. HARRISON, *The Problem of the Pastoral Epistles* (1921). See also EDUARD SCHWEIZER, *Church Order in the New Testament* (1961). (*Hebrews*): Concerning the Christology of Hebrews and the idea of the "wandering people of God," see ERNST KAESEMANN, *Das wandernde Gottesvolk*, 3rd ed. (1959). An approach to the eschatology of Hebrews and an origin connected with followers of Stephen is found in WILLIAM MANSON, *The Epistle to the Hebrews* (1951). (*Catholic Epistles*): For the typical admixture of parenesis, apocalyptic, and the general address of the Catholic Epistles, see CARL ANDRESEN, "Zum Formular frühchristlicher Gemeindebriefe," *ZNW*, 56:233–259 (1965). The similarity of style of the Catholic Epistles to later Christian Greek literature is treated in A. WIFSTRAND, "Stylistic Problems in the Epistles of James and Peter," *Studia Theologica*, 11:35–60 (1948). (*James*): For a solution to the apparent contradiction of Paul and James concerning "works," see JOACHIM JEREMIAS, "Paul and James," *Expository Times*, 66:368–371 (1954–55); for clarification of special passages with a modern technique similar to rabbinic methodology, see ROY B. WARD, "The Works of Abraham: James 2:14–26," *Harvard Theological Review*, 61:283–290 (1968) and "Partiality in the Assembly: James 2:2–4," *ibid.*, 62:87–97 (1969). (*I Peter*): For a date in Trajan's time, see JOHN KNOX, "Pliny and I Peter: A Note on I Pet. 4:14–16 and 3:15," *Journal of Biblical Literature*, 72:187–189 (1953); an interpretation of the Descent into Hell is found in B. REICKE, *The Disobedient Spirits and Christian Baptism: A Study of I Peter iii, 19 and Its Context* (1946). (*II Peter and Jude*): For motivation for the writing of II Peter, see ERNST KAESEMANN, "An Apologia for Primitive Christian Eschatology," in *Essays on New Testament Themes* (1964). (*Johannine Epistles*): For speculations as to authorship, date, and nature of the situation of the Johannine Epistles, see W.F. HOWARD, "The Common Authorship of the Johannine Gospel and Epistles," *Journal of Theological Studies* (1947). (*Revelation*): Concerning liturgical style and content in Revelations, see GUENTHER BORNKAMM, "On the Understanding of Worship; B," in *Early Christian Experience* (1969). For a study of Revelation as a creative revelatory poem with unity throughout, drawing upon apocalyptic imagery of its time, see AUSTIN M. FARRER, *A Rebirth of Images* (1949, reprinted 1963). A general survey of apocalypticism and apocalypses from 200 BC into the early Christian era is found in DAVID S. RUSSELL (*op. cit.*).

Apocrypha: EDGAR HENNECKE, *Neutestamentliche Apokryphen in deutscher Übersetzung* (1959; Eng. trans., *New Testament Apocrypha*, 2 vol., 1963–65), a standard work; MONTAGUE R. JAMES, *The Apocryphal New Testament* (1924, reprinted 1955), convenient but obsolete.

Biblical literature in liturgy: ABRAHAM Z. IDELSOHN, *The Jewish Liturgy and Its Development* (1932, reprinted 1967); JOSEPH H. HERTZ, *The Authorized Daily Prayer Book*, rev. ed. (1948), Hebrew and English with historical notes and commentary; FAN S. NOLI, *Three Liturgies of the Eastern Orthodox Church* (1955); DONALD ATTWATER, *Eastern Catholic Worship* (1945), eight Uniate liturgies in English; JOSEF A. JUNGMANN, *Missarum Sollemnia: Eine Genetische Erklärung der römischen Messe*, 2 vol. (1958; Eng. trans., *The Mass of The Roman Rite: Its Origins and Development*, abridged ed., 1959); CLEMENT J. MCNASPY, *Our Changing Liturgy* (1966), reforms following Vatican II; GREGORY DIX, *The Shape of the Liturgy* (1945); BARD THOMPSON, *Liturgies of the Western Church* (1961), includes the main Protestant traditions.

(J.C.Ry./L.F./R.F./N.M.Sa./
H.G.D./D.Fl./K.St./E.T.Sa./R.M.G.)

Bibliography

Bibliography, the art or science of the description of books, has acquired special importance in the 20th century because of the need for effective organization of the records of human communication in the face of the enormous growth of publishing activity and the need, especially in undeveloped countries, for informed access to the world's scientific and technical information. It has been said that without bibliography, the records of civilization would be an uncharted chaos of miscellaneous contributions to knowledge, unorganized and inapplicable to human needs.

The word bibliography, in its literal sense, derived from the Greek *bibliographia* (2nd century AD), means the writing of books, and it was so defined in the 17th century; since the 18th century, it has been used to denote the systematic description and history of books. It is now commonly used in two widely divergent, though basically connected senses: (1) the listing of books, arranged according to some system (in this sense it is called enumerative, systematic, or descriptive bibliography); and (2) the study of books as material objects; *i.e.*, the study of the material of which books are made and the manner in which they are put together (in this sense commonly called critical bibliography). It is the function of bibliography to provide useful information for the student, in the one case supplying him with information about material for study, in the other helping him to establish the place of a book (or a piece of writing) in an author's production and its quality and authenticity as a text for study.

Kinds of bibliographies

DESCRIPTIVE BIBLIOGRAPHY

The tasks of the compiler of a bibliography are (1) to find out what books on a particular subject exist; (2) to describe them item by item; and (3) to assemble the resulting entries into useful arrangements for reference and study. The need for lists of this kind arises as soon as the number of books in any subject is too great to be easily remembered. Among the earliest lists of books are those compiled by certain writers as guides to their own books—*e.g.*, the celebrated physician Galen compiled, in the 2nd century AD, a description in subject order of his own writings (*De propriis libris liber*, "A Book about My Own Books"); the Venerable Bede attached at the end of his *Ecclesiastical History of the English People*, in 731, an autobiographical note with a list of his writings; Erasmus also published a catalog of his own writings in narrative form in 1523. Early bibliographies, however, were not confined to such autobibliographies; as early as the last decade of the 4th century AD, the idea of attaching lists of works to lives of ecclesiastical writers was adopted by St. Jerome in his *De viris illustribus* ("Concerning Famous Men").

The invention of printing in the 1440s made possible the rapid multiplication of books, thereby increasing the need for guides to the resultant literature and also bringing about changes in the manner of compilation. The earliest substantial bibliography after the invention of printing was that of Johannes Tritheim, abbot of Sponheim in the diocese of Mainz, who in *Liber de scriptoribus ecclesiasticis* ("Book about Ecclesiastical Writers," 1494) included in chronological order, with an alphabetical index, about 1,000 ecclesiastical writers and their books. In 1545 a German-Swiss writer and naturalist Conrad Gesner, who has been known as the father of bibliography, published his *Universal Bibliography* (*Bibliotheca Universalis*) of all Latin, Greek, and Hebrew writers, living and dead; this was followed three years later by a second volume, *Pandectarum sive Partitionum universalium libri XXI* ("Twenty-one Books of Encyclopedias or Universal Divisions [of Knowledge]"), in which the entries, arranged alphabetically in the earlier volume, are rearranged under 21 subject headings. Although Gesner was not the earliest descriptive bibliographer, his attempts at universality and classification earn him his fame. Gesner's idea of universality remained an ideal until recent times; in the face of the formidable problems of size, cost, and complexity, however, it has receded into the background of human aspirations. The Institut International de Bibliographie, founded in 1895 in Brussels by Henry Lafontaine and Paul Otlet with the object of creating a universal bibliography of books and articles in periodicals, arranged according to a specially designed system of classification, the Universal Decimal Classification (UDC), has assembled a card catalog of many millions of entries; but because of the immense bulk of the material and the growing cost of compilation, it is unlikely that it will ever reach the goal set by its founders. The hopes for universal bibliographies have been largely replaced by the published catalogs of such great comprehensive libraries as the British Museum (1961–67), the Bibliothèque Nationale (beginning in 1897), the Library of Congress, and, most promising of all, the United States National Union Catalogue, maintained in the Library of Congress and, in the early 1970's, in the course of being printed in an estimated 625 volumes. Investigations were (early 1970's) being made into the problems involved in programming and computerizing such catalogs, and if the complexities of programming multilanguage material can be solved, if the necessary financing can be found, and if world cooperation in the production and standardization of catalog entries can be assured, it seems possible that the ideals of the Brussels Institute may yet to some extent be realized.

Hopes for a universal bibliography

Gesner's achievement was remarkable; besides writing many other books, including a number of bibliographical works, his universal bibliography with its appendix (1555) describes and classifies some 15,000 works by about 3,000 writers and provides alphabetical and subject indexes. His contemporaries turned their attention to detailed bibliographies of writings on particular subjects or in particular languages or relating to particular countries or places. Thus, John Bale, bishop of Ossory, published the earliest national bibliography, *Illustrium maioris Britanniae scriptorum . . . summarium* ("Summary of the Writings of the Most Eminent Britons," 1548), which lists British writers chronologically and sets out their writings in detail. The vastly improved extended second edition of this work, *Scriptorum illustrium maioris Brytanniae . . . catalogus* (1557), acknowledges Bale's indebtedness to John Leland, Henry VIII's library keeper and antiquary, who had made an antiquarian tour through England in 1534–42 and had made a survey of the writings of British authors. Other national bibliographies followed, such as those of the Italian Antonio Francesco Doni, *La Libraria del Doni Fiorentino Nella quale sono scritti tutti gl'autori vulgari* ("Library of . . . All Secular Authors," 1550; 2nd ed. also 1550); *La seconda del Doni* (1551); the Dutch theologian Cornelius Loos, *Illustrium Germaniae scriptorum catalogus* (1582); the French bibliographer François Grudé de la Croix du Maine, *Premier Volume de la bibliothèque du Sieur de la Croix du Maine* ("The First Volume of the Library of . . .," 1584). The catalogs of the German book fairs held in Frankfurt and Leipzig each spring and autumn, beginning in 1564 and continuing until 1749, while not bibliographies in the strict sense, were nevertheless widely used as foundation material by early German bibliographers.

Catalogs of the German book fairs

Types of descriptive bibliographies. *National bibliographies.* Most countries now have national bibliographies, in a majority of cases published officially by the national library and based on copies of national publications deposited in accordance with provisions of copyright acts. Notable exceptions are the United States and The Netherlands, whose national bibliographies are published commercially; East Germany, whose *Deutsche Nationalbibliographie* is published by the Deutsche Bücherei at Leipzig, an organization maintained since 1913 by German publishers; and the United Kingdom, whose *British National Bibliography* is published by a council representing libraries, publishers, and booksellers. These national bibliographies aim at, and attain, a high degree of completeness and promptitude. The *British National Bibliography* (beginning in 1950), for example, is published

weekly and cumulated quarterly and annually; it is arranged in a classified order according to the decimal classification of the U.S. librarian Melvil Dewey, with an alphabetical index of authors, titles, and subjects. The *Bibliographie de la France* (beginning in 1811), published weekly, is arranged in a classified order with an annual index of authors, titles, and subjects. The East German *Deutsche Nationalbibliographie* (beginning in 1931) is published in two parts: (1) commercial publications, published weekly and (2) publications of societies and institutions, published semi-monthly. The West German *Deutsche Bibliographie* (beginning in 1947) is published weekly, with author and catchword index; this work has been published by computer since the beginning of 1968 —the first such work to be so produced.

Bibliographies of books published in particular countries are of great value to students. Outstanding examples of such bibliographies are Charles Evans, *American Bibliography* (1903–34), covering the period 1639–1799; *A Short-Title Catalogue of Books Printed in England, Scotland and Ireland, and of English Books Printed Abroad, 1475–1640*, compiled by A.W. Pollard and G.R. Redgrave (1926; new edition in preparation), and its continuation by D.G. Wing, *Short-Title Catalogue . . . 1641–1700* (1945–51; new edition in preparation). While not strictly bibliographies, the British Museum catalogs of its holdings of pre-1600 books from several European countries are similarly valuable because of the richness of the museum's collections.

Personal bibliographies. Personal bibliographies may consist of no more than a simple list of an author's works; *e.g.*, those attached to articles in the *Dictionary of National Biography*. But they may be more elaborate—*e.g.*, F.W. Ebisch and L.L. Schücking, *A Shakespeare Bibliography* (1931; supplement 1937); Michael Sadleir, *Trollope: A Bibliography* (1928); Bertha Coolidge Slade, *Maria Edgeworth* (1937). Personal bibliographies are sometimes based on private collections—*e.g.*, *A Stevenson Library* (1951–64) based on the collection of Edwin J. Beinecke; T.J. Wise's bibliographies, based on his own collections, of Tennyson, Swinburne, and others. A variant of personal bibliography consists of a narrative setting out an author's works with a comprehensive account of the circumstances surrounding the composition and publication of each work. An example is J.E. Norton, *A Bibliography of the Works of Edward Gibbon* (1940).

Subject bibliographies. Subject bibliographies vary in size, scope, and method, according to the purpose they are designed to serve. The following are examples of bibliographies that aim at offering comprehensive guidance: A. and A. de Backer and A. Carayon, *Bibliothèque de la Compagnie de Jésus*, ("Library of the Society of Jesus," 9 vol.; new ed. 1890–1909); F.W. Bateson, *The Cambridge Bibliography of English Literature* (4 vol., 1940; new ed. 1970 ff.); *Bibliography of British History:* C. Read, *Tudor Period, 1485–1603* (1933), G. Davies, *Stuart Period, 1603–1714* (1928), S. Pargellis and D.J. Medley, *The Eighteenth Century, 1714–1789* (1951); P. Caron, *Manuel pratique pour l'étude de la Révolution française* ("Practical Manual for the Study of the French Revolution"; latest ed. 1947); F.C. Dahlmann and G. Waitz, *Quellenkunde der deutschen Geschichte,* ("Published Sources of German History"; 9th ed. 1931–32); W.W. Greg, *A Bibliography of the English Printed Drama to the Restoration* (4 vol., 1939–59); G. Lanson, *Manuel bibliographique de la littérature française moderne* ("Bibliographic Manual of Modern French Literature"; latest ed. 1947); F. Madan, *The Early Oxford Press* (3 vol., 1895–1931); L.-N. Malclès, *Les Sources du travail bibliographique* ("Sources of Bibliographic Work," 3 vol., 1950–52). Bibliographies of publications on particular subjects, frequently including articles on periodicals, are numerous and cover many branches of human knowledge and interests.

Bibliographies of bibliographies

Bibliographical guides. The proliferation of bibliographies has led to the publication of bibliographical guides—bibliographies of bibliographies. Among these should be mentioned *Bibliotheca Bibliographica* (1866), compiled by Julius Petzholdt, librarian of the Royal Library at Dresden, Germany and *World Bibliography of Bibliographies* (5 vol.; 1965–66) by Theodore Besterman. Guides to current publications in a number of subject fields are published annually sometimes in narrative form—*e.g.*, *The Year's Work in English-Studies* (1921 ff.) and *The Year's Work in Librarianship* (1929 ff.; later title *Five Years' Work in Librarianship*, 1958 ff.)—and sometimes in list form—*e.g.*, *Internationale Bibliographie des Buch- und Bibliothekswesens* ("International Bibliography of Book and Library Systems," 1926–41).

Methods of compilation. *Research.* The method of compiling a bibliography and the amount of detail included vary according to the author's purpose, his view of the importance of the subject, and his knowledge of it. The task of compiling a bibliography of any subject is a matter for a specialist in that subject. Having determined his objective and what material exists, the author must proceed to describe and arrange it as usefully for his purpose as he can. To ensure that the descriptions are accurate and consistent, they should be made from the works themselves. If, for any reason, this is impossible, the source of the information given should be stated.

For most bibliographies it is usually necessary to give only the author, short title, place of imprint, and date; but these should be transcribed accurately from the book or article. Notes may be added on the scope and quality of the text and the location of a copy. In an author bibliography a chronological arrangement is often adopted, enabling the user to follow the development of the author's work. In such cases, later editions are listed with the earliest edition but additional mention is made at the appropriate chronological point. Subject bibliographies may be arranged in a systematic order to bring out the features important for the author's purpose, or a recognized classification system, such as the Dewey Decimal Classification or Universal Decimal Classification, may be used.

Full descriptions are often given for early or rare books and for detailed author bibliographies. Elaborate rules have been evolved for compiling such descriptions, which make it possible for a skilled bibliographer to reconstruct from the text before him the makeup and appearance of a book. Such descriptions include information about the details of publication, number of copies printed, price, and binding. Semi-facsimile transcriptions are sometimes given, illustrating the various types of font used and the spacing of the title page. Modern methods of reproduction, however, make it possible to reproduce title pages photographically and thus obviate the use of such expedients, which in any case are not entirely satisfactory. The use of reproductions, however, has certain disadvantages, such as high cost and difficulty of layout, and, more important, reproductions themselves may be misleading, especially when the quality of the printing and paper of the original are poor. In 17th-century printing, for example, it is sometimes difficult to distinguish a mark of punctuation from a flaw in the paper, and a reproduction may give a wrong impression. If facsimiles are used, they should be as carefully edited as the text and the source of the reproduction should be stated.

Use of reproductions

Use of computers. The advent of computers and the application of data-processing techniques to library procedures such as cataloging have great potential value in bibliography. If a catalog is maintained in machine-readable form, it can be printed out in a number of different arrangements, or a variety of sublistings can be produced, with a relatively small increase in costs over the cost of printing the original arrangement. This method depends on the planning of requirements in advance and on the work of the programmer. It is thus possible for specialized lists, taken from the contents of a particular library, to be made available to a wide range of users. The best known and most highly sophisticated retrieval project of this kind is the Medlars project (Medical Literature Analysis and Retrieval System), which has as its main objective the publication of the *Index Medicus* (a monthly listing of current articles from about 2,300 biomedical journals throughout the world) but which also

supplies printed listings on particular subjects in the field of medicine.

The full benefit of these methods will accrue when it is economically and administratively possible to contemplate the storage of the enormous amounts of information available in the large comprehensive libraries.

In 1968 the Library of Congress began the production and distribution of cataloging data in machine-readable form. By 1969 the catalog record already consisted of 80,000 entries, and the annual increment was estimated at more than 25,000. The scope of this enterprise was limited initially to monographic works in the English language, but with the intention of expanding the coverage to include other languages. The creation of the machine record of material already in the library before 1969, which is necessary if a full system of machine-readable records is to be created, was being investigated both in the Library of Congress and elsewhere in other large comprehensive libraries. The problems likely to cause most difficulty are the labour involved, the great cost of the operation, and the magnitude of the data storage. Clearly such costly and complex operations have to be related to a realistic assessment of the benefits to be derived therefrom. Such assessments are in keeping with an interest, current in the early 1970s, in research library management. The rapid rise in library costs has emphasized the need for libraries, particularly large research libraries, to study whether and how they can apply accepted management techniques to the fuller exploitation of the bibliographical riches of their own institutions.

CRITICAL BIBLIOGRAPHY

New approach to English literature

A valuable approach to the study of English literature, particularly, though not exclusively, of the 17th and 18th centuries, was developed about the turn of the 20th century out of a concern with Shakespearean texts in general and in particular with the nature and authority of the so-called quarto texts and their relation to the text printed in the First Folio. Editors had previously tended to assume that all the early texts of Shakespeare had equal authority and that it was perfectly proper to select readings which, in their view, improved the text of the author. The earlier editors were handicapped, in addition, by their ignorance of printing and book production in Shakespeare's time and by a lack of understanding of the nature of the texts they were dealing with. The change that took place stemmed from a variety of causes, of which the most important were probably two: (1) the close association in studies of English literature, especially drama, of three eminent scholars, A.W. Pollard, R.B. McKerrow, and W.W. Greg, all of them interested in bibliography; and (2) the remarkable progress made by Robert Proctor, a colleague of Pollard's in the British Museum, in the study of incunabula, the name given to books printed in the first 50 years after the invention of printing (*i.e.*, before 1501). The earliest printed books display the very individual characteristics of their printers in the type used, the typesetting, and the layout of the page, and it was found that, by studying these characteristics, useful deductions could be made about the place of individual books in a printer's total production and hence their approximate dates; in cases in which books contained no indication of the printer, it was often possible, using the same means, to assign a book to a particular printer. Important stages in the study of these earliest printed books, which had always been prized for their early date, their great aesthetic quality, and their rarity, were marked by a study of William Caxton, England's earliest printer, by William Blades, *The Life and Typography of William Caxton* (1861–63), in which typographical details were used to arrange Caxton's publications in chronological order; and by the work of a great antiquarian scholar and librarian, Henry Bradshaw, who made a special study of 15th-century books printed in The Netherlands. Bradshaw laid down an important principle: let the book speak for itself. This new method was still more widely applied by Robert Proctor, who succeeded in allotting the large collections of incunabula in the British Museum and the Bodleian Library at Oxford to places of origin, by countries and towns, and to printers. Proctor's work firmly established this method of study, and the definitive catalog of the incunabula in the British Museum, which he began, demonstrated the valuable results that can be obtained by the meticulous examination of all the features of a book—paper, type, makeup, ornamentation, sewing, binding, manuscript notes, and marks of ownership.

Examination of the physical features of a book

It was against this background and, applying these methods to the literary and dramatic work of English authors of the 16th and 17th centuries, that Greg, McKerrow, and Pollard developed what is now called critical, or analytical, bibliography, which has been succinctly described as the study of books as tangible objects. By using the evidence of the physical features of books, it was often found possible to arrive at conclusions about the place of a book in an author's production that are beyond the scope of literary judgment.

One of the earliest and most illuminating examples of the application of this method concerns the question of the priority between the two 1609 issues of Shakespeare's play *Troilus and Cressida*. These differ in the preliminary matter: the title of one describes the play as having been acted at the Globe Theatre; in the other, the title page makes no mention of a production of the play, and an epistle to the reader stresses this fact by stating that the piece had never been "clapper-clawed with the palmes of the vulger." Literary considerations would suggest that the latter issue was the earlier. An examination of the books themselves, however, shows that the title page of the former is in its original state, while the latter's original title page had been cut away and two leaves substituted, containing the new title page and the address to the reader, thus demonstrating, without question, the order of the two.

In his edition of Beaumont and Fletcher's *Elder Brother* (1905), Greg similarly provided an irrefutable demonstration of the correct chronological order of the two quarto editions of 1637. In Q1 an improperly adjusted quad or space lead had produced a mark above the line before the word "young" in Act I, scene 2, line 72, which was mistaken by the compositor printing Q2 for an apostrophe and printed as such. Hence, Q2 was shown to have been printed from Q1 and not from the author's manuscript.

The success of such demonstrations led to a much more detailed study of the physical features of books and the bibliographical deductions that can be made from them. The methods of critical bibliography have been substantially developed and have been widely applied to books of later periods, and it has been shown that, even in the sophisticated machine age, careful bibliographical examination has a significant role to play in determining the reliability of an author's text and the place of a book in an author's production.

Exposure of false first editions

The method was spectacularly displayed in the exposure of a number of alleged first editions of poems, essays, and other minor productions of well-known 19th-century authors, such as the "Reading" sonnets of Elizabeth Barrett Browning—*i.e.*, an edition of these sonnets, dated 1847, with Reading as the place of imprint. These books, more than 50 in number, were shown, in *An Enquiry into the Nature of Certain Nineteenth Century Pamphlets* (1934), by John Carter and Graham Pollard, to have been printed on kinds of paper, made from esparto grass or wood pulp, known not to have been in use at the dates shown on the title pages, and using printing types that were first cut much later.

It has been clearly demonstrated that machine-printed books of the 19th and 20th centuries produce textual problems rivalling those found in hand-printed books. While editorial procedures and the principles governing the interpretation of bibliographical evidence remain the same, new printing methods force bibliographers to devise new procedures to deal with the new problems presented to them.

A widely used guide for the bibliographical study of English books up to the end of the 17th century is provided by R.B. McKerrow's *Introduction to Bibliography*

for Literary Students (1927). The publications of the Bibliographical Society of London, especially its journal, *The Library* (1889 ff.), contain valuable information about bibliographical methods and their application. A useful account of the work of the society and a study by F.P. Wilson of the application of bibliographical techniques to the works of William Shakespeare, "Shakespeare and the New Bibliography," is contained in *The Bibliographical Society, 1892–1942: Studies in Retrospect,* (1945). Other bibliographical journals that may be usefully consulted are: *The Papers of the Bibliographical Society of America* (1904 ff.), the *Publications* of the *Oxford Bibliographical Society* (1922 ff.), and the *Transactions of the Cambridge Bibliographical Society* (1949 ff.). *Studies in Bibliography* (1948 ff.), published by the Bibliographical Society of the University of Virginia, is especially valuable.

BIBLIOGRAPHY

The works here noted are additional to those given in the text. The general works are for the most part designed for students or beginners in the art of descriptive bibliography. Those on critical bibliography are of a much more advanced nature and demonstrate the depth of the relationship between bibliography and textual criticism.

General works: THEODORE BESTERMAN, *The Beginnings of Systematic Bibliography*, 2nd ed. rev. (1936); A. ESDAILE, *A Student's Manual of Bibliography*, 3rd rev. ed. (1954); L.N. MALCLES, *Les Sources du travail bibliographique*, 3 vol. (1950–58); GEORG SCHNEIDER, *Handbuch der Bibliographie*, 4th ed. (1930, reprinted 1969); J.H. SHERA and M.E. EGAN (eds.), *Bibliographic Organization* (1951); RONALD STAVELEY, I.C. MCILWAINE, and JOHN H. ST. MCILWAINE, *Introduction to Subject Study* (1967); R. STOKES, *The Function of Bibliography* (1969); CURT F. BUHLER *et al.*, *Standards of Bibliographical Description* (1949); J.D. COWLEY, *Bibliographical Description and Cataloguing* (1939).

Critical bibliography: R.B. MCKERROW, *An Introduction to Bibliography for Literary Students* (1927); F.T. BOWERS, *Bibliography and Textual Criticism* (1964) and *Principles of Bibliographical Description* (1949); O.M. BRACK, JR. and W. BARNES (eds.), *Bibliography and Textual Criticism: English and American Literature, 1700 to the Present* (1969).

(F.C.F.)

Bicycle

The bicycle, a light, two-wheeled, steerable machine propelled by human power, is said to be the most efficient means yet devised to convert human energy into propulsion. It is inexpensive and easily mass-produced. First made in the early 19th century, it developed as an important means of transportation and as the basis for a worldwide sport and industry. Bicycles are the most numerous road machines in many countries. Cycle touring is most widely practiced in England but has a large and increasing number of followers in Germany, France, and other European countries. Bicycles are also widely used for racing in France, Germany, Belgium, and to a lesser extent in England and the United States (see CYCLING). A bicycle rider can quite easily maintain a pace of 10–12 miles (16–19 kilometres) per hour; *i.e.*, about four times a walking pace. Cycle touring is popular in many countries and is often organized in large clubs.

EARLY BICYCLES

The first known patent of a machine that could have resembled a bicycle was held by Jean Theson, who at Fontainebleau on Feb. 4, 1645, was given a 30-year privilege to "put into use a small body on four wheels driven without horses but by two seated men." A similar four-wheeled machine was seen in Paris in 1779, according to the *Journal de Paris.* Built by François Blanchard and M. Masurier, it was evidently large and ponderous.

The first two-wheeled machine of which there is evidence was the *draisienne*, invented by Baron Karl de Drais de Sauerbrun and exhibited in Paris, April 6, 1818. It was made of wood, and the seated rider propelled himself simply by paddling his feet against the ground. Steerable, crude, and clumsy, it worked after a fashion. Copies of the machine were made in Britain by Denis Johnson and in America, where it enjoyed a brief vogue. Not until Kirkpatrick Macmillan, a blacksmith of Dumfriesshire, Scotland, completed four years of experiments in 1839 did a self-propelled bicycle appear.

The *draisienne*

Macmillan's machine had wheels rimmed with iron and, though lighter in appearance than the *draisienne*, it was still heavy (Figure 1, left). With a steerable front wheel about 30 inches (75 centimetres) in diameter and a driven rear wheel of about 40 inches (100 centimetres) in diameter, it could move at a brisk pace. In 1842 Macmillan successfully challenged a post carriage.

Instead of pedals attached directly to the rear wheel, Macmillan's vehicle had two swinging cranks mounted at the front. The rider rested his feet on the cranks and swung them back and forth, moving a pair of rods that were linked to two levers, located on either side of the rear wheel. Pressing down on one crank pulled the rod forward, which in turn drew the uppermost of the levers forward in an arc, thus turning the wheel and bringing the opposing lever uppermost for the cycle to be repeated with the other foot.

The machine never became popular and, though copied, passed out of fashion. Thus while Macmillan may fairly be claimed as the inventor of the bicycle, the first usable mechanism that survived in principle was the work of two Frenchmen, Pierre Michaux and his son Ernest. In Paris in 1861 the Michaux family built a machine on which they attached two cranks to the front wheel. The cranks could be rotated by the rider's feet, an arrangement that, according to Henry Michaux in 1893, was an adaptation of the crank handles of a vertical grindstone the inventors had seen. Their machine immediately caught on, though its wood and iron frame gave it the sobriquet of "Boneshaker." That year they made only

Science Museum, London, Crown copyright

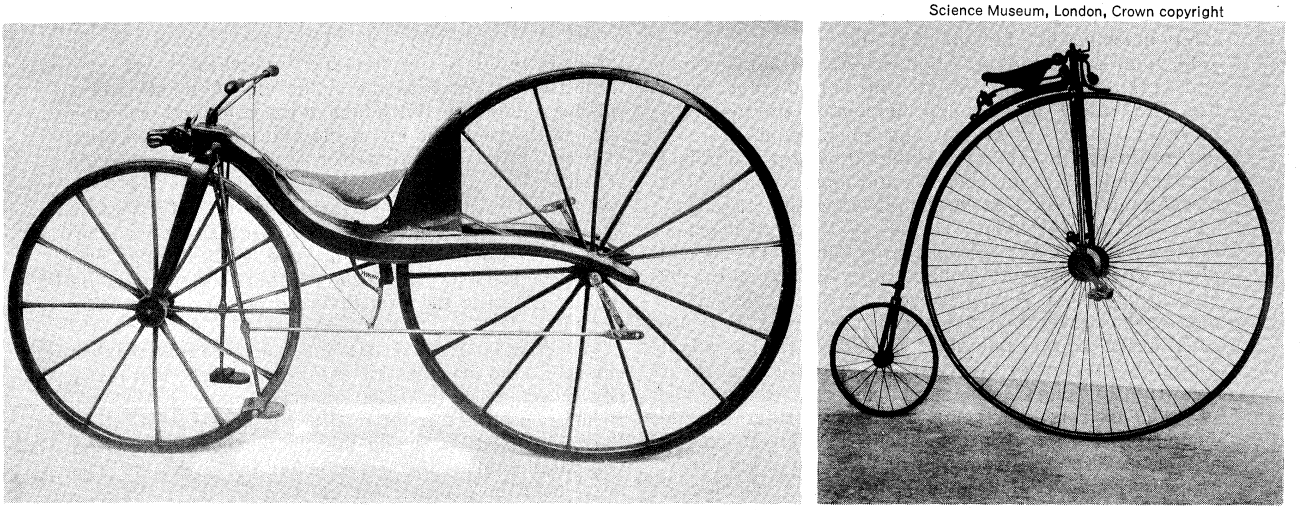


Figure 1: (Left) Macmillan's Hobbyhorse, *c.* 1839. (Right) Ordinary bicycle, 1883.

two machines. In 1862, when it had already been copied in Munich (the machine is preserved in the Deutsches Museum there), they made 142 of the *vélocipèdes*, as they had come to be known. By 1865 the Michaux family was making 400 a year. In 1866 their mechanic, Pierre Lallement, emigrated to the U.S., where with James Carroll of Ansonia, Connecticut, he took out the first U.S. patent.

First bicycle race

On May 31, 1868, the first recorded bicycle race was won in the Parc de Saint-Cloud by James Moore, an Englishman, who also won the first road race (Paris to Rouen) in November, 1869. Reportedly, he rode a 160-pound machine that had solid rubber tires and ball bearings, and he covered the 83 miles in 10 hours 25 minutes, some 45 minutes faster than the second-place racer. Some 200 bicycles started the race. At the first cycle show in Paris that year, primitive freewheels and variable gears were shown.

Rowley B. Turner of the Coventry Sewing Machine Company, England, rode one of the Michaux bicycles from Coventry station to the factory and persuaded the management to make 400 of them. Originally intended to be sold in France, where capacity was not enough for export, the outbreak of the Franco-Prussian War left Turner to sell them in England. James Starley, an inventive young foreman of the Coventry Company who immediately began development work on the primitive machines, became known in England as the father of the bicycle industry. A statue of him is standing today in Coventry.

Starley set out to reduce the weight of the clumsy *vélocipèdes*. In 1870 he made a bicycle with a large front wheel and a small rear wheel (Figure 1, right), derisively nicknamed "penny-farthing" after the largest and smallest English copper coins of the period. He developed a gear that allowed the wheel to be turned twice for each revolution of the pedals. He lightened the wheels by making them of iron with wire spokes under tension. His spokes were a single reel of wire looped through holes in the rim and the hub to which he applied tension by screwing up the threads, an arrangement improved further by the introduction in 1874 of eyed and threaded nipples to hold the spokes individually. Later that year Starley thought of tangential spoking—as distinct from radial—to ease the sideway stresses on the spokes. A party of riders rode these high bicycles from London to John O'Groats, some 690 miles, in 15 days. These machines typically weighed about 50 pounds but could be built as light as 21 pounds for track racing, with a driving wheel varying from 40 to 60 inches in diameter, according to the owner's leg length.

The safety bicycle

Also in 1874 H.J. Lawson built a rear-drive machine with an endless chain between the driving sprocket and the rear wheel. It had two medium-sized wheels of equal diameter. Called the safety bicycle, it had decisive advantages in stability, braking, and mounting over the high front-wheeled "ordinary," which, however, survived for some years. Solid rubber tires improved its stability, but safety machines like Lawson's became more common. In 1888 John Boyd Dunlop, a Belfast veterinarian studying and working with James Moore, introduced a pneumatic tire. The "safety" and Dunlop's pneumatic tire made the bicycle the success it became and killed manufacture of the "ordinary." By 1893 the diamond-pattern frame had been established as the cheapest and most common. The newer models could freewheel and were easily braked.

The next improvement was the introduction of gears. Patents based on the epicyclic principle, using sun and planet wheels inside an annulus ring, were taken out by H. Sturmey and J. Archer between 1901 and 1906. Sturmey-Archer gears, first two speed, then three, were located inside the rear hub of a bicycle and weighed about 2 pounds. Devices of this type also were made in other countries, all using the same principle.

Dérailleur gears

Dérailleur gears—*i.e.*, gears that moved or derailed the chain from one sprocket to another—were less successful at first because mud from the road interfered with their operation, but eventually they proved highly reliable and convenient. Another important innovation came from an offshoot of the bicycle, the tricycle: the differential, a device permitting two driven rear wheels to maintain way while rotating at differing speeds, as in rounding a corner. Bicycle design from 1893 to 1962 was more or less static: only refinements to existing designs were made until 1962, when the cross frame was introduced.

MODERN BICYCLES

Frames. Except for the so-called looped frame, designed to provide clearance for women's skirts, bicycles are produced in two basic types, the diamond frame and the cross frame. They differ widely in principle.

Diamond frame. The diamond frame, the only type made in volume until 1962, consists of three triangles and a pair of forks (Figure 2, top). The rider straddles the main triangle. The other two triangles hold the rear wheel, one on either side. The forks, which can be turned from side to side, hold the front wheel.

By courtesy of Raleigh Industries, Ltd.

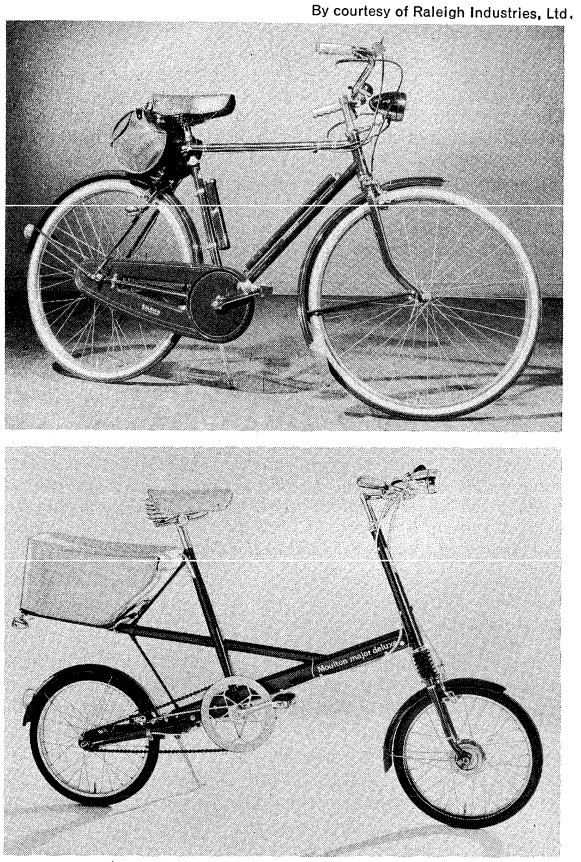


Figure 2: (Top) Modern diamond-frame touring bicycle. (Bottom) Small-wheeled Moulton cross-frame bicycle with patented suspension.

The rider sits on a saddle at the top of the seat tube. Besides holding the down tube at the front and the chain stays at the back, the lug at the bottom of the seat tube also contains the bearings for the cranks and driving sprocket.

Early patterns had straight inclined forks, but it became apparent that curved forks would furnish two advantages: they would absorb some road shock through natural springiness and, under the weight of the rider, would tend to return to the straight-ahead position after being turned.

Cross frame. In 1962 Alexander Moulton, an English engineer, used the cross frame to construct the first really new bicycle since the safety bicycle (Figure 2, bottom). It had one large tube of oval cross section as its main horizontal member. From this main member projected two parallel tubes: the seat tube and the head tube. The latter held the bearings for the steering forks. Parallel with the main tube at the rear were two tapered forks to

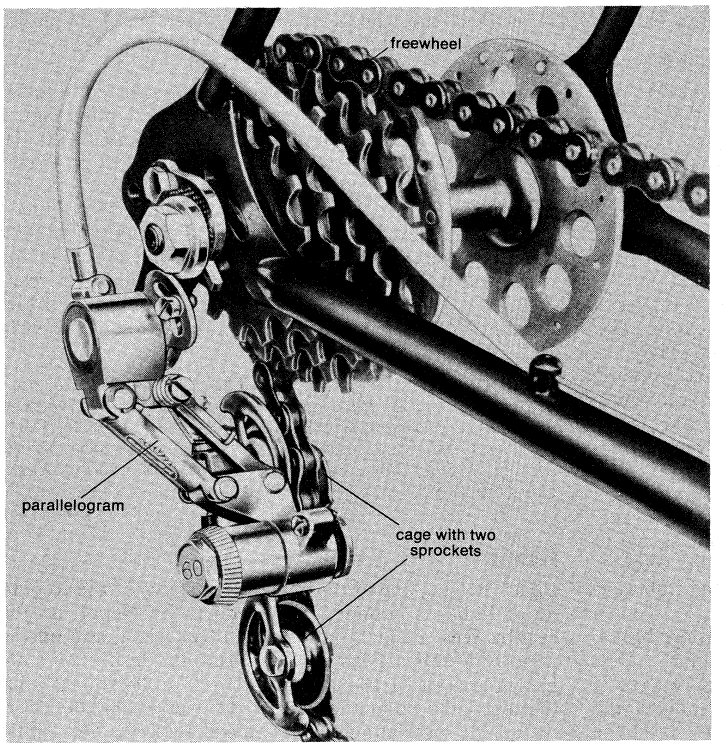


Figure 3: *Dérailleur* variable transmission. When the parallelogram is straightened by pulling the cable, the two sprockets move sideways, allowing the chain to be derailed from one to another of the five sprockets on the freewheel.
The Cyclo Gear Co., Ltd.

hold the rear wheel so arranged as to pivot on each side of the tube about a point approximately three inches forward of the end of the main tube. A small block of rubber was bonded to the end of the tube and to a crosspiece between the tapered forks. Vertical movement of the fork ends subjected the rubber block to compression and shear stress. The front forks were mounted so the rider could turn them while the forks moved vertically. The vertical movement was controlled by a column of rubber mounted inside a coil spring, which provided a form of damping (checking of vibration or oscillation).

Advantages of the Moulton frame

Moulton originally wished to reduce inertia by using very high pressure (about 120 lb/sq in. [8 atm]) tires on small wheels. He adopted 16-inch (400 mm) wheels, as against the more normal 26- or 27-inch (660–686 mm) and included the rubber suspension to reduce the jolting resulting from small wheels and hard tires. The drive was conventional. The suspension, helped by the open frame with no top tube, was a success. Moulton designed it so that the saddle could swiftly be adjusted for riders of varying heights. Behind the seat tube and integral with the frame, he mounted a baggage carrier that, because of the small wheels, was nearer the ground than on an orthodox bicycle and was more stable with heavy loads.

A later development of the design replaced the block of rubber at the end of the main tube by a small metal triangle, free to move slightly up and down, separated from the seat tube by a hard rubber ball, an arrangement that gave a smoother ride.

Within a few years every manufacturer was producing a small-wheeled bicycle, but none enjoyed the Moulton's success, due primarily to its patented suspension.

Wheels and tires. A bicycle wheel consists of a rim to hold the tire, wire spokes under tension, and a ball-bearing hub. Spokes are laced tangentially and maintained under tension by nipples in the rims that can be adjusted to keep the rim true.

Hubs. The hubs may be held in the frame either by nuts screwed onto a spindle or by a cam-action lever utilizing a hollow spindle. The spindle also holds the cones of the ball races, which are used for adjustment and to take up wear in the races that are integral with the hub barrel.

Rims. Rims come in a variety of sizes, from 14 to 28 inches in diameter, and in three basic types: Westwood, Endrick, and tubular. Westwood rims are simple in form, broad with a depression for the spoke nipples in the centre of the section. Rolled sides contain the wired edges of the tires. Endrick rims are flat sided for use with caliper brakes, which grip the rims. Tubular rims are made in the shape of a flat U with a hollow section formed by extrusion. They are used with tubular racing tires, which are attached to the rim with an adhesive rather than wired on. Occasionally racing rims are made from a light wood.

Tires. Tires have an outer cover of canvas and rubber and an inner inflatable tube. The outer cover may either be stitched together to contain the inner tube, as in a racing tire, or have circles of light, springy wire in the edges to keep the cover circular. While outer covers normally contain a large proportion of natural rubber, inner tubes most frequently are made from butyl.

Racing tires

In specialized racing tires the inner tubes are normally of ultrathin natural rubber, and in the interests of lightness and strength, the canvas may be of finer cotton or silk. These tires are usually run at pressures up to 150 pounds per square inch (10 atm). The most common type of tire valve utilizes a restricting cone actuated by the pressure in the tire.

Chains. Although experiments have been made with shaft transmission and bevel gears, as well as with belt transmission, bicycles generally are driven by chains. Roller chains are assembled from pin links and roller links. A pin link consists of two side plates connected by two tightly fitted pins. A roller link consists of two side plates connected by two tightly fitted bushings on which hardened steel rollers are free to rotate. When assembled the pins are a free fit in the bushings. The chains require liberal quantities of oil or grease to protect them from rust and to keep them lubricated, but experiments with plated chains seem to indicate that a dry chain, capable of taking great loads, may not be far off.

Variable transmissions. Devices to improve the speed of the bicycle for a given pedalling rate are of two types: the *dérailleur* (Figure 3), which requires movement of the chain from sprocket to sprocket, and the epicyclic, which

alters the speed of the driver sprocket relative to the rim of the wheel.

Dérailleur gears. A *dérailleur* gear consists of a mechanism to move the chain from one sprocket wheel to another of different size. By varying the size of the driving sprockets, the rear wheel can be made to pass through more or fewer revolutions for each turn of the crank. Cyclists can use up to six sprocket wheels on the rear freewheel and three on the cranks, providing 18 different ratios. The mechanism is spring loaded to absorb chain slack and is controlled by a cable from the frame or handlebars.

Epicyclic gears. Epicyclic gears are made in England, Germany, Japan, and the U.S., in two-, three-, four-, and five-speed models incorporated in the rear hub. They are controlled by levers or twist-grip mechanisms, utilizing cables mounted on the handlebars or on the frame tubes. A two-speed hub gear is also made in brake-hub form, controlled by rotating the sprocket in the reverse direction.

Coaster hub brake

The Sturmey-Archer Company in the 1920s patented a combination three-speed gear and coaster hub brake. Back-pedalling causes a phosphor-bronze split sleeve to expand against the hub shell, resulting in smooth, weatherproof braking action. In this form the hub gear still had to be changed by lever and cable.

Brakes. The most common brake is the coaster brake. It operates on the rear wheel and leaves the rider's hands free. The coaster brake can be installed with a two-speed gear, also free from the handlebars. It is the only brake unaffected by oil or water and in normal use lasts the life of the bicycle with no more than occasional adjustment. Besides the internal coaster, bicycle brakes are made in three principal types: caliper, rod, and drum.

Caliper brakes function by squeezing two blocks of high-friction material against the sides of the wheel rim. Rod brakes consist of two similar pads of friction material that are pulled against the rim from the inside by rods connected to handlebar levers. Drum brakes may be operated by cable or rod connected to a lever that, when rotated, opens two arcs of friction material mounted on shoes inside a steel drum on the hub.

Saddles. Saddles are commonly made of plastic and felt, with or without extra spring loading. Leather, the original saddle material, is still desirable because it absorbs perspiration. The simplest form of saddle uses spring wire made from high-tensile steel, with a curved cantle (upward projecting rear) plate at the rear and a bolt at the front peak. A saddle in this form, of good quality leather and reasonably used (*i.e.*, not allowed to be soaked continually with water) assumes the shape of the rider's bottom and should last his riding lifetime. About 90 percent of all saddles are built up with a girder steel frame of similar shape on which is mounted a hammock of small steel springs and padding of various sorts, covered by polyvinyl chloride sheet. This type of saddle is easily cleaned, waterproof, inexpensive, and easily mass-produced. A third form, which was becoming popular in the 1970s, is the plastic saddle made by injection molding. It is the least expensive of all but has the disadvantage of resisting changes in shape because of its tough, formed nature. Careful design has made it very popular for racing, in which its resistance to deformation through rain is appreciated.

Lights. Battery lamps, employing tungsten-filament bulbs and dry cell batteries, have long been used. The most common form of lighting is that provided by a small, portable dynamo (generator) mounted on the stays or forks of the bicycle and driven by being forced against the side of the tire. As the tire revolves, a milled wheel on the head of the dynamo turns at a speed sufficient to generate about half an ampere of 6-volt power when the bicycle is travelling about ten miles per hour. Many such devices incorporate the front lamp or the rear lamp in the dynamo itself, running one lamp through a single wire and using the frame as a ground return.

Tools, luggage, and small children are commonly carried in baskets or panniers frequently bolted to lugs that are brazed onto the frame. The rider is protected from spray from the wheels by simple sheet metal or plastic guards bolted to the frame and held off the wheel by wire stays.

Production of bicycles. Bicycles are produced in many countries, but the principal builders are the United States, Japan, and Great Britain. In the U.S. some 6,000,000 bicycles were made each year in the 1970s; in Great Britain, some 700,000 were sold and a further 890,000 exported; and the Japanese bicycle industry produced about 4,500,000 bicycles a year, of which 1,000,000 were exported. The largest manufacturing group in the world was Raleigh Industries, producing some 80 percent of British cycles, accessories, and gears in a group of factories at Nottingham. Total world production in the early 1970s was over 36,000,000.

BIBLIOGRAPHY. C.F. CAUNTER, *The History and Development of Cycles: As Illustrated by the Collection of Cycles in the Science Museum*, 2 vol. (1955–58); P.L. SUMNER, *Early Bicycles* (1966); A.J. PALMER, *Riding High: The Story of the Bicycle* (1956); *Bicycling*, formerly *American Cycling Magazine* (monthly).

(A.H.G.)

Bihār

Bihār (Behar), one of the constituent states of the Republic of India, is bounded on the north by Nepal, on the east by West Bengal, on the south by Orissa, and on the west by Madhya Pradesh and Uttar Pradesh. The state has an area of 67,184 square miles (174,008 square kilometres) and a population (1971) of over 56,000,000; it is one of the most populous of the Indian states. Its capital is Patna, and the summer capital is Rānchī.

Bihār occupied an important position in the early history of India; for centuries it was the principal seat of imperial powers and the main focus of Indian culture and civilization. Under the British it formed a part of the Bengal presidency until 1912, when the province of Bihār and Orissa was formed. In 1936 Bihār was separated from Orissa and constituted a separate province, and, on India's independence in 1947, it became a state. In 1948 the small states of Saraikela and Kharsāwān were merged with Bihār. In 1956, when the Indian states were reorganized on a linguistic basis, 3,140 square miles of territory was transferred from Bihār to West Bengal.

The state is naturally divided into two parts—the alluvial plains of the Ganges (Ganga) in the north and the Chota Nāgpur Plateau in the south. The plains districts are popularly known as Bihār proper and the plateau districts as Chota Nāgpur. Bihār proper is further divided into two parts by the Ganges, and it is customary to call these North Bihār and South Bihār. This designation of the central portion of Bihār as South Bihār is somewhat misleading but has the sanction of usage.

The name Bihār (Vihār) dates back to about AD 1200 and was given by the Muslim invaders who were struck by the large number of Buddhist monasteries (*vihāras*) they saw, especially near the then-capital of Odaṇtapurī (present-day Bihārsharif). (For an associated physical feature, see GANGES RIVER.)

HISTORY

The ancient kingdoms

In the early Vedic period (beginning about 1500 BC) several kingdoms existed in the Bihār plain. North of the Ganges was Videha, one of the kings of which was the father of Princess Sītā, the heroine of the *Rāmāyaṇa*, one of the two great national epic poems of India. During the same epoch, the capital of the ancient kingdom of Magadha was Rājagṛha (modern Rājgīr), about 45 miles southeast of Patna; to the east was the kingdom of Aṅga, with its capital at Campā (near Bhāgalpur). A new kingdom later arose in southern Videha, with its capital at Vaiśālī. By about 700 BC, the kingdoms of Vaiśālī and Videha were replaced by a confederacy of the Vṛjjis—said to be the first republican state known in history. It was in Magadha, in the 6th century BC, that Buddha developed his religion and that Mahāvīra, who was born at Vaiśālī, founded the religion of Jainism.

In about 475 BC the capital of the Magadha Empire was located at Pāṭaliputra (modern Patna), where it remained

under Aśoka (emperor of India from about 273 to 232 BC) and the Guptas (a dynasty of emperors who ruled India in the 4th and 5th centuries AD) until the onslaught of the Huns in the middle and late 5th century. In the 6th–7th century AD, the city was devastated by the migration of the Son River—the Chinese pilgrim Hsüan-tsang recording that in AD 637 the city had few inhabitants. It regained some of its glory, but it is doubtful that it ever served as the capital of the Pāla Empire (which lasted from about 775 to 1200). During the ensuing Muslim period (about 1200 to 1765), Bihār had little independent history, remaining a provincial unit until 1765, when it came under British rule and—together with Chota Nāgpur—was merged with the state of Bengal.

Originally, Chota Nāgpur was mostly forest clad and was ruled by chiefs of various aboriginal tribes. Though British authority was only gradually established in the plains to the north during the second half of the 18th and the beginning of the 19th century, occasional revolts against them took place in Chota Nāgpur, the most important being the Ho revolt of 1820 to 1827 and the Munda uprising of 1831 to 1832.

Later, Bihār was an important centre of the Indian movement of 1857 to 1859 against British political authority. Bihār subsequently played an active role in the successive phases of Indian nationalism. Mahatma Gandhi (*q.v.*), the nationalist leader who advocated nonviolent resistance, first launched the *satyāgraha* (passive-resistance) movement against the tyranny of the European indigo planters in the Champaran region of northern Bihār. Rajendra Prasad, who played a leading part in the freedom movement and was elected the first president of independent India, was born in the Saran district, northwest of Patna.

THE LANDSCAPE

Physiography. Bihār, as has been pointed out, consists of two physiographic units—the Gangetic Plains in the north and the Chota Nāgpur Plateau in the south. The plains form part of the middle Gangetic Plains and are in turn divided into north and south plains by the Ganges River. Except for the Himalayan foothills in the extreme northwest, the north Ganges Plain forms a flat alluvial country, less than 250 feet above sea level and liable to flooding. The Ghāgara, the Gandak, the Bāghmati, the Kosi, the Mahānanda, and other rivers flow down from the Nepal Himalayas and make their way to the Ganges in frequently changing channels. Depressions and lakes mark the abandoned courses of streams. The soil consists mostly of new alluvium—chalky and light-textured (mostly sandy loam) west of the Burhi (Old) Gandak River and nonchalky and heavy-textured (clay and clay loam) to the east.

The plains and the plateau

The south Ganges Plain is more diversified than the north, and many hills rise from the level alluvium. The rivers, with the exception of the Son, are all small; their water is diverted into irrigation channels. Beyond the low-lying, treeless country immediately to the south of the Ganges levee (embankment), there are no swamps or waterlogged areas. The soil consists mainly of older alluvium, composed of a darkish clay or yellowish loam, with poor, sandy soils predominating toward the south of this region.

The Chota Nāgpur Plateau, a series of plateaus, hills, and valleys, covers the southern half of Bihār and consists mostly of crystalline rocks. The main plateaus, Hazāribāgh and Rānchī, are separated by the faulted, sedimentary coal-bearing basin of the Dāmodar River, and they average about 2,000 feet in height. In the west there are more than 300 dissected but flat-topped plateaus over 3,000 feet high, known as *pats*. The highest point in Bihār is formed by the conical granite peak of Parasnāth—4,477 feet (1,365 metres) high—in Hazāribāgh; it is sacred to both the Jaina religion and to the Santāl tribe. In the extreme northwest, beyond the Son Valley, lies the Kaimur Plateau, with horizontal sandstone strata underlain by limestone. In the Dāmodar Valley, the soil is sandy; the typical soil of the plateau is red soil, and there are patches of lateritic (red, leached, iron-bearing) soil on the uplands.

Climate. There are three well-defined seasons: the hot-weather season, lasting from March to mid-June; the rainy season, from mid-June to October; and the cold-weather season, from November to February. May is the hottest month, with the mean temperature exceeding 90° F (32° C), except in the extreme north and the plateaus of Rānchī and Hazāribāgh. The normal annual rainfall varies from 40 inches in the west central part to more than 60 inches in the extreme north and in the southwest. The rainfall on the plateau—over 50 inches—is heavier than on the plains. Nearly all the rain (85–90 percent) falls between June and October, and nearly 50 percent of the annual rain falls in July and August. The cold weather season is the pleasantest part of the year.

Vegetation and animal life. The natural vegetation is deciduous forest, but less than one-fifth of the total area is forested; most forests occur in the Himalayan foothills and on the Chota Nāgpur Plateau. In the Himalayan foothills, valuable sal (a resin-yielding species) is found, and bamboo, reeds, and grass are widespread. Chota Nāgpur forms a rich sal area; other timbers include some that are used for the production of lac (a resinous substance used to make varnishes), while tussah silkworms are fed on the leaves of the asan tree (*Terminalia tomentosa*). Mahua (an East Indian tree) yields sweet, edible flowers, also used in the distillation of liquor. Bamboo and sabai (a valuable Indian fibre grass also known as bhabar) of Chota Nāgpur supply raw materials for paper manufacture. Common trees of the plain are the banyan, pipal, and palmyra palm.

Tigers, leopards, elephants, and bears are found only in the more inaccessible forests.

THE PEOPLE

Population. *Demography.* According to the census of 1971, the population of Bihār was over 56,000,000—an increase of about 21 percent since 1961. The average density is a little over 800 persons per square mile, which is almost twice as high as the Indian average. Nearly three-fourths of the population is concentrated in the cultivated plains. The average density per square mile is over 1,200 in North Bihār, over 800 in South Bihār, and under 500 in Chota Nāgpur. North Bihār is one of the most congested tracts in India, and the density in the Muzaffarpur, Saran, and Darbhanga districts exceeds 1,500 persons per square mile. Density decreases toward the east because of the ravages of the Kosi River, which is liable to flood or to change its course, but the position has begun to change as a result of the harnessing of the river. On the South Bihār Plain, a highly developed system of irrigation supports a large population. Density declines toward the south.

Population densities

Settlement in Chota Nāgpur is confined largely to river valleys, deforested peneplains (areas reduced almost to plains by erosion), and mineral and industrial belts. Density exceeds 1,300 people per square mile in the coal-mining and industrial district of Dhānbād, but elsewhere it is between 200 and 400 and is determined primarily by the extent of land available for cultivation.

Rural–urban distribution and characteristics. The great majority of the people, about 90 percent, live in villages. Nearly 80 percent of the total population live in villages of small (less than 1,000) and medium (1,000 to 5,000) size. Compact or clustered villages are usually found in the plains, while dispersed rural settlement is characteristic of the plateau. Aboriginal tribes are concentrated in Chota Nāgpur, especially in the districts of Rānchī (where they constitute 62 percent of the population), Singhbhūm (47 percent), and Santāl Parganas (38 percent). Santāl, Oraon, Munda, and Ho are the principal tribes and together constitute four-fifths of the total tribal population of 4,200,000.

With little more than 10 percent of its population classified as urban, Bihār is one of the most rural states in India. Bihār has more than 160 towns, including nine cities with populations of over 100,000 (1971 census)

—Patna (474,000), Gayā (180,000), Bhāgalpur (173,000), Muzaffarpur (127,000), Darbhanga (132,000), Rānchī (176,000), Monghyr (102,000), Bihār (100,000), and Jamshedpur (341,000). These cities, together with the town groups of Dhānbād–Jharia–Sindri (171,000; 1971 figures) and Bokāro Steel City-Chas (108,000; 1971 figures), account for over two-fifths of the total urban population.

Employed workers constituted 41 percent of the total population in 1961. Three-fifths of the total nonemployed urban population consists of females. Even in the tribal belt of Chota Nāgpur, where both men and women participate in economic activities, the high participation of females in work is confined mainly to rural areas. Of the rural working population, more than 80 percent is engaged in agriculture and only 6 percent in industry, mostly household (cottage) industry.

Religions and linguistic patterns. Hindus constitute about 85 percent and Muslims 12 percent of the population. Christianity and tribal religions are confined to Chota Nāgpur. Muslims are important (about 17 percent) in North Bihār, particularly in Purnea in the northeast. Of the tribal population, about 70 percent are Hindus, 10 percent are Christians, and 20 percent adhere to tribal religions. The Ho is the only tribe in which the majority follow tribal religion; Kharia is the only tribe in which the majority are Christians. Christianity is significant among the Mundas and Oraons (26 and 24 percent).

Indo-European languages

Indo-European languages—including Hindi; the dialects of Bhojpuri, Maithili, and Magahi; and Urdu—are spoken by almost 90 percent of the population. Hindi is spoken by about 44 percent, the Bhojpuri, Maithili, and Magahi dialects by about 35 percent, and Urdu by about 9 percent of the people. Bhojpuri is spoken in the western districts of Shāhābād, Saran, and Champaran; Maithili, in Darbhanga and Saharsa; Magahi, in Patna, Gayā, and Monghyr. Austro-Asiatic (Mundari, Santali, Ho) and Dravidian (Oraon) languages are spoken by about 7 percent of the total population and are confined to the aboriginal tribes.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Bihār has a bicameral legislature. The upper house, the Legislative Council, consists of 96 members, and the lower house, the Legislative Assembly, consists of 318 elected members. Appointed by the president of India, the governor is the constitutional head of the state and functions on the advice of the chief minister, who is the head of the Council of Ministers. The bureaucratic hierarchy located in the Patna secretariat is headed by a chief secretary.

The state is divided into four administrative divisions and 17 districts consisting of Patna division (Patna, Gayā, and Shāhābād districts); Tirhut (Saran, Champaran, Muzaffarpur, and Darbhanga); Bhāgalpur (Monghyr, Bhāgalpur, Saharsa, Purnea, and Santāl Parganas); and Chota Nāgpur (Hazāribāgh, Rānchī, Palāmau, Dhānbād, and Singhbhūm). Local administration is the responsibility of a divisional commissioner in each of the divisions, a district magistrate and collector in each of the districts, and a subdivisional officer in each of 58 subdivisions.

There is a separate administration for development, consisting of a commissioner at the state level, a district magistrate assisted by a district development officer at the district level, and a block (district subdivision) development officer assisted by committee, or *pañcāyat samiti*, at the local level. Below the blocks are *grām pañcāyat*s, or elected village councils. The rural administration (*pañcāyati rāj*) has a similar district, block, and village organization.

The police administration is headed by an inspector general, assisted by superintendents at the district level. There is a high court at Patna, with a chief justice and other judges. Below the high court are district courts, subdivisional courts, *munsif*s' (subordinate judicial officers') courts, and village councils.

Education and public health. *Literacy rates.* About 20 percent of the population is literate, males greatly outnumbering females in literacy, with only three literate females for every 14 literate males. Excluding children below five years of age, the literacy rate is 22 percent—35 percent for males and 8 percent for females. South Bihār is the most literate part, while North Bihār, mostly rural and agricultural, is the least literate. In urban areas the literacy rate is 45 percent, as against 15 percent in rural areas. Over one-half of males and one-third of females in urban areas are literate. In rural areas, about 28 percent of the males and only about 6 percent of the females are literate.

Primary, middle, and secondary schools. There has been a remarkable expansion of education since the early 1950s. The number of primary schools rose from fewer than 24,000 in 1951 to more than 44,000 by the 1970s; middle schools from about 2,000 to about 7,000; and secondary schools from 650 to more than 2,000. Enrollment has increased but is still unsatisfactory, especially among the girls and in the higher age groups. Bihār spends less than one-sixth of its total budget on education, more than half of it on elementary education. The per capita expenditure on education is one of the lowest in India—3.9 rupees (Rs. 7.5 = $1 U.S.; Rs. 18 = £1 sterling, on December 1, 1971), or about 50 cents a head. Education is free for children up to age 11. Vocational and technical institutions are sponsored by government departments.

Higher education. Higher education is the responsibility of the six universities: Patna University, the oldest and most important, with over 12,000 students; Bihār University, at Muzaffarpur; Bhāgalpur University; Rānchī University; and Magadh University, at Bodh Gayā. Darbhanga Sanskrit University provides teaching and research facilities in all branches of Sanskrit learning.

The universities

Health. Medical facilities, though improving, are still inadequate outside the towns. Villages are served mainly by allopathic (traditional medical) and ancient Hindu medical (Ayurvedic and Unanī) dispensaries. There are more than 130 allopathic hospitals and almost 600 primary health centres. Large and well-equipped hospitals and medical colleges are located at Patna, Darbhanga, Rānchī, and Jamshedpur, and a medical college has been established at Bhāgalpur. Smallpox, respiratory diseases, dysentery, and diarrhea figure prominently among the causes of death. Plague, cholera, and malaria have been largely controlled. A tuberculosis sanatorium, a mental hospital, and a leprosarium are all located near Rānchī.

ECONOMY

Economic resources. *Minerals.* The Chota Nāgpur Plateau is the richest mineral belt in India, and Bihār produces nearly 35 percent by value of all the minerals mined in the country. Bihār produces almost the entire national output of copper, kyanite (an alumina–silica mineral used in the manufacture of heat-resistant porcelain), and phosphate, half the output of bauxite (a source of aluminum) and mica, almost half the output of coal and china clay, one-fourth of fireclay, and one-fifth of iron ore. Coal accounts for 85 percent of Bihār's mineral production; coal mining employs nearly 200,000 persons. The principal coalfields, all in the Dāmodar Valley, supply nearly all the coking coal of India. Singhbhūm, together with districts in adjoining Orissa state, contains one of the world's richest and best hematite-iron-ore deposits, which have an iron content of from 60 to 68 percent. Copper is mined and smelted near Ghātsīla, in the Singhbhūm district. High-grade bauxite is mined in the Lohārdaga region of Rānchī. The Kaimur hills in Shāhābād provide 60 percent of Bihār's limestone, mostly used for cement manufacture. A mica belt, extending for over 90 miles from east to west and from 12 to 16 miles from north to south, mainly in Hazāribāgh, produces high-quality muscovite (a light-coloured mica). China clay is abundant in the Singhbhūm, Santāl Parganas, and Bhāgalpur districts, and fireclay is found in the Dāmodar coalfields. The kyanite deposits of Kharsāwān in Singhbhūm are the largest in the world. Singhbhūm is also important for manganese, chromite, apatite (rock phosphate, a source of fertilizer), and uranium. Extensive

Mineral resources

pyrite (an iron disulfide mineral used in the manufacture of sulfuric acid) deposits have been found near Amjore, Shāhābād.

Agricultural resources. Out of a total area of about 42,823,000 acres, about 21 percent is forest covered, 5 percent is barren land, 9 percent is put to nonagricultural use, 11 percent is fallow, 1 percent is planted with trees and groves, 1 percent is used for pasture and grazing, 3 percent is cultivable waste, and the net area sown is 49 percent. Pressure of population has pushed cultivation to the furthest limits, and little remains to be developed. Per capita cultivated land is less than half an acre.

The crops

The transitional nature of the climatic zone is reflected in the cropping pattern, which shows a mixture of wet and dry crops. Rice is everywhere the dominant crop, but corn (maize), wheat, barley, gram, oilseeds, and pulses (leguminous plants, such as peas, beans, and lentils) are important supplementary crops. Jute, a crop of the hot, moist lowlands, is found only in the easternmost plain districts. There are three harvests in a year: *bhadai* —sown from May to June and harvested in Bhado (August to September); *aghani*—sown in mid-June and harvested in the month of Aghan (December); and *rabīʿ*—harvested in spring. Rice, essentially an *aghani* crop, covers 50 percent of the total cropped area. Areas of densest cultivation lie east of the Burhi Gandak, in North Bihār, and in the irrigated area of Patna division. In the plateau region, rice is cultivated in the valleys.

Maize, a *bhadai* crop, is grown principally in the northwest, west of the Burhi Gandak. In the plateau it is the most important food crop after rice and is grown in almost every village. Wheat, the most valuable *rabīʿ* crop, is grown only in the plains, and barley is grown mainly in the northwestern districts. Pulses are the most important crop after rice in the plains. Sugarcane is grown in a fairly well defined belt in the northwest.

Fruits and vegetables are extensively grown. Muzaffarpur and Darbhanga are particularly noted for mangoes, bananas, and litchi fruits. Vegetables are important in the vicinity of large towns. The potato-growing area near Bihārsharif, in Patna district, produces the best variety of seed potato in India. Chilies and tobacco are important cash crops on the banks of the Ganges.

Industrial resources. Despite rich mineral resources, Bihār is one of the least industrialized states in India, since minerals are sent out of the state unprocessed. Less than 10 percent of the working population is employed in industry, as against more than 75 percent in agriculture; more than 70 percent of the industrial workers are engaged in household industry. Of the remaining industrial workers, only 40 percent are employed in the organized sector, principally in steel and other metal-based industries and food-processing industries.

Industry

Regional economic development. Regional industrial distribution shows heavy concentration in the two plateau districts of Singhbhūm and Dhānbād. It is, however, possible to recognize several significant zones of economic development. Singhbhūm, the richest mineral-bearing district, is important for heavy industries. Jamshedpur, the seat of ironworks and steelworks, has also attracted a number of satellite engineering industries. Copper is smelted at Maubhandār near Ghātsīla. Chaibāsa manufactures cement from Jamshedpur slag, and there is sheet-glass manufacturing at Kāndra.

The Dāmodar Valley has coal-washing plants, coking plants, and large coal-burning thermal-power plants at Bokāro, Chandrapura, and Patratu. Sindri produces fertilizers and cement, and there is another cement factory at Khelari. Other important manufactures include alumina at Muri and sheet glass at Bhurkunda (Barka Kāna), in Hazāribāgh. Rānchī, just south of the Dāmodar Valley, with its heavy-machine building, foundry forge, and heavy-machine-tool plants, has grown into an important centre of heavy engineering. A large steel plant is nearing completion at Bokāro.

The Bihār mica belt has no significant manufacturing industry; processing is by skilled hand labour. Factories in this area are located mostly near Giridih, Kodarma, and Jhumri Tilaiya.

Dālmianagar is the focus of the Son Valley industrial area and has paper, cement, sugar, chemical and other industries. Japla and Kalyanpur produce cement.

Light engineering and consumer industries have developed in the trading centres situated along the south bank of the Ganges at Patna, Mokameh, Monghyr–Jamālpur, and Bhāgalpur. North of the Ganges lies the petrochemical complex at Barauni, connected by a pipeline with the Assam oil fields. Monghyr is noted for cigarette manufacture, and Bhāgalpur is noted for silk. The sugar belt of Tirhut manufactures four-fifths of the state's sugar. Rice milling is situated in the extreme north, where paddy (rice) for husking also comes from the adjoining Nepal Terai (a marshy lowland area). In the jute-growing area, there are jute mills at Katihār and Samastipur.

Transport and communication. *Railways.* The waterways, once important, are now of little significance. North Bihār is served by a narrow-gauge railway. Because of the dense population, the railways carry a heavy load of traffic. They run parallel with the rivers because of the difficulty of constructing bridges, and communication between important towns is consequently often long and tedious.

The rail network

South Bihār has better rail facilities. Three lines starting from Calcutta and passing through Sāhibganj, Madhupar and Gomoh converge onto the plain. There are road and rail bridges on the Son at Dehri and at Koilwar. A rail and road bridge over the Ganges at Mokameh connects South with North Bihār.

In Chota Nāgpur the railways naturally follow the terrain. Extensive rail development has taken place in the coal-mining areas of the Dāmodar Valley. One of the railroads connecting Calcutta and Bombay passes through Jamshedpur via the Sanjai–Koel gap. There are, however, large areas only poorly served by rail transport. The mica-mining area in Hazāribāgh is not directly served by any railroad.

Roads. Of the total motorable-road mileage of 27,000, only about 7,000 miles are paved. In North Bihār most of the roads are unpaved. Dense population, inadequate rail facilities, and the requirements of the sugar industry have resulted in a dense network having been constructed toward the west, although the eastern half is poorly provided with roads. In South Bihār the important towns are well served with roads, but paved roads are frequently interrupted by unpaved stretches. Road accessibility is best around Patna and Gayā. The Chota Nāgpur Plateau has good roads, built principally during World War II. The Grand Trunk Road, a national highway, cuts across the plateau through Dhānbād and Dehri. Another national highway connects Patna with Rānchī and Jamshedpur.

CULTURAL LIFE AND TRADITIONS

The cultural regions of Bihār show a close affinity with the linguistic regions. Maithili is the language of old Mithilā (the area of ancient Videha, now Tirhut), which is dominated by orthodoxy and Brahminical way of life. Maithili is the only Bihar dialect with a literary history; one of the earliest and most celebrated writers of Maithili was Vidyāpati (15th century), noted for his lyrics of love and devotion.

Bhojpuri, a virile dialect spoken by an active and enterprising people, has hardly any literature but does have a considerable oral folk literature. Magahi too has a rich folkloric tradition. The Bihār Plain has also contributed significantly to modern Hindi and Urdu literature.

Tribal customs and traditions

Most tribal villages have a dancing floor (*akhāra*), the sacred grove (*sarna*)—where worship is offered by a village priest—and a bachelor's dormitory (*dhumkuria*). All the tribes believe in one supreme being. The *hāt*, or weekly market, plays an important part in tribal economy. Tribal festivals (such as Sarhūl), a spring festival (Sohrai), and a winter festival (Mage Parab) are occasions of great festivity. Tribal culture is fast changing under the impact of external influences, such as Christianity, industrialization, new communication links, tribal welfare schemes, and community development projects.

Places of religious and cultural interest abound in the plains. Nālandā is the seat of the ancient celebrated Nālandā monastic university; nearby Rājgīr, with its ancient and modern temples and shrines, is visited by people of many faiths; Pāvāpurī is the place where Mahāvīra, the founder of Jainism, attained Nirvāna (freedom from an endless cycle of reincarnation). Gayā is an important place of Hindu pilgrimage. Nearby Bodh Gayā, where Buddha attained Enlightenment, is Buddhism's holiest place. The holy town of Deoghar is well-known for its Baidyanāth temple. Hariharkshetra, near Sonepur, north of Patna, is famous for one of the oldest and largest animal fairs in India, held in November.

BIBLIOGRAPHY. P. DAYAL, *Bihar in Maps* (1953), a source of reference for Bihār geography: "The Bihar Plain: A Regional Study," *Trans. Indian Council of Geographers*, vol. 5 (1968); R.R. DIWAKAR (ed.), *Bihar Through the Ages* (1959), an authoritative cultural and political history up to the post-independence period with introductory chapters on geology and geography; SIR JOHN HOULTON, *Bihar: The Heart of India* (1949), a popular account of history and culture; NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of Bihar*, vol. 1 (1959), a guide to economic resources and potentialities; *Census of India, 1961*, vol. 4 *Bihar*, pt. 1-A, *General Report on the Census* (1969), and pt. 9, *Census Atlas of Bihar* (1968); E. AHMAD, *Bihar: A Physical, Economic and Regional Geography* (1965), sections on population and settlement; DIRECTORATE OF STATISTICS AND EVALUATION, BIHAR, *Bihar Through Figures* . . . (annual).

(P.D.)

Bikini

Bikini is an atoll in the Ralik (western) chain of the Marshall Islands in the central Pacific Ocean. Lying north of the equator, at 12° N, 165° E, it is 225 miles (360 kilometres) northwest of Kwajalein (the world's largest coral atoll) and 190 miles (305 kilometres) east of Eniwetok Atoll. It forms part of the Trust Territory of the Pacific Islands, since 1947 a United Nations strategic area trusteeship administered by the United States. Bikini was a site for peacetime atomic explosions conducted for experimental purposes by the United States between 1946 and 1958. The population of the island, numbering less than 200, was moved to other islands before the tests began; some of the islanders and their descendants are to return to the atoll shortly. The word "bikini," describing a very brief two-piece swimsuit for women, which came into vogue at the time of the atomic tests, has entered the international vocabulary. (For detailed coverage of the Pacific region, see PACIFIC ISLANDS and PACIFIC OCEAN; for coverage of the system by which it is administered, see PACIFIC ISLANDS, TRUST TERRITORY OF THE.)

Natural environment. Because of its association with nuclear testing and related environmental investigations, Bikini is one of the most thoroughly examined atolls in the world. It consists of a small group of islands of which the average elevation is only some seven feet above low tide level. The land area of the group amounts to little more than two square miles of dry land, distributed about the edges of an oval lagoon 25 miles long and 15 miles wide, with a water area of 243 square miles and a maximum depth of about 200 feet. The largest islands are Bikini, which forms a narrow arc to the northeast, and Enyu (or Eneu), a slightly smaller island situated at the southeast curve of the reef circling the lagoon. The principal entrance to the lagoon, near Enyu Island, is Enyu Passage, which is nine miles wide. Bikini stands in the path of the westward-flowing North Equatorial Current and in the region of the northeast trade winds. Storms occur infrequently, and the average mean monthly temperature varies only from 80° to 83° F (27°–28° C).

History. Bikini, before World War II known as Escholtz Atoll, was placed on Pacific charts by Otto von Kotzebue, the Russian circumnavigator, who in 1825 sailed past it without landing. Kotzebue noted no inhabitants, but even then Bikini almost certainly had been inhabited for many generations by people of Malayo-Polynesian origin. The Bikinians, together with other Micronesians, were subjected to western and other influences that entered the Pacific in the 19th and 20th centuries, but their small island community was only lightly touched by change until after World War II. The Bikinian social system was matrilineal in character, and property was held in matrilineal lines, although community leadership was provided by male representatives of the principal families. The major source of food and fibre was the coconut palm. The pandanus (a member of the screw pine family of plants) and arrowroot (a tropical American plant whose roots yield a nutritious starch) were also common, and the lagoon supplied the islanders with fish, clams, crabs, and other seafoods.

When Japan was driven from the Marshall Islands in 1944, the islands and atolls, Bikini among them, came under the administration of the United States Navy. Soon after World War II, and before the United States had been confirmed in 1947 in its United Nations strategic trusteeship of the Pacific Islands, Bikini in 1946 became the theatre for Operation Crossroads, a vast military-scientific experiment to determine the impact of atomic bombs on naval vessels.

Operation Crossroads test site

Further tests, some of them thermonuclear, were conducted from 1954 to 1958, when Bikini, together with Eniwetok Atoll (190 miles to the west), constituted the Pacific Proving Ground of the United States Atomic Energy Commission.

The atomic tests. In Operation Crossroads, during which the fourth and fifth atomic explosions in history took place, two atomic bombs of the type used in World War II, each of which released energy equivalent to 20,000 tons of TNT, were detonated over and under a target fleet of some 70 ships anchored in Bikini lagoon. The first test, an air drop, took place on July 1, 1946. During the second test, an underwater explosion that took place on July 25, the bomb was detonated at a depth of 90 feet. The explosion produced a million-ton hollow column of water 2,000 feet in diameter, which rose to a height of more than a mile above Bikini lagoon before falling back in a storm of waves, steam, radioactivity, and debris. Radioactivity in the lagoon remained intense for days. Because the products of the atomic explosion had been intermixed with water, the problems presented by radioactivity were far larger, in fact and in implication, than had been anticipated.

In 1947, a year after these explosions, Bikini was subjected to an intense examination by a further scientific survey sponsored by the U.S. Navy. Although radioactivity was found to be present at low levels, scientists could not determine its biological effects and thus could not be certain that Bikini was again habitable. The U.S. Atomic Energy Commission, which then took over nuclear programs, sponsored additional surveys at Bikini in 1948 and 1949 to gather data on the biological disposition of radioactivity.

Further tests, 21 in all, were conducted at Bikini in 1954, 1956, and 1958. The first test at Bikini, held on March 1, 1954, was a 15-megaton thermonuclear explosion, 750 times more powerful than the Operations Crossroads tests. From this explosion radioactive fallout, descending from a cloud 114,000 feet high, unexpectedly spread radioactive dust in a long plume across the ocean east of Bikini, seriously contaminating Rongelap Atoll nearby, as well as a Japanese fishing vessel cruising in the area. This tragic mishap revealed the global nature of radioactive contamination, and led to a widening and intensification of radioenvironmental research. The Rongelap people and their contaminated atoll were of concern for years. The people, hurriedly evacuted from Rongelap in 1954, were repatriated in 1957 but remained under medical observation. Meanwhile the exposure of the Japanese fishermen to "Bikini ashes" created concern in Japan lest the radioactivity from Bikini had poisoned waters flowing toward the homeland. The Japanese scientific community organized a Bikini Waters Investigation Team that in 1954 conducted a seven-week cruise directly into and through the Eniwetok and Bikini waters for the purpose of sampling their radioactivity; useful technical information was thereby obtained. Further surveys were made by U.S. scientific teams in 1955, 1956, and 1958.

Testing at Bikini ceased when a moratorium on nuclear tests was declared in 1958.

With the cessation of testing, Bikini was left for a time undisturbed and unoccupied. The islanders, however, still hoped to return. In 1946 the Bikini community had been moved to Rongerik, a neighbouring atoll, but by 1948 it became evident that Rongerik could not provide enough food for them. After several months spent on Kwajalein, the community was established at Kili Island, in the southern Marshalls, a former German copra plantation that was lush with vegetation but that lacked a protected lagoon for fishing. In 1956 the United States awarded the Bikinians more than $300,000 in compensation. The community grew in numbers. By the early 1970s nearly 400 persons were living on Kili, while some 200 persons living elsewhere also claimed land rights on Bikini.

Scientific surveys of Bikini were conducted under the auspices of the U.S. Atomic Energy Commission in 1964 and 1967. These established that some islands had been altered by erosion or silting, and that the stands of coconut eradicated by atomic heat and blast had not grown again. Particularly on Bikini itself, however, a heavy growth of scrub vegetation had occurred. Radioactivity had declined to levels that were determined to be acceptable for human habitation. On the basis of these studies, it was concluded that the Bikinians could be returned to the Bikini-Enyu Island complex and that, if certain precautions were observed, other islands could be used for sources of food, such as birds, turtles, and certain plants. In 1968 U.S. President Lyndon Baines Johnson announced that Bikini, no longer needed for nuclear operations, was to be prepared for return to the Bikinians.

Resettling the island

Under the auspices of the administration of the U.S. Trust Territory of the Pacific Islands and other U.S. agencies, a resettlement program was projected over a six-year period, at a cost of $3,000,000. By 1969 the islands had been cleared of nuclear test debris and scrub vegetation. As a radiological precaution, the top two inches of Bikini's topsoil were removed before the planting of pandanus trees. Salvageable buildings were reconditioned, and a 4,500-foot airstrip on Enyu was repaired. In 1970 coconut and food crop planting was begun under the direction of Trust Territory agriculturalists, the work itself being done by men from Bikini. Sixty dwellings are to be built, each standing on its own "weto," or parcel of land. When repatriation occurs, Bikinians who wish to remain on Kili will be allowed to do so. In view of the special interest taken in Bikini because of its nuclear past, a return to the coconut economy of other times is unlikely. Further scientific studies are to be conducted, and tourism may be encouraged.

BIBLIOGRAPHY. For detailed accounts of Operation Crossroads tests, see W.A. SHURCLIFF, *Bombs at Bikini* (1947), and D.J. BRADLEY, *No Place to Hide* (1948). For accounts of scientific surveys of Bikini and environs, see K.O. EMERY, J.I. TRACY, JR., and H.S. LADD, "Geology of Bikini and Nearby Atolls," *Prof. Pap. U.S. Geol. Surv. 260-A* (1954). For narrative accounts of operations and studies at Pacific Proving Ground, see N.O. HINES, *Proving Ground: An Account of the Radiobiological Studies in the Pacific, 1946–1961* (1962). For summary analyses of phenomena of nuclear detonation including observations at Bikini and a record of announced detonations, see S. GLASSTONE (ed.), *The Effects of Nuclear Weapons*, rev. ed. (1964).

(N.O.H.)

Billiard Games

The games of billiards take many forms, but fundamentally all are games of skill played with various numbers of balls and a long stick, called a cue, on a rectangular table (traditionally slate topped) covered with a smooth, tight-fitting cloth and bordered by rubber-cushioned rails. Carom, or French billiards, games are played with three balls on a table that has no pockets. The other principal games are played on tables that have six pockets, one at each corner and one at the centre of each of the long sides: these games include English billiards, played with three balls; snooker, played with 21 balls and a cue ball; and pocket billiards, or pool, played with 15 balls and a cue ball. There are numerous varieties of each game, and particularly of carom and pocket billiards.

Billiards has often been referred to as an ancient and honoured game, and many references to it occur in numerous literary works and other books. Unlike many other well-known games, the popularity of billiards over the years has been greatest at the two extreme ends of the so-called social scale. It has always been considered a game of great prestige, and fine billiard tables are frequently present in the most elegant homes of the world. At the same time, billiards is considered in many places to be a very low-class game with a most unfortunate reputation.

Popularity of billiard games

The various billiard games continue to grow in popularity in many countries of the world. Although it was predicted that television would cause games such as billiards to lose their popularity, quite the contrary has happened, and billiard shows on television have, in fact, contributed to the popularity of this sport. The growth of billiards in private homes continues to keep pace with the increasing affluence of many nations. In North America, for instance, more than 500,000 homes have billiard tables installed in family and recreation rooms. The development of modern public billiard halls in many cities has also been substantial. It was reported that over 9,000 new billiard parlors were built in Japan in one year alone in the early 1960s. In addition, there has been continued development of billiards, both in public halls and private homes, in England, Sweden, and France in recent years. With the development of modern billiard tables in new styles using fine woods and brightly coloured billiard cloths, the game has taken on a new look. Modern materials have replaced the traditional slate bed of the table, and this has made it possible to produce an inexpensive but good-quality table for home use.

HISTORY

Origins and early history. The origins of billiards are difficult to trace, and many people have attempted to pinpoint them without success. Some accounts trace the origins of billiards back to the 6th century BC, quoting the Scythian philosopher Anacharsis, who mentions watching a game similar to billiards while travelling in Greece. Other dubious, if not bizarre, resemblances allegedly exist in games played by the Greeks and Romans. In this connection, Jacques Bonhomme, author of economic and social studies on France, writing in 1885, remarked with irony that had billiards been a Roman diversion, Horace would assuredly have devoted an ode to it and Nero would have been diverted from his famed incendiary exploit by so agreeable a pastime.

Attempts to trace origins

Some authorities have placed the origins of the game in France around 1452, and it is known that Clément Marot, who died in 1544, made mention of the game in one of his poems. Many other countries, among them England, China, Italy, and Spain, have been credited with housing the invention of the game; but, in fact, nothing is really known about the origin of billiards. It may be inferred that it developed from a variety of games in which propelling a ball was a main feature.

Billiards has been related to such games as *shovilla bourde*, or shuffleboard, a game popular in the time of Henry VIII, in which objects are shoved by hand or with a stick or other implement so that they come to rest on spaces marked on a table, the floor, or an outdoor court. Some features of the game suggest *paille-maille*, or pall-mall, an early ball and mallet game played on the ground with hoops and believed to be the origin of croquet.

The confusion in the nomenclature of games in general during the different stages of their development is another obstacle to accurate research. Etymology offers little aid, because the derivation of the word billiards from the French *billard*, or *billart* (Old French), meaning a curved stick (crosse, cue), not to mention the diminutive *bille* (stick), is uncertain. "For myself," French billiard champion Maurice Vignaux wrote in *Le Billard* (1895), "*bille*, meaning a ball, is the key, and *billard* comes from it. And *bille*, whence comes this? From *pila* (Latin) says

one authority; and *pila* is derived from . . ." After this he confesses to being lost in an etymological fog.

Literary and artistic evidence

Research concerning the 16th and 17th centuries is on surer ground, however, for allusions to the game were by then plentiful, and old prints, both English and French, exist to provide corroborative evidence. Poets and dramatists such as Edmund Spenser in *Mother Hubberds Tale* (1591); George Chapman (1598), who makes a character say, "Go, Aspasia, send for some ladies who could play with you at chess, at billiards and at other games"; Shakespeare (1623), with his anachronistic "Let us to billiards . . ." from *Antony and Cleopatra*; Ben Jonson (1637); and later Samuel Johnson in his *Dictionary* (1755) all cite the game.

English diarist John Evelyn (1620–1706) made a habit of noting the billiard tables in country mansions he visited. There are also historical allusions, such as the complaint of Mary, Queen of Scots, in 1576, during her captivity, that her billiard table had been taken away. An interesting reference is to be found in the English poet Francis Quarles's *Emblemes* (1635), a highly and popularly regarded book of symbolic pictures accompanied by expositions in verse, in which one of the engravings depicts two angels playing a table game with balls, hoops, and maces (an early form of the cue), the table having pockets.

In some of the old prints, a number of obstacles are seen on the tables, such as hoops, ivory pegs or "kings," forts, and batteries, the player's ball being required either to circumvent such objects without knocking them over or to pass through them, as in the case of hoops. A French print shows the Duchess of Burgundy playing billiards with a mace in 1694. The first description of billiards in English is to be found in Charles Cotton's *Compleat Gamester* (1674).

Apparently, billiards has been popular with the noble and elite from the very beginning of the game. King Henry III, the first of many French monarchs to adopt the sport, installed a primitive table in his château at Blois in the middle of the 16th century.

Development of equipment. *Cues.* The first cues were equipped with curved heads and were similar to a modern hockey stick. These were eventually replaced by the type of cue known as a mace, which consisted of a stick or rod similar to the modern cue except that it had a wooden block on the end, usually covered with baize or leather, and this was used to push the balls where desired.

From square to bevelled cue tips

The cue superseded the mace in the late 1700s, and about 1760 cues with perfectly flat ends came into vogue. Twenty-five years later, a cue cut obliquely at the small end or slightly rounded at one side was produced to enable the player to hit a ball below the centre. Still another change was adopted toward the end of the century, when the point of the cue was bevelled all around, thus making a still broader surface. A leather cue tip invented by a Captain Mingaud of the French infantry followed in 1806, when the advantage of using chalk was discovered.

The modern billiard cue is normally a round, tapered stick, usually of wood and, depending on the size of the table on which it is used, ranging in length from approximately 40 to 60 inches (100 to 150 centimetres) and in diameter from approximately 1.25 inches (32 millimetres) at the large end to about 0.5 inch (13 millimetres) at the small end. The large, or butt, end of the cue is usually fitted with a rubber-cushioned bumper and the small end, with which the ball is struck, is fitted with a plastic, fibre, or ivory reinforcement to which is cemented a leather cue tip that strikes the ball. These developments have resulted in a cue with great durability and uniformity. Though aluminum and plastic cues have been introduced, the wooden cue is still supreme.

Chalk. Players at one time rubbed the leather tips of their cues against whitewashed ceilings to make the leather less smooth and give it a fine coating of lime, or chalk, thus reducing its tendency to slip off the ball if the ball was not struck near its centre. Later, chalk in small cubes was developed to be applied uniformly to the cue tip permitting the players to strike the cue ball off centre on purpose to impart a spinning motion, called "side" in Great Britain and "English" in the United States.

Balls. Until the late 19th century, the most widely used billiard balls were of elephant-tusk ivory. In 1868 a United States printer and co-inventor of celluloid, John Wesley Hyatt, in his search for a better billiard ball, discovered that a mixture of nitrocellulose, camphor, and a small amount of alcohol when properly prepared becomes thermoplastic (*i.e.*, soft when heated) and could be molded in a hydraulic press. After cooling at ordinary atmospheric pressure the ball became hard and strong. This discovery heralded the beginning not only of the composition billiard ball but also of the plastic industry. In the 1920s a new type of plastic billiard ball produced from cast phenolic resin proved to be much more durable and offered a greater brilliance of colours. Although balls made of ivory were credited by the outstanding players with being more sensitive, they were gradually replaced by the new composition balls, which were found to be unaffected by weather and which, unlike ivory balls, did not become imperfect and untrue with age.

The table. In early versions of the game, tables were made of wood and equipped with iron hoops (similar to those used in the lawn game of croquet) or obstacles that later gave way to holes cut in the top of the table. At first, these holes were cut in the centre of the table, but as time passed they were moved to the corners and the sides. In the French game of carom, however, tables without holes continue to be used.

In general, the appearance of the billiard table has followed the styling and materials that were used by the cabinetmakers and furniture manufacturers of the various eras. Modern plywoods and Formica surfaces have been utilized, and plastic has become an even more important ingredient in the manufacture of the table. In many countries, however, the traditional mahogany tables with heavy inlays are still considered to be the finest. The basic structural parts of the table have undergone far less change. One such change involved the introduction of a slate bed instead of the wooden bed and, then, in smaller home tables, the replacement of slate with modern pressboard or chipboard. The purpose of the bed is to provide a level and true surface.

Introduction of slate beds and rubber cushions

The cloth used to cover the table bed developed early in the history of billiards and consisted of a finely woven, green woollen cloth with lustrous nap. Because of its appearance, billiard cloth has also been referred to as a felt instead of a woven fabric. The usual green colour used for billiard cloth, which was the only acceptable colour for over 100 years, is gradually being replaced by the more bright and vibrant colours being used in modern decorating, such as red, blue, and gold, to mention a few.

The rubber cushions that are attached to the rails around the table to contain the balls have seen very little change since John Thurston, an English cabinetmaker, founded a billiard-equipment firm in 1799 and in 1835 introduced an India-rubber cushion to replace the old stuffed cushion built up with layers of list (strips of cloth) or felt. This rubber cushion was a great improvement, but cold weather affected its resilience. In 1839, however, the process of vulcanization was discovered, and in 1845 Thurston obtained letters patent for applying the processes to billiard cushions.

TYPES OF BILLIARD GAMES

The game of billiards is played in many different ways, using variations of the equipment described above. Different forms of the game tend to be played in certain groups of countries or areas of the world, although many of the games cross many national boundaries. The game of carom, which was probably the original game, is still played primarily in France and other European countries, to a lesser degree in the United States, and has many players in Japan, Indonesia, the Philippines, Taiwan, and South Korea and in Central America, South America, Africa, and the Middle East. The game of English billiards is most popular in England and the former empire countries. The game of pocket billiards,

or pool, which uses six large pocket openings, is primarily the game played on the American continents and, in recent years, has been played in Japan. The game of snooker is primarily British and is played to a small degree in the Americas.

Carom, or French billiards, games. Carom billiards is played on a table usually five feet by ten feet (1.5 by three metres) or 4.5 by nine feet (1.4 by 2.7 metres). It has no pockets. The game is played with three balls, two white and one red, with one of the white balls having a small red dot, or spot, to distinguish it. One of the white balls (plain or spot) serves as the cue ball for each player, the red ball and the other white ball serving as his object balls. In play, the object is to stroke the cue ball so that it hits the two object balls in succession, scoring a carom, or billiard, which counts one point; in some carom games the cue ball must also touch a cushion or cushions one or more times to complete a carom. Scoring a carom also entitles the player to another shot and his turn, or inning, continues until he misses, when it becomes his opponent's turn.

Three-ball games on tables with no pockets

All carom games are begun by lagging (see below) to determine rotation of play, except that if more than two players are involved, rotation may be determined by drawing lots. When the game is played by two players (or sides), each player (or side) selects a cue ball and places it on the table within the head string, an imaginary line between the second diamonds from the head of the table (for plan of a carom table see Figure 1). The red

Starting the game

By courtesy of Billiard Congress of America

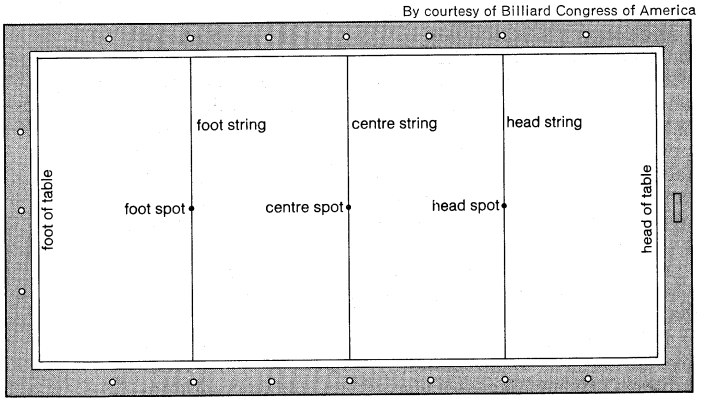


Figure 1: Plan of the carom table.

ball, meanwhile, has been spotted on the foot spot. Each player strokes his cue ball, one to the left of the red ball and the other to the right, to the foot cushion and return: this is known as "lagging," and the player whose ball comes to rest nearest the headrail wins the lag. The winner has his choice of cue balls and also has the right either to break (shoot first) or assign the break shot to his opponent.

Three-cushion billiards. Three-cushion billiards, or three-cushion carom, is the classic form of the game. Its essence lies in the requirement that the cue ball, in addition to striking both object balls, must also touch a cushion or cushions three different times to complete a carom, or scoring sequence for one point. Such a three-cushion carom is scored (1) when the cue ball strikes an object ball and then strikes three or more cushions before striking the second object ball; (2) when the cue ball strikes three or more cushions before contacting the two object balls; (3) when the cue ball strikes a cushion, then the first object ball, then two or more cushions, and then the second object ball; (4) when the cue ball strikes two or more cushions, then the first object ball, then one or more cushions, and finally the second object ball.

The number of cushions struck does not necessarily mean three different cushions; a count may be executed on one cushion with the required number of contacts by the cue ball.

Straight-rail billiards. In straight-rail, the player must drive his cue ball against the two object balls to score a carom for one point. He may contact both object balls at the same time or alternately. Cushion contact is not required, although a count is legal if one or more cushions are contacted. When object balls are crotched (standing in the corner of the table no more than 4.5 inches [11.5 centimetres] out from either rail), three counts are allowed. The player then must drive an object ball out of the crotch. Failure to do so is a miss and ends the inning.

Balkline billiards. The balkline game rules were developed because experts at carom learned how to gather balls and to nurse them along the rails and into the corners for long runs of scoring. In order to restrict this type of carom, balklines were drawn on the table either 14 or 18 inches (36 or 46 centimetres) from the sides and ends, and anchor squares, small seven-inch squares, were marked out where the balklines met the rails. In the games known as 14.1 or 18.1 balkline billiards, only one carom (scoring sequence in which the player's cue ball must contact both object balls, cushion contact being optional as in straight-rail billiards) can be made in a balk area or an anchor area before the balls are driven to another area for further scoring; in 14.2 or 18.2 balkline billiards, two caroms are permitted in a balk or another zone before the ball must be driven out.

Purpose of balklines

Other variations. Other variations of the carom games include bank billiards, in which the cue ball must contact one or more cushions before contacting the first object ball, and red ball, in which the red ball is always the first object ball.

Prominent players. Prominent players of three-cushion billiards have included Americans Jacob Schaefer, Sr., in the late 1800s, and William Hoppe, who dominated world carom title play from 1906 until his retirement in 1952, and his two closest rivals, Welker Cochran and Jacob Schaefer, Jr. More recently a Belgian player, Raymond Ceulemans, world amateur champion for eight consecutive years, has been unrivalled.

Hoppe was also predominant in balkline billiards, first winning the 18.1 world title from Maurice Vignaux of France in 1906. Hoppe and his rivals Cochran and Schaefer, Jr., dominated the game throughout its most prominent era, the years 1919–32, becoming so proficient that audiences became bored, and the championships were discontinued. In Europe, Roger Conti, considered the inventor of modern balkline technique, was outstanding. More recently the European amateurs Jean Marty, Ceulemans, and Ludo Dielis have been able to exceed the records of the professional players of the 1919–32 era.

Governing bodies. The international governing body of carom billiards is the World Billiards Union (Union Mondiale de Billard), with headquarters in Brussels and administrative centre in Barcelona. Its associated and affiliated organizations are the Asiatic, North American, South American, Africano-Levantine, and European confederations, which conduct world championships in accordance with an agenda established by the international association.

Pocket billiard games. There are two basic types of billiard games: the one just described, which is played on a table without pockets, and pocket billiards, which is played on a table with six pockets. The pocket billiards games are more widely played throughout the world and have a greater number of variations, these variations being principally defined by geographical location. In British-oriented countries, for example, the most popular games are English billiards and snooker, which are also played in Europe, whereas in the Western hemisphere and to a limited degree in the Orient, the American type known as pocket billiards, or pool, is the more popular game.

The English billiards game and snooker are played on a table much larger than that used in the American pocket billiards game.

Pocket billiards tables

English billiards and snooker are normally played on a table six feet, 1.5 inches by 12 feet (two metres by 3.7 metres). The American game of pocket billiards is usually played on a table 4.5 by nine feet (1.4 by 2.7 metres), although in special championships the table is sometimes

five feet by ten feet (1.5 metres by three metres), the size of the normal American snooker table, and in some areas of the United States and South America the tables are as small as four by eight feet (1.2 metres by 2.4 metres). Pocket openings on the English billiards and snooker table are much narrower than on the American pocket billiards table. All other aspects of the equipment are very much as noted in the earlier section on billiard equipment. It should be noted that in the British form of snooker, the balls are $2\frac{1}{16}$ inches (5.2 centimetres) in diameter, instead of $2\frac{1}{8}$ inches (5.4 centimetres), which is common in America. For plan of English billiard table see Figure 2; of pocket billiard (pool) table see Figure 3.

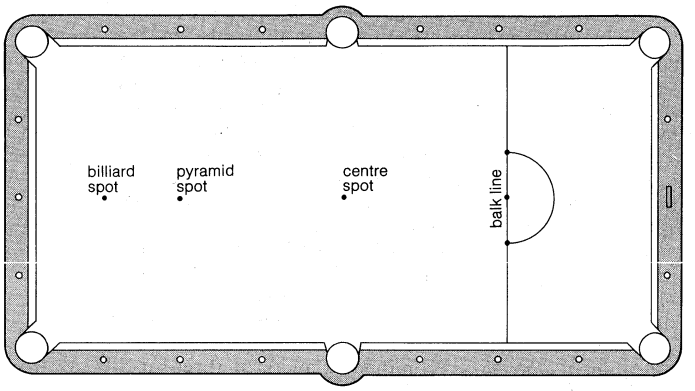


Figure 2: Plan of English billiard table.

English billiards. The game of English billiards, which is played on the relatively large English billiard table, is played with three balls as in carom—a plain white, a white with a spot, and a red. The length of the game is determined either by a period of time or by a number of points that may be agreed to by the players. The game is usually played with two players or two pairs of players, but it may be played by any convenient number. There are three ways of scoring: (1) the losing hazard, or loser, a stroke in which the striker's cue ball is pocketed after contact with another ball; (2) the winning hazard, or pot, a stroke in which a ball other than the striker's cue ball is pocketed after contact with another ball; and (3) the cannon, British term for carom, a scoring sequence in which the striker's cue ball contacts the two other balls successively or simultaneously. Two points are scored for a losing hazard off white, for a winning hazard if white is potted, or for a cannon; three points are scored for a losing hazard off red or for a winning hazard if red is potted. The object of the game is to score more points than one's opponent. The skill involved consists of leaving one scoring stroke after another. A player continues at the table for as long as he succeeds in scoring.

Losing and winning hazards

Snooker. The English form of snooker is played on the same table and with the same size balls used for English billiards. The game is played with 22 balls, made up of one white ball (the cue ball), 15 red balls, and six numbered coloured balls including one yellow, 2; one green, 3; one brown, 4; one blue, 5; one pink, 6; one black, valued at 7 points. The American form of snooker uses the same 22 balls, having the same values.

The games are played in basically the same fashion. The player must first pocket a red ball and then try to pocket any colour he may choose, scoring the value of the ball he has pocketed. He then alternately pockets red and coloured balls. Each red ball when pocketed remains in the pocket, while the colours when pocketed, as long as any reds remain on the table, are placed on their respective spots. Play continues until only the six colours remain on the table. Finally, the six coloured balls must be pocketed in the order of their values beginning with the yellow 2. When the last ball is pocketed the game is ended. This pocketing of the 21 object balls constitutes a frame, and a match consists of any selected number of frames. When a player cannot hit the ball that the rules require him to play at because of obstruction by another ball or balls, he is said to be snookered and loses his turn. This situation gives the game its name.

Pocket billiards, or pool. While the game of pocket billiards, or pool, is virtually unknown in the British countries or in Europe, it is by far the most popular of all the billiard games, at least in terms of the number of its players. In pocket billiards, 15 numbered object balls are used in addition to one white ball. Points are scored by driving the numbered object balls into any of the six pockets. The balls numbered from 1 to 8 are in solid colours, those from 9 to 15 are striped.

Many varieties of the game are possible, the most popular probably being that known as rotation or "Chicago," in which the object is to pocket the balls in rotation, starting with the lowest number. The numbers of the balls are added up to determine the winner of the game. Other well-known games of pocket billiards are fifteen ball, eight ball, baseball, golf, and 14.1 continuous, which is also known as the championship game. The names that have been coined for these and other variants are almost limitless. Of special interest among the pocket games is 14.1 continuous pocket billiards, the object of which is to make a continuous run of 14 balls in one turn, leaving the 15th ball on the table. The 14 pocketed balls are then racked again. The player then attempts to pocket the 15th ball and break the rack balls in order to continue the run. In championship play the first player to score 150 points wins the game. The skill required to pocket 14 balls without an error is very high indeed and this is a game only for highly skilled players.

Rotation and other pool games

Variant games. There are a number of variations of the game of pocket billiards, most of these being played on smaller size tables. One of the more popular forms is referred to as bumper pool. In the bumper pool table there are generally three holes, one in the middle of either end, but not set back into the rail as the normal pocket is, and a third in the centre of the table. In these tables there are a number of cushioned pegs or bumpers which are so placed on the playing surface that the player must bank off the cushion in order to put the ball in the hole.

Bumper pool and home billiards

Other variations of the regular billiard games noted above might be referred to as home billiard games as reflected by the great popularity of small billiard tables, which have become increasingly acceptable in the home due to the lack of space and the high cost of the larger tables. These smaller tables have been sold in great numbers. The games are played in the same fashion as the regular billiard games noted above but with the size of the tables reduced down to as little as 2.5 by five feet (0.8 by 1.5 metres). The balls and cues have also been reduced in size. The rules of the games, however, are the same as noted previously.

Prominent players. The first outstanding player in English billiards was Edward (better known as Jonathan) Kentfield, who held the title of champion for 24 years from 1825, and his influence on the accessories of the game probably was greater than that of any other man. His part in bringing about the improvement in tables, cushions, balls, and cues, was supreme. In 1849 John

By courtesy of Billiard Congress of America

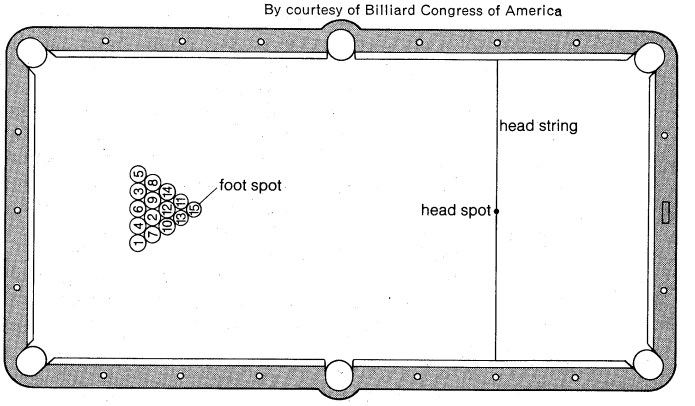


Figure 3: Plan of pocket billiard table.

Roberts, Sr., challenged Kentfield, but the latter, being concerned about his advanced age, refused to settle the issue by a match, and so Roberts was acknowledged champion. He in turn defeated all rivals until 1870, when, in the first official championship, he was beaten by a younger player, William Cook. Two months later, Cook lost the champion's title to John Roberts, Jr., who dominated the game in every respect for the rest of the 19th century. In more recent years, two of the best known names in professional English billiards have been Walter Lindrum of Australia and Joe Davis, who was also outstanding in snooker.

Famous American players in pocket billiards include Alfredo DeOro and Thomas Hueston, who held sway around the turn of the 20th century, with Ralph Greenleaf dominating in the 1920s. Willie Mosconi won the first of his many world titles in 1941 and is the holder of virtually every pocket billiard record. Luther Lassiter and Arthur Cranfield, Jr., dominated in the 1960s.

Governing bodies. The world governing body of English billiards and of the English form of snooker is the Billiards Association and Control Council. Formed in 1919, it represents an amalgamation of the former Billiards Association, established in 1885, and the Billiards Control Club, established in 1908. It frames the conditions and rules of both games and publishes the official organ, *Billiards and Snooker*.

The principal governing body of the American pocket billiards games, including the American form of snooker, is the Billiard Congress of America, which has been the ruling body since 1948. The congress maintains the rules of the games and sanctions championship tournaments including the U.S. Open Pocket Billiards Championship, regarded as the world championship.

For lists of world champions and records, see the *Ready Reference Index*.

BIBLIOGRAPHY. BILLIARD CONGRESS OF AMERICA, *Official Rules and Record Book* (1970); and BILLIARDS ASSOCIATION AND CONTROL COUNCIL, *Handbook and Rules* (1970–71), the official rule books for all pocket and carom billiard games, and for English billiards and snooker games, respectively; the most comprehensive and precise accounts of the various types of games, equipment, terms, champions, and records; CLIVE COTTINGHAM, JR., *The Game of Billiards* (1964), an interesting general account of pocket and carom billiard games; JOE DAVIS, *How I Play Snooker* (1949), a famous snooker player's approach to the game; WILLIE MOSCONI, *Pocket Billiards* (1971) and *Winning Pocket Billiards* (1965); and GEORGE SULLIVAN and IRVING CRANE, *The Young Sportsman's Guide to Pocket Billiards* (1964), books by two well-known champions giving their approach to the game for both beginners and more advanced players.

(B.E.C.)

Biochemistry

Biochemistry is the study of substances found in living organisms and of the changes they undergo during development and life of the organism. It deals with the chemistry of life, and as such it draws on the techniques of analytical, organic, and physical chemistry, as well as those of physiologists concerned with the molecular basis of vital processes. All chemical changes within the organism—either the degradation of substances, generally to gain necessary energy, or the buildup of complex molecules necessary for life processes—are collectively termed metabolism. These chemical changes depend on the action of organic catalysts known as enzymes, and enzymes, in turn, depend for their existence on the genetic apparatus of the cell. It is not surprising, therefore, that biochemistry enters into the investigation of chemical changes in disease, drug action, and other aspects of medicine, as well as in nutrition, genetics, and agriculture.

The term biochemistry is synonymous with two somewhat older terms: physiological chemistry and biological chemistry. Those aspects of biochemistry that deal with the chemistry and function of very large molecules (*e.g.*, proteins and nucleic acids) are often grouped under the term molecular biology. Biochemistry is a young science, having been known under that term only since about 1900. Its origins, however, can be traced much further back; its early history is part of the early history of both physiology and chemistry.

HISTORICAL ASPECTS

The particularly significant past events in biochemistry have been concerned with placing biological phenomena on firm chemical foundations.

Prelude to biochemistry

Before chemistry could contribute adequately to medicine and agriculture, however, it had to free itself from immediate practical demands in order to become a pure science. This happened in the period from about 1650 to 1780, starting with the work of Robert Boyle and culminating in that of Antoine-Laurent Lavoisier, the father of modern chemistry. Boyle questioned the basis of the chemical theory of his day and taught that the proper object of chemistry was to determine the composition of substances. His contemporary John Mayow observed the fundamental analogy between the respiration of an animal and the burning, or oxidation, of organic matter in air. Then, when Lavoisier carried out his fundamental studies on chemical oxidation, grasping the true nature of the process, he also showed, quantitatively, the similarity between chemical oxidation and the respiratory process. Photosynthesis was another biological phenomenon that occupied the attention of the chemists of the late 18th century. The demonstration, through the combined work of Joseph Priestley, Jan Ingenhousz, and Jean Senebier, that photosynthesis is essentially the reverse of respiration was a milestone in the development of biochemical thought.

In spite of these early fundamental discoveries, rapid progress in biochemistry had to wait upon the development of structural organic chemistry, one of the great achievements of 19th-century science. A living organism contains many thousands of different chemical compounds. The elucidation of the chemical transformations undergone by these compounds within the living cell is a central problem of biochemistry. Clearly, the determination of the molecular structure of the organic substances present in living cells had to precede the study of the cellular mechanisms whereby these substances are synthesized and degraded.

There are few sharp boundaries in science, and the boundaries between organic and physical chemistry, on the one hand, and biochemistry, on the other, have always shown much overlap. Biochemistry has borrowed the methods and theories of organic and physical chemistry and applied them to physiological problems. Progress in this path was at first impeded by a stubborn misconception in scientific thinking—the error of supposing that the transformations undergone by matter in the living organism are not subject to the chemical and physical laws that apply to inanimate substances and that consequently these "vital" phenomena cannot be described in ordinary chemical or physical terms. Such an attitude was taken by the vitalists, who maintained that natural products formed by living organisms could never be synthesized by ordinary chemical means. The first laboratory synthesis of an organic compound, urea, by Friedrich Wöhler in 1828, was a blow to the vitalists but not a decisive one. They retreated to new lines of defense, arguing that urea was only an excretory substance—a product of breakdown and not of synthesis. The success of the organic chemists in synthesizing many natural products forced further retreats of the vitalists. It is axiomatic in modern biochemistry that the chemical laws that apply to inanimate materials are equally valid within the living cell.

At the same time that progress was being impeded by a misplaced kind of reverence for living phenomena, the practical needs of man operated to spur the progress of the new science. As organic and physical chemistry erected an imposing body of theory in the 19th century, the needs of the physician, the pharmacist, and the agriculturalist provided an ever present stimulus for the application of the new discoveries of chemistry to various urgent practical problems.

Two outstanding figures of the 19th century, a German, Justus von Liebig, and a Frenchman, Louis Pasteur, were

The contributions of Liebig and Pasteur

particularly responsible for dramatizing the successful application of chemistry to the study of biology. Liebig studied chemistry in Paris and carried back to Germany the inspiration gained by contact with the former students and colleagues of Lavoisier. He established at Giessen a great teaching and research laboratory, one of the first of its kind, which drew students from all over Europe.

Besides putting the study of organic chemistry on a firm basis, Liebig engaged in extensive literary activity, attracting the attention of all scientists to organic chemistry and popularizing it for the layman as well. His classic works, published in the 1840s, had a profound influence on contemporary thought. Liebig described the great chemical cycles in nature. He pointed out that animals would disappear from the face of the Earth if it were not for the photosynthesizing plants, since animals require for their nutrition the complex organic compounds that can be synthesized only by plants. The animal excretions and the animal body after death are also converted by a process of decay to simple products that can be re-utilized only by plants.

In contrast with animals, green plants require for their growth only carbon dioxide, water, mineral salts, and sunlight. The minerals must be obtained from the soil, and the fertility of the soil depends on its ability to furnish the plants with these essential nutrients. But the soil is depleted of these materials by the removal of successive crops; hence the need for fertilizers. Liebig pointed out that chemical analysis of plants could serve as a guide to the substances that should be present in fertilizers. Agricultural chemistry as an applied science was thus born.

In his analysis of fermentation, putrefaction, and infectious disease, Liebig was less fortunate. He admitted the similarity of these phenomena but refused to admit that living organisms might function as the causative agents. It remained for Pasteur to clarify that matter. In the 1860s Pasteur proved that various yeasts and bacteria were responsible for "ferments," substances that caused fermentation and, in some cases, disease. He also demonstrated the usefulness of chemical methods in studying these tiny organisms and was the founder of what came to be called bacteriology.

Later, in 1877, Pasteur's ferments were designated as enzymes, and, in 1897, the German chemist E. Buchner clearly showed that fermentation could occur in a press juice of yeast, devoid of living cells. Thus a life process of cells was reduced by analysis to a nonliving system of enzymes. The chemical nature of enzymes remained obscure until 1926, when the first pure crystalline enzyme (urease) was isolated. This enzyme and many others subsequently isolated proved to be proteins, which had already been recognized as high-molecular-weight chains of subunits called amino acids.

The mystery of how minute amounts of dietary substances known as the vitamins prevent diseases such as beriberi, scurvy, and pellagra became clear in 1935, when riboflavin (vitamin B_2) was found to be an integral part of an enzyme. Subsequent work has substantiated the concept that many vitamins are essential in the chemical reactions of the cell by virtue of their role in enzymes.

In 1929 the substance adenosine triphosphate (ATP) was isolated from muscle. Subsequent work demonstrated that the production of ATP was associated with respiratory (oxidative) processes in the cell. In 1940 F.A. Lipmann proposed that ATP is the common form of energy exchange in many cells, a concept now thoroughly documented. ATP has been shown also to be a primary energy source for muscular contraction.

Isotopes as tools

The use of radioactive isotopes of chemical elements to trace the pathway of substances in the animal body was initiated in 1935 by two U.S. chemists: R. Schoenheimer and D. Rittenberg. That technique provided one of the single most important tools for investigating the complex chemical changes that occur in life processes. At about the same time, other workers localized the sites of metabolic reactions by ingenious technical advances in the studies of organs, tissue slices, cell mixtures, individual cells, and, finally, individual cell constituents, such as nuclei, mitochondria, ribosomes, lysosomes, and membranes.

In 1869 a substance was isolated from the nuclei of pus cells and was called nucleic acid, which later proved to be deoxyribonucleic acid (DNA), but it was not until 1944 that the significance of DNA as genetic material was revealed, when bacterial DNA was shown to change the genetic matter of other bacterial cells. Within a decade of that discovery, the double helix structure of DNA was proposed by the Nobel Prize winners J.D. Watson and F.H.C. Crick, providing a firm basis for understanding how DNA is involved in cell division and in maintaining genetic characteristics.

Advances have continued since that time, with such landmark events as the first chemical synthesis of a protein, the detailed mapping of the arrangement of atoms in some enzymes, and the elucidation of intricate mechanisms of metabolic regulation, including the molecular action of hormones.

SCOPE

A description of life at the molecular level includes a description of all the complexly interrelated chemical changes that occur within the cell—*i.e.*, the processes known as intermediary metabolism. The processes of growth, reproduction, and heredity, also subjects of the biochemist's curiosity, are intimately related to intermediary metabolism and cannot be understood independently of it. The properties and capacities exhibited by a complex multicellular organism can be reduced to the properties of the individual cells of that organism, and the behaviour of each individual cell can be understood in terms of its chemical structure and the chemical changes occurring within that cell. When all the chemical changes within a cell are completely described and understood, man will have achieved as complete an understanding of life as can be achieved by the intellect alone. Living processes are sufficiently complex, however, to guarantee the biochemist enough unsolved problems to last into the unforeseeable future.

Chemical composition of living matter. Every living cell contains, in addition to water and salts or minerals, a large number of organic compounds, substances composed of carbon combined with varying amounts of hydrogen and usually also of oxygen. Nitrogen, phosphorus, and sulfur are likewise common constituents. In general, the bulk of the organic matter of a cell may be classified as (1) protein, (2) carbohydrate, and (3) fat, or lipid. Nucleic acids and various other organic derivatives are also important constituents. Each class contains a great diversity of individual compounds. Many substances that cannot be classified in any of the above categories also occur, though usually not in large amounts.

Proteins are fundamental to life, not only as structural elements (*e.g.*, collagen) and to provide defense (as antibodies) against invading destructive forces but also because the essential biocatalysts are proteins. The chemistry of proteins is based on the researches of the German chemist Emil Fischer, whose work from 1882 demonstrated that proteins are very large molecules, or polymers, built up of about 24 similar building blocks, known as amino acids. Proteins may vary in size from small—insulin with a molecular weight of 5,700 (based on the weight of a hydrogen atom as 1)—to very large—molecules with molecular weights of more than 1,000,000. The first complete amino acid sequence was determined for the insulin molecule in the 1950s. By 1963 the chain of amino acids in the protein enzyme ribonuclease (molecular weight 12,700) had also been determined, aided by the powerful physical techniques of X-ray-diffraction analysis. In the 1960s, Nobel Prize winners J.C. Kendrew and M.F. Perutz, utilizing X-ray studies, constructed detailed atomic models of the proteins hemoglobin and myoglobin (the respiratory pigment in muscle), which were later confirmed by sophisticated chemical studies. The abiding interest of biochemists in the structure of proteins rests on the fact that the arrangement of chemical groups in space yields important clues regarding the biological activity of molecules.

Protein size

Carbohydrates include such substances as sugars, starch, and cellulose. The second quarter of the 20th century witnessed a striking advance in the knowledge of how living cells handle small molecules, including carbohydrates. The metabolism of carbohydrates became clarified during this period, and elaborate pathways of carbohydrate breakdown and subsequent storage and utilization were gradually outlined in terms of cycles (*e.g.*, the Embden–Meyerhof glycolytic cycle and the Krebs cycle). The involvement of carbohydrate in respiration and muscle contraction was well worked out by the 1950s. Refinements of the schemes continue.

Fats, or lipids, constitute a heterogeneous group of organic chemicals that can be extracted from biological material by nonpolar solvents such as ethanol, ether, and benzene. The classic work concerning the formation of body fat from carbohydrates was accomplished during the early 1850s. Those studies, and later confirmatory evidence, have shown that the conversion of carbohydrate to fat occurs continuously in the body. The liver is the main site of fat metabolism. Fat absorption in the intestine, studied as early as the 1930s, still is under investigation by biochemists. The hormonal control of fat absorption is known to depend upon a combination action of cortical hormones of the adrenal glands and of salt levels in the body. Abnormalities of fat metabolism, which result in disorders such as obesity and rare clinical conditions, are the subject of much biochemical research. Equally interesting to biochemists is the association between high levels of fat in the blood and the occurrence of arteriosclerosis ("hardening" of the arteries).

Significance of nucleic acids

Nucleic acids are large, complex compounds of very high molecular weight present in the cells of all organisms and in viruses. They are of great importance in the synthesis of proteins and in the transmission of hereditary information from one generation to the next. Originally discovered as constituents of cell nuclei (hence their name), it was assumed for many years after their isolation in 1869 that they were found nowhere else. This assumption was not challenged seriously until the 1940s, when it was determined that two kinds of nucleic acid exist: deoxyribonucleic acid (DNA), in the nuclei of all cells and in some viruses; and ribonucleic acid (RNA), in the cytoplasm of all cells and in most viruses.

The profound biological significance of nucleic acids came gradually to light during the 1940s and 1950s. Attention turned to the mechanism by which protein synthesis and genetic transmission was controlled by nucleic acids (see below *Genes*). During the 1960s, experiments were aimed at refinements of the genetic code. Promising attempts were made during the late 1960s and early 1970s to accomplish duplication of the molecules of nucleic acids outside the cell—*i.e.*, in the laboratory. Several preliminary steps in the so-called creation of life in the test tube were qualified successes. Experiments to alter the DNA content of cells were also under way. By such means, defective genes might someday be corrected.

Nutrition. Biochemists have long been interested in the chemical composition of the food of animals. All animals require organic material in their diet, in addition to water and minerals. This organic matter must be sufficient in quantity to satisfy the caloric, or energy, requirements of the animals. Within certain limits, carbohydrate, fat, and protein may be used interchangeably for this purpose. In addition, however, animals have nutritional requirements for specific organic compounds. Certain essential fatty acids, about ten different amino acids (the so-called essential amino acids), and vitamins are required by many higher animals. The nutritional requirements of various species are similar but not necessarily identical; thus man and the guinea pig require vitamin C, or ascorbic acid, whereas the rat does not.

That plants differ from animals in requiring no preformed organic material was appreciated soon after the plant studies of the late 1700s. The ability of green plants to make all their cellular material from simple substances —carbon dioxide, water, salts, and a source of nitrogen such as ammonia or nitrate—was termed photosynthesis. As the name implies, light is required as an energy source, and it is generally furnished by sunlight. The process of photosynthesis itself is primarily concerned with the manufacture of carbohydrate, from which fat can be made by animals that eat plant carbohydrates. Protein can also be formed from carbohydrate, provided ammonia is furnished (see PHOTOSYNTHESIS).

Similarity of chemical changes in living things

In spite of the large apparent differences in nutritional requirements of plants and animals, the patterns of chemical change within the cell are the same. The plant manufactures all the materials it needs, but these materials are essentially similar to those that the animal cell uses and are often handled in the same way once they are formed. Plants could not furnish animals with their nutritional requirements if the cellular constituents in the two forms were not basically similar (see NUTRITION).

Digestion. The organic food of animals, including man, consists in part of large molecules. In the digestive tracts of higher animals, these molecules are hydrolyzed, or broken down, to their component building blocks. Proteins are converted to mixtures of amino acids, and polysaccharides are converted to monosaccharides. In general, all living forms use the same small molecules, but many of the large complex molecules are different in each species. An animal, therefore, cannot use the protein of a plant or of another animal directly but must first break it down to amino acids and then recombine the amino acids into its own characteristic proteins. The hydrolysis of food material is necessary also to convert solid material into soluble substances suitable for absorption. The liquefaction of stomach contents aroused the early interest of observers, long before the birth of modern chemistry, and the hydrolytic enzymes secreted into the digestive tract were among the first enzymes to be studied in detail. Pepsin and trypsin, the proteolytic enzymes of the gastric and pancreatic juice, respectively, continue to be intensively investigated.

The products of enzymatic action on the food of an animal are absorbed through the walls of the intestines and distributed to the body by blood and lymph. In organisms without digestive tracts, substances must also be absorbed in some way from the environment. In some instances simple diffusion appears to be sufficient to explain the transfer of a substance across a cell membrane. In other cases, however (*e.g.*, in the case of the transfer of glucose from the lumen of the intestine to the blood), transfer occurs against a concentration gradient. That is, the glucose may move from a place of lower concentration to a place of higher concentration.

In the case of the secretion of hydrochloric acid into gastric juice, it has been shown that active secretion is dependent on an adequate oxygen supply (*i.e.*, on the respiratory metabolism of the tissue), and the same holds for absorption of salts by plant roots. The energy released during the tissue oxidation must be harnessed in some way to provide the energy necessary for the absorption or secretion. This harnessing is achieved by a special chemical coupling system. The elucidation of the nature of such coupling systems is an objective of the biochemist (see DIGESTION AND DIGESTIVE SYSTEMS).

Extent of knowledge about blood

Blood. One of the animal tissues that has always excited special curiosity is blood. Blood has been investigated intensively from the early days of biochemistry, and its chemical composition is known with greater accuracy and in more detail than that of any other tissue in the body. The physician takes blood samples to determine such things as the sugar content, the urea content, or the inorganic-ion composition of the blood, since these show characteristic changes in disease (see BLOOD DISEASES).

The blood pigment hemoglobin has been intensively studied. Hemoglobin is confined within the blood corpuscles and carries oxygen from the lungs to the tissues. It combines with oxygen in the lungs, where the oxygen concentration is high, and releases the oxygen in the tissues, where the oxygen concentration is low. The hemoglobins of higher animals are related but not identical. In invertebrates, other pigments may take the place and function of hemoglogin. The comparative study of these

compounds constitutes a fascinating chapter in biochemical investigation (see COLORATION, BIOLOGICAL).

The proteins of blood plasma also have been extensively investigated. The gamma-globulin fraction of the plasma proteins contains the antibodies of the blood and is of practical value as an immunizing agent. An animal develops resistance to disease largely by antibody production. Antibodies are proteins with the ability to combine with an antigen (*i.e.*, an agent that induces their formation). When this agent is a component of a disease-causing bacterium, the antibody can protect an organism from infection by that bacterium. The chemical study of antigens and antibodies and their interrelationship is known as immunochemistry (see IMMUNITY).

Metabolism and the study of hormones. The cell is the site of a constant, complex, and orderly set of chemical changes collectively called metabolism. Metabolism is associated with a release of heat. The heat released is the same as that obtained if the same chemical change is brought about outside the living organism. This confirms the fact that the laws of thermodynamics apply to living systems just as they apply to the inanimate world. The pattern of chemical change in a living cell, however, is distinctive and different from anything encountered in nonliving systems. This difference does not mean that any chemical laws are invalidated. It resides rather in the extraordinary complexity of the interrelations of cellular reactions (see METABOLISM).

Hormones, which may be regarded as regulators of metabolism, are investigated at three levels, to determine (1) their physiological effects, (2) their chemical structure, and (3) the chemical mechanisms whereby they operate. The study of the physiological effects of hormones is properly regarded as the province of the physiologist. Such investigations obviously had to precede the more analytical chemical studies. The chemical structures of thyroxine and adrenaline are known. The chemistry of the sex and adrenal hormones, which are steroids, has also been thoroughly investigated. The hormones of the pancreas, insulin and glucagon, and the hormones of the hypophysis (pituitary gland) are peptides (*i.e.*, compounds composed of chains of amino acids). By the mid-1970s, the structures of most of these hormones had been determined or were close to elucidation. The chemical structures of the plant hormones, auxin and gibberellic acid, which act as growth-controlling agents in plants, were also known.

Status of hormonal study

The first and second phases of the hormone problem thus have been well, though not completely, explored, but the third phase is still in its infancy. It seems likely that different hormones exert their effects in different ways. Some may act by affecting the permeability of membranes; others appear to control the synthesis of certain enzymes. Evidently some hormones also control the activity of certain genes (see HORMONE).

Genes. Genetic studies have shown that the hereditary characteristics of a species are maintained and transmitted by the self-duplicating units known as genes, which are composed of nucleic acids and located in the chromosomes of the nucleus. One of the most fascinating chapters in the history of the biological sciences contains the story of the elucidation, in the mid-20th century, of the chemical structure of the genes, their mode of self-duplication, and the manner in which the deoxyribonucleic acid (DNA) of the nucleus causes the synthesis of ribonucleic acid (RNA), which, in turn causes the synthesis of protein. Thus, the capacity of a protein to behave as an enzyme is determined by the chemical constitution of the gene (DNA) that directs the synthesis of the protein. The relationship of genes to enzymes has been demonstrated in several ways. The first successful experiments, devised by the Nobel Prize winners George W. Beadle and Edward L. Tatum, involved the bread mold *Neurospora crassa;* the two men were able to collect a variety of strains that differed from the parent strain in nutritional requirements. Such strains had undergone a mutation (change) in the genetic makeup of the parent strain. The mutant strains required a particular amino acid not required for growth by the parent strain. It was then shown that such a mutant had lost an enzyme essential for the synthesis of the amino acid in question. The subsequent development of techniques for the isolation of mutants with specific nutritional requirements led to a special procedure for studying intermediary metabolism. (see GENE.)

Evolution and origin of life. The exploration of space beginning in the mid-20th century intensified speculation about the possibility of life on other planets. At the same time, man was beginning to understand some of the intimate chemical mechanisms used for the transmission of hereditary characteristics. It was possible, by studying protein structure in different species, to see how the amino acid sequences of functional proteins (*e.g.*, hemoglobin and cytochrome) have been altered during phylogeny (the development of species). It was natural, therefore, that biochemists in the 1970s should look upon the problem of the origin of life as a practical one. The synthesis of a living cell from inanimate material was not regarded as an impossible task for the future (see LIFE; EVOLUTION).

The coming of space biochemistry

METHODOLOGY AND INSTRUMENTATION

Like other sciences, biochemistry aims at quantifying, or measuring, results, sometimes with sophisticated instrumentation. The earliest approach to a study of the events in a living organism was an analysis of the materials entering an organism (foods, oxygen) and those leaving (excretion products, carbon dioxide). This is still the basis of so-called balance experiments conducted on animals, in which, for example, both foods and excreta are thoroughly analyzed. For this purpose many chemical methods involving specific colour reactions have been developed, requiring spectrum-analyzing instruments (spectrophotometers) for quantitative measurement. Gasometric techniques are those commonly used for measurements of oxygen and carbon dioxide, yielding respiratory quotients (the ratio of carbon dioxide to oxygen). Somewhat more detail has been gained by determining the quantities of substances entering and leaving a given organ and also by incubating slices of a tissue in a physiological medium outside the body and analyzing the changes that occur in the medium. Because these techniques yield an overall picture of metabolic capacities, it became necessary to disrupt cellular structure (homogenization) and to isolate the individual parts of the cell —nuclei, mitochondria, lysosomes, ribosomes, membranes—and finally the various enzymes and discrete chemical substances of the cell in an attempt to understand the chemistry of life more fully.

Centrifugation and electrophoresis. An important tool in biochemical research is the centrifuge, which through rapid spinning imposes high centrifugal forces on suspended particles, or even molecules in solution, and causes separations of such matter on the basis of differences in weight. Thus, red cells may be separated from plasma of blood, nuclei from mitochondria in cell homogenates, and one protein from another in complex mixtures. Proteins are separated by ultracentrifugation—very high speed spinning; with appropriate photography of the protein layers as they form in the centrifugal field, it is possible to determine the molecular weights of proteins.

Exploiting electrical charges of molecules

Another property of biological molecules that has been exploited for separation and analysis is their electrical charge. Amino acids and proteins possess net positive or negative charges according to the acidity of the solution in which they are dissolved. In an electric field, such molecules adopt different rates of migration toward positively (anode) or negatively (cathode) charged poles and permit separation. Such separations can be effected in solutions or when the proteins saturate a stationary medium such as cellulose (filter paper), starch, or acrylamide gels. By appropriate colour reactions of the proteins and scanning of colour intensities, a number of proteins in a mixture may be measured. Separate proteins may be isolated and identified by electrophoresis, and the purity of a given protein may be determined. (Electrophoresis of human hemoglobin revealed the abnormal

hemoglobin in sickle-cell anemia, the first definitive example of a "molecular disease.")

Chromatography and isotopes. The different solubilities of substances in aqueous and organic solvents provide another basis for analysis. In its earlier form, a separation was conducted in complex apparatus by partition of substances in various solvents. A simplified form of the same principle evolved as "paper chromatography," in which small amounts of substances could be separated on filter paper and identified by appropriate colour reactions. In contrast to electrophoresis, this method has been applied to a wide variety of biological compounds and has contributed enormously to research in biochemistry.

The general principle has been extended from filter paper strips to columns of other relatively inert media, permitting larger scale separation and identification of closely related biological substances. Particularly noteworthy has been the separation of amino acids by chromatography in columns of ion-exchange resins, permitting the determination of exact amino acid composition of proteins. Following such determination, other techniques of organic chemistry have been used to elucidate the actual sequence of amino acids in complex proteins. Another technique of column chromatography is based on the relative rates of penetration of molecules into beads of a complex carbohydrate according to size of the molecules. Larger molecules are excluded relative to smaller molecules and emerge first from a column of such beads. This technique not only permits separation of biological substances but also provides estimates of molecular weights.

Perhaps the single most important technique in unravelling the complexities of metabolism has been the use of isotopes (heavy or radioactive elements) in labelling biological compounds and "tracing" their fate in metabolism. Measurement of the isotope-labelled compounds has required considerable technology in mass spectroscopy and radioactive detection devices.

A variety of other physical techniques, such as nuclear magnetic resonance, electron spin spectroscopy, circular dichroism, and X-ray crystallography, have become prominent tools in revealing the relation of chemical structure to biological function.

APPLIED BIOCHEMISTRY

Clinical biochemistry. An early objective in biochemistry was to provide analytical methods for the determination of various blood constituents because it was felt that abnormal levels might indicate the presence of metabolic diseases. The clinical chemistry laboratory now has become a major investigative arm of the physician in the diagnosis and treatment of disease and is an indispensable unit of every hospital. Some of the older analytical methods directed toward diagnosis of common diseases are still the most commonly used—for example, tests for determining the levels of blood glucose, in diabetes; urea, in kidney disease; uric acid, in gout; and bilirubin, in liver and gallbladder disease. With development of the knowledge of enzymes, determination of certain enzymes in blood plasma has assumed diagnostic value, such as alkaline phosphatase, in bone and liver disease; acid phosphatase, in prostatic cancer; amylase, in pancreatitis; and lactate dehydrogenase and transaminase, in cardiac infarct. Electrophoresis of plasma proteins is commonly employed to aid in the diagnosis of various liver diseases and forms of cancer. Both electrophoresis and ultracentrifugation of serum constituents (lipoproteins) are used increasingly in the diagnosis and examination of therapy of atherosclerosis and heart disease. Many specialized and sophisticated methods have been introduced, and machines have been developed for the simultaneous automated analysis of many different blood constituents in order to cope with increasing medical needs.

Diagnostic tests

Food and pharmaceutical biochemistry. Analytical biochemical methods have also been applied in the food industry to develop crops superior in nutritive value and capable of retaining nutrients during the processing and preservation of food. Research in this area is directed particularly to preserving vitamins as well as colour and taste, all of which may suffer loss if oxidative enzymes remain in the preserved food. Tests for enzymes are used for monitoring various stages in food processing.

Biochemical techniques have been fundamental in the development of new drugs. The testing of potentially useful drugs includes studies on experimental animals and man to observe the desired effects and also to detect possible toxic manifestations; such studies depend heavily on many of the clinical biochemistry techniques already described. Although many of the commonly used drugs have been developed on a rather empirical (trial-and-error) basis, an increasing number of therapeutic agents have been designed specifically as enzyme inhibitors to interfere with the metabolism of a host or invasive agent. Biochemical advances in the knowledge of the action of natural hormones and antibiotics promise to aid further in the development of specific pharmaceuticals.

BIOCHEMICAL SOCIETIES AND PUBLICATIONS

In 1871 R. Maly began issuing a chemistry annual, the *Jahresbericht über die Fortschritte der Tierchemie*, which appeared until 1919. With the assistance of biochemists in other countries, Maly abstracted from various scientific journals published all over the world all those papers that had primarily a biochemical content. The first scientific journal devoted exclusively to the subject of biochemistry was the *Zeitschrift für Physiologische Chemie*, founded by F. Hoppe-Seyler in 1877. The U.S. *Journal of Biological Chemistry* was founded in 1905, and the British *Biochemical Journal* in 1906. In the same year another German journal, *Biochemische Zeitschrift*, also was founded. The French *Bulletin de la Société de Chimie Biologique* first appeared in 1914, and the *Journal of Biochemistry* (Japan) appeared in 1922. With temporary interruptions during World Wars I and II, there has been a steady increase in number of new journals containing biochemical papers and in the volume of material published by the older journals as well.

Paralleling the growth of journals was the growth of scientific societies. The American Society of Biological Chemists was founded in 1906 by a group of members of the Physiological Society, spurred, perhaps, by the formation in 1905 of a biochemical section of the American Chemical Society. The British Biochemical Society, founded in 1911, was also an offshoot of the Physiological Society. These two biochemical societies own and control the *Journal of Biological Chemistry* and the *Biochemical Journal* (London), respectively. These and other societies from some 30 countries are bound together by the International Union of Biochemistry, which has the important function of standardizing journals and nomenclature, as well as arranging congresses and symposia for international exchange of research progress.

BIBLIOGRAPHY. JOSEPH NEEDHAM (ed.), *The Chemistry of Life: Eight Lectures on the History of Biochemistry* (1970), provides brief development of the important areas of photosynthesis, enzymes, microbiology, neurology, hormones, vitamins, and other topics. Two selected biographical treatments that couple human interest with development of biochemical areas are FRITZ A. LIPMANN, *Wanderings of a Biochemist* (1971); and DAVID KEILIN, *The History of Cell Respiration and Cytochrome* (1966). Two monographs that describe areas in the forefront of biochemistry are JAMES D. WATSON, *Molecular Biology of the Gene*, 2nd ed. (1970); and THOMAS H. JUKES, *Molecules and Evolution* (1966). Two recent and attractively illustrated textbooks are ALBERT L. LEHNINGER, *Biochemistry: The Molecular Basis of Cell Structure and Function* (1970); and ROBERT W. MCGILVERY, *Biochemistry: A Functional Approach* (1970). Other standard and somewhat more comprehensive texts are ABRAHAM WHITE, PHILIP HANDLER, and EMIL L. SMITH, *Principles of Biochemistry*, 5th ed. (1973); and EDWARD S. WEST *et al.*, *Textbook of Biochemistry*, 4th ed. (1966). A useful set of paperbacks is *Essays in Biochemistry*, published for the Biochemical Society (1965–). More comprehensive survey treatises are MARCEL FLORKIN and ELMER H. STOTZ (eds.), *Comprehensive Biochemistry* (1962–); and MARCEL FLORKIN and HOWARD S. MASON (eds.), *Comparative Biochemistry*, 7 vol. (1960–64). Current publications include *Annual Review of Biochemistry* (1932–), *Advances in Enzymology* (1941–), and many journals containing reports of original research.

(E.H.St./B.V.)

Bioelectricity

Bioelectricity refers to the generation or action of electric currents or voltage in biological processes. Most studies are involved with nerve or muscle tissue; with organs such as heart, brain, eye, ear, stomach, and some glands; with electric organs in some fish; and with potentials associated with damaged tissue. Free electrons that are developed (*e.g.*, by friction) in living things tend to migrate to the ground, in some cases escaping as a spark, perhaps to a door handle or other grounded object. Voltages so generated are not considered bioelectric potentials but manifestations of electrostatics (see ELECTRICITY). Bioelectric phenomena are mainly fast signalling in nerves or triggering of physical processes in muscles or glands. Possibly because fairly efficient electrochemical systems evolved early, there is some similarity among the nerves, muscles, and glands of all organisms possessing them. Bioelectric phenomena are found in all living organisms, particularly where rapid responses are needed. Electric activity in living tissue is a cellular phenomenon, dependent upon the cell membrane. The membrane acts like a capacitor, storing energy as electrically charged ions on opposite sides of the membrane. The stored energy is available for rapid utilization and also stabilizes the membrane system so that small disturbances do not activate it. For a detailed discussion of membranes, see MEMBRANE, BIOLOGICAL.

Range and types of biopotentials. Cells capable of electric activity show a resting potential in which their interiors are negative by about 0.1 volt (sometimes less) compared to the outside of the cell. When the cell is activated, the resting potential may reverse suddenly in sign; as a result, the outside of the cell becomes negative and the inside positive. This condition lasts for a short time, after which the cell returns to its original resting state. This sequence, called depolarization and repolarization, is accompanied by a flow of substantial current through the active cell membrane, so that a "dipole current source" exists transiently. Small currents flow from this source through the aqueous medium containing the cell and are detectable at considerable distances from it, though at very low levels. These currents, originating in active membrane, are functionally significant very close to their site of origin but must be considered incidental at any distance from it. In the electric fish, however, adaptations have occurred so that this otherwise incidental electric current is actually utilized. In some species the external current apparently is used for sensing purposes, while in others it is used to stun or kill prey. In both cases adaptations have evolved that enable voltages from many cells to add up in series, thus assuring that the specialized functions can be performed. Bioelectric potentials detected at some distances from the cells originating them may be as small as the 20 or 30 microvolts of some components of the human electroencephalogram or the millivolt of the human electrocardiogram, or they may be as large as the 600 to 1,000 volt shocks developed by the electric eel.

Generation of current

In addition to the potentials originating in nerve or muscle cells, relatively steady or slowly varying potentials (often called dc) are known. These dc potentials can occur in an area where cells have been damaged and ionized potassium is leaking from them (voltage level, up to 50 millivolts); when one part of the brain is compared with another part (voltage level, to one millivolt); when different areas of the skin are compared (voltage level, up to 10 millivolts); in the vicinity of the stomach and other organs when activity is present (voltage level, up to several millivolts); within pockets (*e.g.*, follicles in the thyroid gland) in glands that are active (voltage level, up to 60 millivolts); and in special structures such as the endotic space in the inner ear (voltage level in the ear, 80 millivolts). In some of these cases, theories have been advanced relating existence of the electric charge to physiological function.

Activity potentials involving representative types of nerve or muscle cells are identified as follows: nerve action potentials, muscle action potentials, muscle endplate potentials, synaptic potentials of nerve cells, action potentials of the ventricular muscle of the heart, action potentials of the auricle cells of the heart, action potentials of Purkinje fibres of the heart, generator potentials from sensory cells, and cochlear potentials recorded from the middle ear. Many of these phenomena are discussed in detail in the articles MUSCLE CONTRACTION and NERVE IMPULSE. Electric-organ potentials produced by certain fish are discussed in a later section of this article.

Clinical medicine makes routine use of the electrical effects originating in active cells of the heart (electrocardiogram), in active skeletal-muscle cells (electromyogram), in brain cells (electroencephalogram), and in the retina of the eye (electro-oculogram). See BIOMEDICAL INSTRUMENTATION for more information concerning application of these effects.

Factors required to establish a bioelectric potential. *Semipermeable membrane.* Atoms and molecules usually carry a net positive or negative charge and can be prevented from moving by one of several types of barriers. In this way a zone that contains barriers to the movement of charged atoms can possess a distinct electric charge. A barrier can be completely impervious to all ionic movement (*e.g.*, the walls of a glass beaker); however, this is not typical of biological material. Most barries are permeable to ions within a specific size range. Cellophane, a semipermeable membrane of this type, can be thought of as possessing pores of a certain size. In a biological system the membrane barrier is not so simple, and its ion-retaining capacity may be closely related to the mechanism responsible for actually accumulating ions. Ions carried through a membrane by complex chemical processes are accumulated faster within the confines of the membrane than they can leak away; in this way the membrane is an effective barrier. Diffusion of various ions through a selectively permeable membrane proceeds at different rates, resulting in a net positive or negative charge in a solution of basically neutral molecules.

Accumulation of potassium ions

In the nerve or muscle cell, potassium ions are accumulated inside the cell and sodium ions are removed from it. Cells contain in solution a complex substance called potassium–proteinate. Potassium holds a positive charge and proteinate a negative charge. The cell membrane (*i.e.*, a biological barrier), very slightly permeable to potassium, is completely impermeable to proteinate. Therefore, leakage of potassium, but not proteinate, occurs, and the interior of the cell is negatively charged. Eventually an equilibrium is attained, since the attractive forces of negatively charged proteinate prevent the escape of all the positively charged potassium ions, despite the diffusive force driving them. This is the basis for the Nernst equation, used to calculate the resting potential of a cell with fair accuracy if the ion concentrations (and therefore the diffusion forces) are known. It is desirable to include in such calculations all the ions involved; the Goldman equation attempts to do this (see MEMBRANE, BIOLOGICAL). In the Goldman equation, ions other than potassium (*e.g.*, sodium, chloride, calcium) are considered, along with their individual permeabilities or activities as determined by the system.

The secretion of a substance into the lumen of an organ or into the follicle of a gland possibly involves a net charge transport that may be fortuitous or have some definite function.

Energy. To carry a particular kind of ion into an area already containing a concentration of the same ion requires the expenditure of energy. Energy, transported in the blood as glucose, is transformed into bond energy in the energy-rich compound adenosine triphosphate, ATP (see METABOLISM). Since the combustion of glucose requires oxygen, the existence of bioelectric potentials is dependent on a supply of oxygen to tissues.

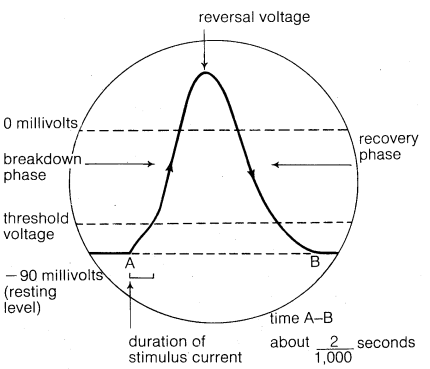
Occurrence of potential differences

Potentials of the type described above occur in resting internal potentials of nerve and muscle cells, in the potentials generated by frog skin, in the sweat-gland po-

tentials that give a voltage to the palms of the hands, in the potentials of the eye, and in various other potentials of glands and organs. It is possible also to detect steady or slowly varying potential differences between areas of the brain or between different parts of the whole body. Whole-body potentials, related in some way to the nervous system, may occur because nerve-fibre membranes are not polarized to the same degree along their entire length at any given time. This variable polarization results in weak, steady currents that flow in tissues outside the nerve fibres. It is possible that such electric fields have an important function; for example, limb regeneration in some lower animals cannot take place in the absence of these electric fields.

Specialized properties of excitable cells. A muscle is an aggregate of contractile cells each a few millimetres or a few centimetres long. A nerve cell consists of a bulbous cell body and a long filament that is sometimes continuous from the spinal cord to the extremity of a limb. The membrane of a nerve cell is similar to that of a muscle cell. Both membranes are "excitable." In an excitable cell the mechanism responsible for the resting potential is supplemented by an additional mechanism that transiently reverses the membrane potential and then returns it to normal. This transient voltage reversal travels as a wave along the length of the cell membrane. In nerve cells it conveys information; in muscle cells it triggers contraction.

Depolarization. The cell membrane in the resting polarized state is a poor conductor of electricity and allows little leakage of charge. Membrane integrity appears to depend upon a near-normal polarization. If a stimulus results in partial depolarization (*i.e.*, decrease in resting potential) of a membrane, the resting potential may be restored by the expenditure of metabolic energy. If the stimulus is sufficiently strong, however, membrane depolarization continues, a decrease in membrane resistance occurs, and charge leaks away faster than it can be restored. This loss of charge results in further decreases both in voltage and in membrane resistance. A threshold exists below which complete depolarization of the membrane occurs (see diagram). When a region on a



Oscilloscope tracing during depolarization of an excitable membrane following application of a stimulus (see text).

cell membrane undergoes the changes described, the depolarization is spread, or propagated, along the membrane by the following mechanism. The low resistance of the depolarized area provides a path for leakage of charge from the membrane next to the depolarized region. The leakage results in a decrease in membrane potential to the threshold voltage, depolarization occurs, and a travelling depolarization wave is generated.

Sodium equilibrium potential. Whereas the resting potential of the membrane depends on its permeability to potassium, depolarization is related to a transient increase in membrane permeability to sodium. The transient potential attained across the depolarized membrane is close to the value calculated for a sodium equilibrium potential using the Nernst equation. It has been possible to trace and separate the sodium and potassium effects and to obtain mathematical expressions describing them. The Nobel Prize winners A.L. Hodgkin and A.F. Huxley did this and obtained an expression for total membrane current in an area following stimulation. A nerve fibre is thought of as similar to a telegraph cable. Both have inner and outer conductors and a cylindrical insulating sheath. The mathematics of the propagation of electric waves in a cable in terms of R, L, G, and C (R, lengthways resistance; L, inductance per unit length; G, leakage; C, capacitance between inner and outer conductors per unit length) were worked out by Lord Kelvin many years ago. Hodgkin and Huxley introduced their expression for membrane current into the cable equation (making $L = O$) to obtain a differential equation that describes the propagation of impulses along a nerve. The equation is used to predict electrical changes with time and along the length of a nerve fibre. In deriving their equations Hodgkin and Huxley found it necessary to introduce factors m^3h (related to sodium conductance) and n^4 (related to potassium conductance) where m, n, and h varied between 0 and 1. A suggested interpretation was that these variables applied to some molecular "stepping stone" that had to disperse through the membrane before the ions themselves could pass. Many people have used Hodgkin and Huxley's techniques to study resting potentials and action currents in a variety of excitable membranes. Basic concepts of the equation have seldom been challenged, but no satisfactory explanation of m, n, and h has appeared.

Hodgkin–Huxley equation

Electrical changes induced in stimulated cells. The characteristic signal transmitted along nerve fibres consists of trains of impulses of the type described above. The membrane recovers from one impulse within a few milliseconds, so that impulse-repetition rates of many hundreds a second are possible, though not usual. Impulses travel as rapidly as 100 metres a second or as slowly as one metre a second in very small fibres. Sometimes a train of impulses begins with a fast repetition rate that slows and quickly fades. In a few modalities, impulse repetition is sustained continuously with a rate proportional to the parameter (*e.g.*, blood pressure) being signalled.

Stimulus. Impulse trains originate in special cells or in specialized endings of nerve fibres. A stimulus (*e.g.*, heat, stretching, a chemical substance) usually acts directly to modify the permeability of an area of membrane. The effect is to change either the internal cell potential or some other factor, so that normal resting conditions cannot be sustained; the membrane continues to fire and recover and fire again until the stimulus is removed or some metabolic readjustment "accommodates" the system to the stimulus. When a stimulus is acting on a specialized sensing cell, the permeability change and partial depolarization are detected as a "generator potential"; the amplitude of the generator potential often has a logarithmic relationship to stimulus strength, at least within a specific range. Impulse firing rate itself tends to be more or less proportional to the generator potential. Most sensing cells, however, have external protective systems (*e.g.*, pupil of the eye operating to reduce light levels, skin-blood-flow changes carrying away heat) so that no general law of proportionality can be expected between nerve firing rate and stimulus levels. In many modalities sensory signalling is geared to alerting the organism to changes in the environment rather than to signalling background levels of physical parameters. For a discussion of sensory receptors, see SENSORY RECEPTION.

Generator potential

Trains of nerve impulses generated in the brain and spinal cord are involved in computational activity of the brain and in actuating muscles. In these cases the stimulus comes from the activity of other nerve cells.

Effect. Nerve impulses are sent to muscles and to secretory organs such as the salivary glands. In the muscle itself the excitable membrane is coupled to the contractile mechanism so that depolarization of a muscle cell causes contraction. The coupling between nerve ending and muscle membrane occurs at a zone called a neuromyal junction. The nerve impulse releases a chemical sub-

stance, acetylcholine, which partially depolarizes the muscle membrane in the area of the neuromyal junction. This area itself is not excitable but acts as a low resistance, partially short-circuiting the nearby excitable membrane, which then discharges and initiates a depolarization wave that can travel along the muscle. In a gland, release of fluid commences when nerve impulses arrive. Other effects of nerve impulses include light production and electric-shock generation. Actions of and in higher animals, however, consist ultimately of muscle contraction and fluid secretion. For detailed information regarding the neuromyal junction, see MUSCLE CONTRACTION; for nerve conduction, see NERVE IMPULSE; and, for light production, see BIOLUMINESCENCE.

Effector organs and systems. *Bioelectric organs in fishes and eels.* Small electrical currents occur in freshwater or saltwater from a wide variety of causes. Variations in salinity, temperature, and water movement contribute to them. Contraction of muscles in moving fish and biological potentials from other causes (*e.g.*, even small animals buried beneath sand) generate appreciable electric fields. Members of an ancient group, in evolutionary terms, the elasmobranchs (*e.g.*, sharks and rays), have special organs for detecting feeble electric currents in water. Experiments indicate that a field as weak as 0.01 microvolts per centimetre is detected by these fishes. Frequently, the underwater environment is either devoid of light or turbid because of suspended matter. It is not surprising therefore, that some modern fish species have also evolved methods for sensing electric fields. Most of these species have evolved further than the elasmobranchs and can probe the environment with self-generated electric current. Any alterations in pattern, as a result of environmental changes or obstacles, are detected by receptor cells. The electric current probably functions also as a signal mechanism between fishes. Possibly the signals have environmental significance (*e.g.*, detection of like species, definition of territory) similar to those of insects, birds, and other animals.

Electric fields generated by fishes are usually either pulsating or oscillating. The mormyrid fishes of Africa generate apparently random clicklike electrical pulses or signals. The South American knife fish *Eigenmania* produces a continuous signal that oscillates at about 500 cycles per second. Most other electric fishes generate fields with lower frequencies. Electric fishes can be located in streams and lakes or identified in aquaria using a pair of earphones and lead wires into the water.

In addition to fishes that use electric fields for gathering information are species that generate substantial electric discharges in order to kill or stun prey or enemies.

Nature of the electric organ

Electric organs are derived from muscle precursors. The basic element of an electric organ is a flattened cell called an electroplaque. Large numbers of electroplaques are arranged in series and in parallel to build up voltage and current-producing capacity of the electric organ. Two modifications of this basic plan have occurred. In the electric eel (*Electricus electricus*) the electroplaque unit is modified muscle contractile tissue and therefore excitable. It can contribute the full reversal potential of about 150 millivolts when activated. In other cases (*e.g.*, some skates and rays) electroplaque units are elaborated from muscle-end-plate zones. Thus they can be depolarized down to zero but cannot be stimulated to reverse their polarization to the sodium equilibrium potential. Electroplaques are stimulated to activity by the action of nerve fibres. Nerve fibres reach only one of the two faces of an electroplaque unit. In the electroplaque unit derived from excitable tissue, the innervated face generates current; when excitation travels around the edges of the cell to the other face, it generates current also but in the opposite direction. In the second type of electroplaque, depolarization occurs only on the innervated face; the external current is generated by the uninnervated side of the cell, which passes external current through the innervated side, where electrical resistance is now low.

It has been shown that the amazingly steady frequency of the continuous electric discharge (as in *Eigenmania*) or the sudden, very large discharge (as in *Electricus electricus*) that generates 600–1000 volts at one ampere (in freshwater) are modulated from the brain. Electroplaques are spread along many centimetres of the body; for this reason precise timing is needed if all discharges are to occur in reasonable synchrony. Nerve impulses that actuate individual electroplaques are appropriately delayed in small fibres to the electroplaques to provide simultaneous action of all units.

Nature of the receptor organ

Electroreceptors respond either to externally generated electric fields or to electric current returning to the fish from its own electric organ. In the first case, the ultimate in electrosensitivity is advantageous, whereas, in the second case, the fish obtains information from tiny variations in the current received. The variations indicate that something has altered the normal field distribution. Very little is known of the mechanism of electroreception, especially concerning the way by which minute changes in a strong, periodic signal are coded into nerve impulses to give meaningful information to the brain.

Electroreceptors are located below skin, which possesses high electrical resistance in the area surrounding (but not over) the receptor. There is a pore, or zone of low electrical resistance, passing through the skin to a sensitive polarized membrane closely associated with a nerve-fibre termination. In receptors that receive current generated by the fish, some structure, perhaps another membrane, in the passage through the skin prevents steady current from flowing to the receptor; however, the system allows variations in the voltage developed to pass to the membrane in the way a series capacitor does in an electronic circuit. The electric field causes the release of a chemical transmitter substance for final stimulation of the sensory nerve fibres. The sensitive polarized membrane is associated with an energy-supply system, and under unusual conditions the membrane generates a sustained electrical oscillation of its own. M.V.L. Bennett has investigated this phenomenon in a number of species, and E.E. Suckling has obtained almost pure sine waves of several millivolts amplitude from the receptor organs of the fish *Gymnarchus*.

Endocrine systems. Homeostatic control of internal body environment is organized by means of innumerable, often interrelated feedback systems. Signalling in these systems is either by special molecules (hormones) from the endocrine glands carried in the blood or by nerve impulses. At many sites, but predominantly in the hypothalamus, there is intimate interrelationship between the two signalling systems; as a result, either hormones stimulate nerves or nerve impulses result in the production of hormones. For further information, see ENDOCRINE SYSTEM.

Central nervous system. The central nervous system, including the brain, carries out a computer-like function in which the nerve cell, under the influence of both stimulatory and inhibitory effects from other nerves synapsing on it, is the basic decision-making element. A nerve impulse and the action at a cell surface are electrochemical events. The memory function, which seems to influence intercellular action, must also be of a molecular nature. The bioelectric action of the brain is perhaps, therefore, best described as an essential correlate of basically chemical or molecular action. For further information, see NERVES AND NERVOUS SYSTEMS.

BIBLIOGRAPHY. E.E. SUCKLING, *Bioelectricity* (1961), a description of bioelectric phenomena and of methods of their detection; *The Living Battery* (1964), an introduction to bioelectricity at a popular level; T.C. RUCH and H.D. PATTON (eds.), *Physiology and Biophysics*, 19th ed. (1965), a textbook of physiology for medical students that provides sound theoretical bases for the material covered; C. CHAGAS and A. PAES DE CARVALHO (eds.), *Bioelectrogenesis* (1961), proceedings of a symposium containing source material on electric fish; H. GRUNDFEST, "Electric Fishes," *Scientific American*, 23:115–124 (Oct. 1960), a readable and authoritative account; M.A.B. BRAZIER, *The Electrical Activity of the Nervous System*, 3rd ed. (1968), a textbook for students; J.C. ECCLES, *The Physiology of Nerve Cells* (1957), an authoritative work, in paperback, by a noted neurophysiologist and Nobel Prize winner;

W.G. WALTER, *The Living Brain* (1953), a popular classic; B.F. HOFFMAN and P.F. CRANEFIELD, *Electrophysiology of the Heart* (1960), a review of studies using the micro-electrode on various types of heart cells; P.H. CAHN (ed.), *Lateral Line Detectors* (1967), proceedings of a conference containing material on electric sensing organs.

(E.E.S.)

Biogeographic Regions

Divisions of the Earth's surface showing differences in the average composition of the flora and the fauna are known as biogeographic regions or, when plants and animals are considered separately, as floristic and faunal, or zoogeographic, regions. Biogeographic regions refer mainly to terrestrial areas—continents and islands, and they are treated in this restrictive sense in this article. The recognition of biogeographic regions was the starting point of biogeography—the study of the distribution and dispersal of plants and animals.

GENERAL FEATURES

The conception of biotic distribution. Biogeographic regions describe the average patterns of distribution of present plant and animal species, as effected by a multitude of historical and current causes.

Determinants of present regional patterns

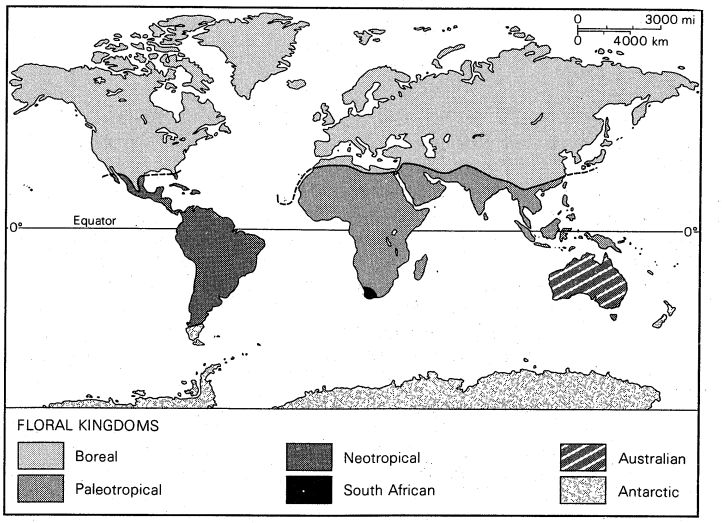
Present distribution patterns of plants and animals, therefore, are the result of: (1) current climatic and geographic conditions and zonations of the world (ecological biogeography); (2) the geological history of climates and landmasses (historical biogeography); and (3) the evolutionary history of the taxon (taxonomic category, such as species or genus) concerned. Adaptability to environmental conditions, rate of dispersal, and age of the taxa under consideration all lead to considerable differences in pattern and extent of distribution. Micro-organisms, for example, disperse more easily across ecological barriers, such as deserts and oceans, than do elephants. Annelid worms, known to have existed on Earth since Permian times (about 280,000,000 years ago), have had more chance to disperse than have cats and, therefore, have influenced the composition of regional faunas for a much longer period of time. These are matters of degree, not of principle; in older groups the results of extermination may be more apparent, whereas, in younger ones, the more recently required adjustments may have led to rapid spread over the accessible parts of the world.

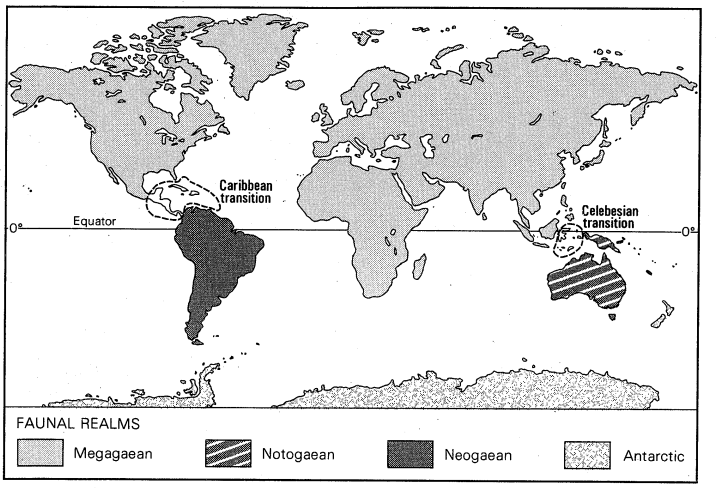
The boundaries of regions. Opposition to the concept of biogeographic regions has centred on the notion of regional boundaries. In view of the historical background of distribution patterns, a current tendency is to abandon biogeographical regional boundaries and to recognize instead overlapping and interdigitating regional floras and faunas. In support of that tendency, current classifications of floristic regions do not coincide with those of zoogeographic regions. Apparently, no true base for a combined biogeographic history of the world has as yet been found. One of the main reasons for the noncorrespondence of animal and plant regions could be the greater ease with which plants disperse across seas. The distribution of animals, however, is still very much influenced by vegetation and depends directly upon the availability of certain plants.

An early attempt to divide the vegetations of the world, based on characteristic plant species of a region, recognized 25 major divisions, or "kingdoms." A later classification was based more on characteristic floras than on characteristic species and more on flowering than on nonflowering plants; the flora of any region was considered to be conditioned by geological history and geographical circumstances and to have developed largely in the region in question. In general the latter scheme, with modifications, is in widespread use; it lists six kingdoms and 37 component regions.

Classification of Floristic Regions of the World (after Good, 1966)

1. Boreal kingdom
   - Arctic and subarctic region
   - Euro-Siberian region
   - Sino-Japanese region
   - Western and central Asiatic region
   - Mediterranean region
   - Macronesian region
   - Atlantic North American region
   - Pacific North American region
2. Paleotropical kingdom
   - African subkingdom
     - North African–Indian desert region
     - Sudanese park–steppe region
     - Northeast African highland and steppe region
     - West African rain-forest region
     - East African steppe region
     - South African region
     - Madagascar region
     - Ascension and St. Helena region (two South Atlantic islands)
   - Indo-Malaysian subkingdom
     - Indian region
     - Continental southeast Asiatic region
     - Malaysian region
   - Polynesian subkingdom
     - Hawaiian region
     - New Caledonian region
     - Melanesian and Micronesian region
     - Polynesian region
3. Neotropical kingdom
   - Caribbean region
   - Venezuelan and Guiana region
   - Amazon region
   - South Brazilian region
   - Andean region
   - Pampas region
   - Juan Fernandez region (three islands off Chile)
4. South African kingdom
   - Cape region
5. Australian kingdom
   - North and east Australian region
   - Southwest Australian region





(Top) Floral kingdoms of the world. (Bottom) Faunal realms of the world.
From W. Neill, *The Geography of Life*

Central Australian region
6. Antarctic kingdom
New Zealand region
Patagonian region
South temperate oceanic islands region

Wallace's zoogeographical regions

Alfred Russel Wallace, the founder of modern zoogeography, clearly stated in his classic book, *The Geographical Distribution of Animals* (1876), both the values and the restrictions of the meaning of zoogeographic regions. Wallace recognized six zoogeographical regions, which he named Palaearctic, Nearctic, Ethiopian, Oriental, Australian, and Neotropical. He emphasized the striking contrast of animal types between Bali and Lombok, two islands near Java; the strait separating these islands is very deep and sharply delimits the Oriental and Australian regions. Together with its continuation northward it has come to be called Wallace's Line.

A current system of zoogeographic regions is based on an 1858 classification of the distribution of perching birds and on Wallace's work on mammal distribution, both of which coincide with the main geological events that occurred in the Tertiary and Quartenary Periods (beginning about 65,000,000 years ago), when the continents of the Earth and the present division of land and sea were shaped. A recent sound classification of faunal distribution includes three realms and five regions.

Classification of Zoogeographic Regions of the World (after Darlington, 1957)

1. Realm Megagaea (or Arctogaea)
Holarctic region—the nontropical part of Eurasia, northern Africa, and North America
Palaearctic subregion (the Old World)
Nearctic subregion (the New World)
Ethiopian region—Africa south of the Sahara, and southwestern Arabia
Ethiopian subregion—continental Africa and southwestern Arabia
Madagascan subregion
Oriental, or Indian, region—tropical southern and southeastern Asia, including the Indo-Australian transition zone, consisting of Celebes, Moluccas, Lesser Sunda, and West-Papuan islands
2. Realm Notogaea
Australian region—Australia, New Guinea, New Zealand, and tropical Pacific islands
Australian subregion—Australia, Tasmania, and New Guinea
New Zealand subregion
Polynesian subregion
3. Realm Neogaea
South American, or Neotropical, region
4. Realm Antarctica
Antarctic region, including Antarctica and most of the subantarctic islands

Although different in many respects, the floral and faunal systems both are hierarchic; *i.e.*, their realms or kingdoms are divided into subgroups, regions, provinces, and districts. Unfortunately, the extent and limits of these units become increasingly disputable with the subtlety of the subdivisions. In an attempt to overcome such problems, the concepts of faunal elements and faunal types have been developed. Even so, the floristic subdivisions are larger in number than the faunal ones; the floristic kingdoms and subkingdoms may perhaps be equivalent to the level of region in zoogeography.

Climatic factors in classification

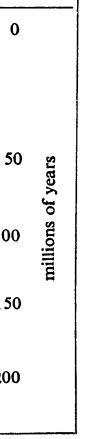
One important aspect of plant geographical regions—climate—is absent, or of only secondary importance, in zoogeography. Climatic influences in temperature and precipitation interfere with the geological and evolutionary aspects of floristic zonation. Consequently, vegetation types are arranged latitudinally, north and south of the Equator, in the following general order: jungle, savanna, desert, steppes, deciduous forest, coniferous forest, and tundra.

THE MEGAGAEAN REALM

Holarctic region, or Boreal kingdom. The Holarctic region includes the cold, temperate, and subtropical northern landmasses of the world, south to the Mexican and Saharan desert regions and the watershed of the Himalayas to southern China. In zoogeography, the border line usually follows the course of the Yangtze River; in plant geography, however, the border is a more southerly line, dividing subtropical and tropical vegetation. The unity of the Holarctic region is accentuated by the mainly circumpolar distribution patterns and dispersal faculties of the northern, or Arctic, flora and fauna and the occurrence of frequent land bridges across the Bering Strait during previous geological periods. The climate of the Bering land area during the last glacial period (which ended during the last 5,000 years) probably enabled a flora and fauna exchange of a tundra and taiga type only between northeastern Asia and northwestern North America. Land connections in the Pleistocene (beginning 2,500,000 years ago) and probably in the Tertiary (which extended from 65,000,000 to 2,500,000 years ago) are thought to have occurred during temperate or even tropical conditions, although evidence is not entirely convincing. The occurrence of land connections over the northern Atlantic is less certain; at least the recent flora and fauna of Europe and eastern North America show little evidence in favour of such a connection. Land connections over the North Atlantic in more remote geological times, however, have been postulated by geologists. The peculiarity of certain distribution patterns can be explained by assuming American–Eurasiatic contacts sometime in the past, but when and where they occurred is uncertain. Apart from Palaearctic (Old World) and Nearctic (New World) subregions, an Arctic subregion is sometimes recognized and may even be considered a region separate from the Holarctic. Almost identical climatic and ecological conditions reign circumpolarly over the ice-covered North Polar Basin and the islands and coastal fringes of the Eurasian and American continents. Animal and plant life face the same severe hazards of permanently frozen ground, long polar nights, short summers, low temperatures, blizzards, and low precipitation. The southern border of the Arctic subregion is limited ecologically by the Arctic treeline (or a line indicating the 10° C or 50° F average temperature in the month of July). A southern birch zone (sub-Arctic subregion) is replaced northwardly by a low-Arctic willow zone (polar willow and numerous saxifrages and heathlike plants) and a still more northern high-Arctic moss and lichen zone. In summer, boggy tundras prevail in the low-Arctic zone, and birds, insects (mainly flies and small beetles), and spiders dominate the fauna. Circumpolar Arctic mammals include the polar bear, Arctic fox, snowshoe hares, wolverine, lemmings, and reindeer. The musk ox is limited to northern Canada, northern Greenland, and Spitsbergen Island. Several species of seal and walrus occur exclusively along Arctic coasts and on pack ice. Circumpolar Arctic birds include the white-fronted goose, rock ptarmigan, snowy owl, and snow bunting. The few birds confined to the Old World Artic include the red-breasted goose and the little stint; those

Sequence and Duration of Geological Time Periods

era	period	epoch	duration in millions of years (approx.)	millions of years ago (approx.)
Cenozoic	Quaternary	recent	approx. last 10,000 years	0
		Pleistocene	2.5	25
	Tertiary	Pliocene	4.5	7
		Miocene	19	26
		Oligocene	12	38
		Eocene	16	54
		Paleocene	11	65
Mesozoic	Cretaceous		71	136
	Jurassic		54	190
	Triassic		35	225



Separation of the Arctic subregion

confined to the New World Arctic include Ross's goose, the semipalmated sandpiper, and the buff-breasted sandpiper. Virtually native Arctic bird groups include some groups of gulls, jaegers and skuas, phalaropes, alcids, and the snowy owl.

The Arctic subregion presents two interesting zoogeographical peculiarities. One relates to the estimated age of the Arctic subregion. Assuming that Arctic conditions in the northern part of the Northern Hemisphere have existed only since the end of the Tertiary (about 2,500,000 years ago), the present Arctic flora and fauna must be of relatively recent origin. The second basic zoogeographic peculiarity involves the relatively uncomplicated and direct food relationships in the Arctic, which provide models for studies on long- and short-term population dynamics and of predator–prey relationships. Lemmings, well-known rodent subjects of population studies, provide the basic food for a large number of avian and mammalian predators.

Extent of the taiga

The coniferous forest belt of Siberia and Canada is virtually circumglobal. It forms the taiga, or boreal, zone characterized by pines, spruces, and larches, interspersed, particularly along river courses, with birches, willows, poplars, and alders. The taiga, which forms the greater part of the Holarctic region, is the most extensive uniform vegetation of the world; it is bordered in the south by mixed deciduous forests in Europe, China, and North America and by wooded and shrub steppes in Central Asia. Lichens and carpet-forming heaths are conspicuous, as are tree-dwelling mammals and birds; reptiles, amphibians, and freshwater fish are very scarce. The Old and New World taigas have many species in common, including the stoat, or ermine, the lynx, the European elk (American moose), the hawk owl, the waxwing, the tree creeper, and the white-winged crossbill; most of these animals have developed special subspecies or races, some of which form interesting borderline cases of geographical species formation, such as the brown bears, the sable and marten, the red deer and wapiti, and the pygmy owls. Duplication of species, resulting from back colonization over the Bering Sea land bridges, has occurred in the brown bears, the three-toed woodpeckers, and the boreal owls. Throughout the entire Holarctic, Pleistocene glaciations have had a dramatic influence on the fauna of the taiga; hence, North America and, particularly, Eurasia are relatively poor in animal species when compared with the richness of the original Tertiary fauna, as reflected by the present faunal composition of the East Asiatic taiga. Over 200 species of birds breed in the East Asiatic taiga, whereas in the westerly extremity, Scandinavia, only about 20 species are found. Of bumblebees more than 30 species occur in Manchuria, more than ten in central Siberia, and only six in northern Europe. The Tertiary richness in number of species of plants and animals is most convincingly apparent in the mixed deciduous woods of Europe, Caucasia, North Iran, China, Japan, and the southwestern and southeastern United States. The significance of the so-called Arcto-Tertiary forest refugia (islands of land untouched by glaciers) is shown most strikingly in the large number of tree species in China and in the southern Allegheny Mountains in the United States.

Forest refugia

The mammalian fauna is extremely poor throughout the Holarctic region. Elephants, mastodonts, tapirs, rhinoceroses, and giraffes, as well as numerous buffalo, deer, and large carnivores became extinct at the time of the Pleistocene glaciations, during which the temperate deciduous woods, with accompanying animal life, were reduced to mere fractions of their original extent. Although shrews, moles, weasels, badgers, foxes, wolves, and squirrels, among the mammals, and hawks, owls, woodpeckers, thrushes, tits, finches, jays, and crows, among the birds, are widely distributed in the Holarctic region, the two subregions—Palaearctic and Nearctic—often contain different species of these animals.

A large number of terrestrial and freshwater animals are considered geographical and systematical relicts, some of which apparently go back to Tertiary times. Classical examples of such relicts are: the garpike of North America (with Cretaceous and Tertiary fossils in Europe, Asia, India, and North America); the bowfin of North America, the only remnant of a large group of fish that flourished in the Jurassic and Cretaceous periods, with Early Tertiary fossils in Europe and North America; and the giant salamanders (*Megalobatrachus*) of Japan and China, with Tertiary fossils in Europe and North America. Lake Baikal, in central Siberia, and Lake Ochrid, on the border of Yugoslavia and Albania, contain numerous noteworthy invertebrate relicts, some of which apparently are of great age.

Reptiles and amphibians, poor in species in the Holarctic, are more numerous in the southern, Mediterranean, and subtropical areas, particularly in North America, where the effect of glacial extermination has been diminished by colonization from the south. Freshwater fish, equally poor in species, are dominated by cypriniforms (carp and minnows).

Intergradation of the Palaearctic with the Oriental fauna and flora is noticeable in China and Japan. Numerous Oriental fishes and the frog *Rhacophorus*, as well as several sunbirds and pittas, have penetrated far into Palaearctic China. In Japan the influence of the Palaearctic fauna is conspicuous from north to south: the Kurile Islands with almost an Arctic component; Hokkaido, with a taiga flora and fauna; the islands of Hondo, with mixed deciduous woods in the north, and Shikoku and Kyushu, with subtropical conditions in the south; and lastly the Ryukyu Islands, where almost tropical conditions prevail. A similar mixture, with Ethiopian (tropical African) elements in northwestern Africa and the Near East is slight; somewhat more extensive is the mixture of Nearctic and neotropical elements in Mexico and the southern United States. The nine-banded armadillo, hummingbirds, and even the Canadian porcupine, must be considered as South American colonists in the Nearctic fauna.

Ethiopian region, or African subkingdom. The Ethiopian region includes Africa south of the Sahara, the southwestern tip of Arabia, Madagascar, and all surrounding islands—except the Canary and Cape Verde islands, which, like northern Africa, have a predominantly Palaearctic fauna. Most of the desert fauna of the Sahara also seem to be of Palaearctic origin. Throughout the Ethiopian region are contrasts between humid lowland forests and dry savanna, bush, or steppes. The flora and fauna of the lowland forests are thought to be geologically the oldest portions of the region. There is a resemblance between the floras of the African and tropical Asiatic forests, and some equatorial and West African animals have been considered as relicts from a time in the Tertiary when a forest and its contained animal life extended uninterrupted from West Africa to southeastern Asia. Alternation of relatively cold and dry and relatively warm and humid (pluvial) periods has influenced the African continent from the Late Tertiary to Recent times. Forest vegetation has extended at least on four occasions across the continent from the Atlantic to the Indian Ocean in such a way as to reduce and split the climatically defined arid areas with savanna, steppe, semidesert, and desert vegetations. It is not clear whether these pluvial and dry periods coincided with Northern Hemisphere glacial and interglacial periods.

Effect of alternating climate

The main vertebrate fauna of the African savanna and steppe vegetation are sometimes considered to have originated from the late Tertiary grassland fauna developed in Central Asia, which, according to a wealth of fossil evidence, extended from China to France, with very rich deposits in northern India and Greece. This fauna consisted of numerous species of antelope, giraffes, rhinoceroses, horses (the three-toed *Hipparion*), elephants, and large carnivores. At least part of the current savanna fauna seems to have been present in Africa long before the flourishing of a savanna fauna in Asia. It is not likely that the periodic restriction of forest communities in Africa in favour of dry, open country, has stimulated the development of large two-legged primates. Various kinds of australopithecine apes, which show

transitional stages between ape and man, support the theory that Africa is the continent of man's origin. At the present time Africa is enduring a dry period, which is diminishing severely the lowland tropical forests and its animals, especially in the narrow Guinean coastal belt and the Congo Basin. It is not too hazardous to foretell that, among others, the following forest species will become rarer and perhaps soon become extinct: the pygmy chimpanzee, the pygmy hippopotamus, the okapi, the Congo peacock, and the bareheaded rock fowl. In former ages numerous forest animals must have disappeared in the same manner; it is on this basis that the remarkable absence in Africa of forest rhinoceroses, tapirs, and gibbons (although occurring in Tertiary Europe and Asia and still surviving in southeastern Asia) can be explained. Although major extinctions have occurred, species formation has been favoured by the isolation of organisms into bands of savanna and forest populations. There are many examples of forest species that have developed savanna forms and vice versa. Nowhere else in the world has a mammalian fauna survived that is so rich in species and so spectacular as that in the East African savannas; they include the following animals, which at present do not occur outside the African continent: otter shrews, golden moles, elephant shrews, bushbabies, gorillas, chimpanzees, hydraxes, zebras, black and square-hipped rhinoceroses, giraffes, and at least 40 species of antelope. It has been stated that elephants, aardvarks, and storks originated in Africa, but the paleontological evidence does no fully support this supposition.

Rarity of endemic Ethiopian mammals and birds

Among mammals and birds, endemic Ethiopian groups are rare; on the family level only the following birds seem restricted to Africa south of the Sahara: hammerhead storks, secretary birds, guinea fowl, touracos, and colies, or mousebirds. Ostriches, so well known in the African bird fauna, have had Pleistocene representatives, possibly belonging to the same species, in the Palaearctic region as far as China. Still, the Saharan desert belt must have acted as a strong barrier to dispersal; groups such as bears, deer, nuthatches, wrens, salamanders, and pit vipers have never been observed in Africa south of the Sahara. Pre-Pleistocene relicts are found mainly in such groups as fish (*e.g.*, the bichirs, lungfish, or *Protopterus*, and the families Pantodontidae and Mormyridae), amphibians (phrynomerid and brevicipitid tree frogs, family Microhylidae), and reptiles (pelomedusid turtles, terrestrial lizards of the family Cordylidae, worm snakes of the families Typhlopidae and Leptotyphlopidae), most of which seem to have their closest relatives in South America. These relict forms have sometimes been used as zoogeographical evidence in favour of the theory of continental drift, which postulates that all continents were at one time united and drifted apart gradually (see CONTINENTAL DRIFT).

The unusual fauna of Madagascar—including about 30 species of tenrecs, about 15 species of lemurs, and the huge ostrichlike elephant birds (Aepyornithidae), which survived well into recent times—present interesting problems. It seems that, although Madagascar is the present home of the lemurs, lemur relatives such as monkeys and apes never succeeded in colonizing the island, a fact that could be interpreted to mean that the separation of Madagascar from Africa (assuming they were originally united) would date at least from Early Tertiary times, before the monkeys and apes appeared. Other well-known Madagascar animals include: two viverrids (the Malagasy civet and the fossa); four endemic bird families, with more than 45 genera; and 27 species of Chamaeleons. The following groups seem not to have reached Madagascar at any time: lungfish, cyprinoid fish, varanid lizards, viperine and elapine snakes, pythons, all carnivorous mammals (except the viverrids), and all hoofed mammals (except one pig and a Pleistocene hippopotamus). Of the insect fauna of Madagascar, 67 percent of the species are claimed endemic (not to occur elsewhere).

Oriental region, or Indo-Malaysian subkingdom. The Oriental region extends from the River Indus and the southern slopes of the Himalayas, through southern China to the Philippines and Indonesia, to the borders of the Australian region. This is more a botanical unit than a zoological one. The borders of the Indo-Malaysian subkingdom extend eastward as far as New Guinea and northward to include the Ryukyu Islands (generally considered to have Palaearctic animals).

The present distribution of such bird families as the hornbills (Bucerotidae), honey-guides (Indicatoridae), and weaverbirds (Ploceidae) seems to favour the recognition of a single palaeotropical biogeographic area (which would include Africa and the Orient), as does the distribution of a species such as the leopard, which occurs, almost without variation, from China to South Africa. On the other hand, the tiger has always been mainly restricted to the Oriental region, with extensions into the Palaearctic (Iran, Turkistan, China, Siberia). The fauna of peninsular India and Ceylon is conspicuously poor; it must have faced extreme periods of extermination through Tertiary volcanic activity and lava outflows, and through Pleistocene aridity. Burma and Indo-China, however, are among the richest areas in the world in animal species, and the natural wealth of Malaysia (Malaya, Sumatra, Borneo, Java) is equally well known. In central India are dry savannas, in the southwest are tropical mountain forests, and at the foothills of the Himalayas are humid luxuriant woodlands. Palaearctic and Chinese elements extend through the Himalayan mountain zones toward the west and through a broken chain of mountains in Indo-China to the south. The mountains of Kashmir form a knot where Palaearctic elements (Kashmir stag and jackdaw) meet with the Indian fauna. By way of the natural corridors outlined above, Palaearctic elements penetrate far into the Oriental region; *e.g.*, the great tit to Borneo, Java, and the Lesser Sunda Islands; the jay (*Garrulus glandarius*), throughout the Himalayas and southward to Burma, northern Thailand, and Indo-China; and the gray-headed woodpecker to the Malayan mountains and northern Sumatra. Large mammals of the Oriental region include the elephant, from Ceylon to Sumatra; the Malayan tapir, in Malaya, Sumatra, and Borneo; three species of rhinoceros, the great Indian, Javan, and Sumatran; several species of buffalo, including the gaur, the kouprey, and the banteng; several kinds of leaf monkeys and langurs and macaques; several gibbons; the orangutan, now restricted to small areas in Sumatra and Borneo; antelope, such as the bluebuck, or nilgai, and the blackbuck, or Indian antelope, from Pakistan and India; three species of pangolin closely related to African species; several lorises; about 20 species of tree shrews; mainly from Malaysia and the Philippines; and two species of colugos, or "flying lemurs." The Oriental region is especially well known for its large numbers of pheasants, woodpeckers, and babblers, and its small number of parrots. The west Indonesian islands (Sumatra, Borneo, and Java) are situated on the continental shelf of Asia (Sunda Shelf) and have been alternately united and separated during the Pleistocene, giving rise to repeated island isolation and, therefore, to active processes of species variation, thereby providing the faunal richness for which these tropical islands are known. The border of the Sunda Shelf was considered by Wallace as forming the border of the oriental region; the line, running between Bali and Lombok, northward between Borneo and Celebes and south of the Philippines, as mentioned above, is now known as Wallace's Line. The islands east of Wallace's Line have a fauna that differs from the westward islands mainly in the absence of minnows, wild cats, dogs, and deer. There is, however, a certain admixture of Australian elements present; marsupials—such as bandicoots and phalangers—cockatoos, and birds of paradise show an increase in abundance from west to east. The border of Australia's continental shelf, known here as the Sahul Shelf, running west of New Guinea and the Papuan islands of Misool, Waigeu, and the Aru Islands, is generally known as Lydekker's Line. The islands between (Celebes, Lesser Sunda Islands, Moluccas), known as Wallacea, form a famous zoogeographical transition area, although no floral division is apparent. A third line, Weber's Line, runs west of the Moluccas and the Kei Islands and denotes the places where the Oriental and Aus-

The richest faunal areas

Wallacea: a transitional area

tralian elements equal each other in numbers. Celebes is the home of such remarkable mammals as the dwarf buffalo, or anoa, and the babirusa (also on the Sula Islands and Buru). There are also two species of monkeys, a tarsier, two squirrels, a palm civet, and numerous birds. During the Pleistocene there even were elephants in this apparently unstable and highly changing part of the Asiatic shelf edge; the colonization route to the Celebes apparently ran mainly from the Philippines.

THE NOTOGAEAN REALM: AUSTRALIAN REGION

The Australian region constitutes the realm Notogaea and includes Australia, Tasmania, and New Guinea on one continental shelf, and the outlying islands of New Caledonia and New Zealand, and all of the Melanesian, Micronesian, and Polynesian islands. The region is characterized by birds of paradise and by egg-laying (monotreme) and pouched (marsupial) mammals. The monotremes include the duck-billed platypus and the echidnas, or spiny anteaters; the marsupials include numerous forms such as kangaroos, dasyures, and koalas. Marsupials and birds of paradise have crossed Lydekker's Line east to Celebes.

Of placental mammals only bats, water rats, and the dingo, or domestic dog, have penetrated the Australian region. The absence of woodpeckers, monkeys, squirrels, weasels, cats, and other large terrestrial predators has stimulated the development of a bizarre bird fauna in New Guinea. A highly differentiated adaptive expansion of marsupial mammals in Australia has provided a remarkably balanced fauna. Whether the present Australian monotremes and marsupials originated from stock that migrated from Asia across a Late Mesozoic or Early Tertiary land bridge between Australia and Asia (which afterwards disintegrated) or whether one or two species colonized the Australian continent by chance is at present impossible to determine. Alternation of climatic periods, comparable to, and probably synchronous with, those in Africa and South America, caused the isolation and constriction of forest communities in southeast and southwest Australia during arid periods, and the reduction of arid areas during moist periods; such climatic changes encouraged the emergence of new species. A humid, tropical flora and fauna still prevails in northeastern Australia, adding a third aspect to the continental wildlife—cassowaries and pythons. The marsupial fauna includes pouched mice, jerboas, and dasyures; the thylacine, or pouched wolf (*Thylacinus cynocephalus*); numbats; pouched moles (*Notoryctes*); bandicoots; cuscuses and phalangers; the koala, or pouched bear; wombats; and more than 50 species of wallabies, kangaroos, and rat kangaroos. There are at least 55 species of marsupials in New Guinea, but no native marsupials in New Caledonia or New Zealand (introductions, of course, have occurred).

Significance of marsupials in South America

The presence of marsupials in South America and of false beech forests (*Nothofagus*) in New Guinea, southeastern Australia, Tasmania, New Caledonia, New Zealand, and southwestern South America suggests the possibility of continental drift of the southern landmasses. At present, typically Australian forms, apart from mammals, include such birds as emus, mallee fowl, frogmouths, lyrebirds, bowerbirds, and Australian magpies; such reptiles as agamid lizards and elapid snakes; such amphibians as leptodactylid and hylid frogs; and such fishes as osteoglossids and a lungfish (*Epiceratodus*). Remarkable by their absence, apart from placental mammals, are pheasants and quail, hornbills, woodpeckers, true finches (Fringillinae), toads, frogs of the genus *Rana*, and cyprinoid fish. Floral elements include many trees adapted to arid conditions; especially conspicuous in species are *Eucalyptus* and *Acacia*. New Caledonia and New Zealand have impoverished Australian faunas, in which, through the absence of large predatory mammals, some noteworthy animals have survived as relicts: the kagu of New Caledonia, the kiwis, two endemic passerine bird families (Acanthisittidae and Callaeidae), a primitive group of frogs (*Leiopelma*), and the archaic reptile tuatara in New Zealand. Other animal groups have developed remarkable terrestrial, partly secondarily flightless, forms, such as the mainly nocturnal parrot, or kakapo; a large rail, the takahe, earlier considered extinct; and several flightless Pleistocene geese and songbirds.

The fauna of the tropical Pacific islands becomes poorer the farther the islands are separated from the Australian–New Guinean centre. The diversity of the fauna is, in addition, clearly related to the size and ecological diversity of the islands. Hurricanes, of regular occurrence in this area, are known to have had devastating effects on the land fauna and to have transported animal life over considerable distances. In the bird fauna, parrots, fruit pigeons, and honey eaters predominate; native rails, rendered flightless through insular isolation and specialization, are nowhere so numerous as they are in these islands.

The Hawaiian Islands lack native land mammals, except for a hoary bat (*Lasiurus*), apparently of American origin, but have a variety of reptiles, amphibians, and freshwater fishes. Birds, however, form the most remarkable asset of the fauna, including the widely varied family of Hawaiian honeycreepers, apparently derived from transoceanic chance colonizations from America, and the Nene, or Hawaiian goose, living on volcanic mountain slopes.

THE NEOGAEAN REALM: NEOTROPICAL REGION

The Neotropical region, which constitutes the realm Neogaea, extends south from the Sonoran (Mexican) transitional zone into South America as far as the temperate and subantarctic zones. The fauna of South America, including tropical Central America, elucidates the main zoogeographical principles in operation around the world. South America, isolated from the other continents from Early until Late Tertiary times, was not connected with North America through narrow land bridges until the end of the Pliocene Epoch (2,500,000 years ago). The fauna that developed is as unique and varied as that of the Australian region. At a time when predatory mammals were on the increase, South America was isolated and its unique fauna was thereby protected. The continent, however, reveals the kinds of animals and plants already present in other continents when it became isolated. Early and pre-Tertiary relicts can be found in the present Neotropical flora and fauna. Among these ancient South American inhabitants, indicators of a former inter-continental connection, postulated according to the continental drift theory, have been sought, particularly in view of similarities between the Neotropical and the Ethiopian floras and faunas.

The uniqueness of the South American fauna

More convincingly than anywhere else, the effects of large-scale faunal contacts are seen in South America. The land connection between South and North America had hardly been accomplished when Pleistocene glaciations started their devastating work on North American plant and animal life, forcing a rapid southward expansion of the northern Tertiary fauna. Soon North American mammals (including mastodonts and sabre-toothed tigers) reached even the southernmost tip of South America, clashing dramatically with the native South American fauna. Finally, the alternation of dry and cool versus moist and warm climates, which occurred also in the other southern continents, apparently led to the isolation of forest and savanna areas and accelerated the production of new species. The rise of the long chain of the Andes during the Tertiary Period further favoured the diversification of flora and fauna in highly distinct climatic mountain zones, which extend today, in Colombia, Ecuador, and Peru, from damp tropical lowlands to alpine snow-capped mountains and windswept high plateaus. Pre-Tertiary relicts of this region include: the lungfish *Lepidosiren* (with a related recent genus in Africa), the osteoglossid fishes *Osteoglossum* and *Arapaima* (claimed to be the largest of the freshwater fish), wormlike salamanders, clawed frogs (with recent African relatives), small bell frogs, side-necked turtles (with recent members living in all southern continents), large boine snakes (including the semi-aquatic anaconda), rheas, tinamous, screamers, the hoatzin, and the marsupial opossums and opossum rats. Resulting from the period of

Tertiary South American isolation are the highly varied groups of edentates, including such members as armadillos, the tree sloths, the anteaters and tamanduas, and extinct glyptodonts and large ground sloths, as well as the platyrrhine monkeys.

The South American monkeys represent a remarkable case of independent but parallel development with the Old World monkeys. The edentates, no less remarkable than the Australian marsupials, show an extremely wide ecological differentiation, which is apparent from the rich fossil evidence of Tertiary expansion of this group. No less noteworthy has been the ecological differentiation of the South American hoofed animals known as Notungulata, of which about 160 genera in 18 families have been described, none of which has survived. There were forms comparable to rhinoceroses, hippopotamuses, tapirs, camels, and horses, including such formidable animals as *Toxodon* and *Macrauchenia*. Apparently all of them became extinct during the Pleistocene, through competition with the North American invaders. All present South American ungulates (llamas, tapirs, deer, pigs) are of North American origin, the expansion occurring during the Pleistocene. Similarly, all early marsupial predatory mammals succumbed after the clash with North American placental predators, which gave rise to the present-day South American wolves, foxes, bears, jaguars, and pumas. The mainly arboreal South American opossums and the small opossum rats are vestiges of a formerly rich South American marsupial fauna. What has been described above for the mammals is representative of what happened to the whole of the fauna.

Extinction of exotic South American mammals

Native South American animals in the Neotropical region not mentioned earlier include: porcupines; cavias, from which the well-known guinea pig has been derived; the semiaquatic rodent capybara; pacas; agoutis; viscachas; hutias and coypus; toucans; hummingbirds; woodcreepers; ovenbirds; antbirds; and tyrant flycatchers. North American invaders not mentioned earlier include: toads, frogs, tortoises, colubrid snakes, squirrels, rabbits, raccoons, and coatis. The Chilean *Nothofagus* flora (southern beech forests), but not the fauna, may hold still recognizable indications of the existence of one former southern land mass (Gondwana Land), and thus confirm the fact of continental drift. The tropical South American fauna is extremely rich in birds and insects, including numerous interesting cases of ecological differentiation and of mimicry.

THE ANTARCTIC REALM: ANTARCTIC REGION

The Antarctic region constitutes the realm Antarctica. Although presently covered by a thick sheet of land ice, with very limited ice-free areas suitable for a free-living land fauna, the Antarctic continent seems to have harboured rich animal and plant life in earlier times, as shown by its fossils: insects and ferns (*Glossopteris*) from the Permian, a labyrinthodont amphibian from the Triassic, and a beetle from the Jurassic. At present, the fauna and flora of the Antarctic and its closest surrounding islands are differentiated enough to warrant its recognition as a separate biogeographical region. Of the arthropod fauna, formed of mites and ticks and springtails, 90 percent of the species seem to be endemic. Penguins, albatrosses, petrels, and seals form the main animal groups of this region. The great skua is the main avian predator, playing the combined role of falcon and vulture. The leopard seal, with its remarkable velocity on land, acts both as a terrestrial and a marine predator on penguins. On the whole, all of the extremely scarce plant and animal life in present-day Antarctica is concentrated on coastal and marine areas. On the sub-Antarctic islands tundra vegetation, with high grasses, heaths (*Azorella*), lichens, and mosses, prevails; the fauna, poor in species, usually is adapted to withstand high winds.

BIBLIOGRAPHY. The following three works form the classical basis of zoogeography: P.L. SCLATER, "On the General Distribution of the Members of the Class Aves," *Proc. Linn. Soc. Lond.*, 2:130–145 (1858); ALFRED RUSSEL WALLACE, *The Geographical Distribution of Animals* (1876); and J.A. ALLEN, "The Geographical Distribution of the Mammalia, Considered in Relation to the Principal Ontological Regions of the Earth, and the Laws That Govern the Distribution of Animal Life," *Bull. U.S. Geol. Geogrl. Surv. Terr.*, 4:313–378 (1878). Comprehensive modern textbooks in which both the problems and the literature pertaining to zoogeographical regions are treated clearly include L.F. DE BEAUFORT, *Zoogeography of the Land and Inland Waters* (1951); PHILIP J. DARLINGTON, *Zoogeography* (1957); and MIKLOS D.F. UDVARDY, *Dynamic Zoogeography* (1969). Plant geographical regions and related literature may be found in L. CROIZAT, *Manual of Phytogeography* (1952); and RONALD GOOD, *The Geography of the Flowering Plants*, 3rd ed. (1964). Rarely have surveys on plant and animal geography been combined, but an attempt, well worth reading by nonspecialists, has been made in DAVID J. DE LAUBENFELS, *A Geography of Plants and Animals* (1970), where further literature references combining the subjects may be found. Another work of interest is PHILIP J. DARLINGTON, *Biogeography of the Southern End of the World* (1965).

(K.H.V.)

Biographical Literature

One of the oldest forms of literary expression, biographical literature seeks to re-create in words the life of a human being, that of the writer himself or of another person, drawing upon the resources, memory and all available evidences—written, oral, pictorial.

In the present article, the subject is treated under four main heads: aspects of biographical literature, forms of biography, forms of autobiography, and historical development in Western and other literatures.

ASPECTS OF BIOGRAPHICAL LITERATURE

Historical aspects. Biography is sometimes regarded as a branch of history, and earlier biographical writings —such as the 15th-century *Mémoires* of the French councellor of state, Philippe de Commynes, or George Cavendish's 16th-century life of Thomas Cardinal Wolsey—have often been treated as historical material rather than as literary works in their own right. Some entries in ancient Chinese chronicles included biographical sketches; imbedded in the Roman historian Tacitus' *Annals* is the most famous biography of the emperor Tiberius; conversely, Sir Winston Churchill's magnificent life of his ancestor John Churchill, first duke of Marlborough, can be read as a history (written from a special point of view) of Britain and much of Europe during the War of the Spanish Succession (1701–14). Yet there is general recognition today that history and biography are quite distinct forms of literature. History usually deals in generalizations about a period of time (for example, the Renaissance), about a group of people in time (the English colonies in North America), about an institution (monasticism during the Middle Ages). Biography focusses upon a single human being and deals in the particulars of his life.

Both biography and history, however, are concerned with the past, and it is in the hunting down, evaluating, and selection of sources that they are akin. In this sense biography can be regarded as a craft rather than an art: techniques of research and general rules for testing evidence can be learned by anyone and thus need involve comparatively little of that personal commitment associated with art.

A biographer in pursuit of an individual long dead is usually hampered by a lack of sources: it is often impossible to check or verify what written evidence there is; there are no witnesses to cross-examine. No method has yet been developed by which to overcome such problems. Each life, however, presents its own opportunities as well as specific difficulties to the biographer: the ingenuity with which he handles gaps in the record—by providing information, for example, about the age that casts light upon the subject—has much to do with the quality of his resulting work. James Boswell knew comparatively little about Dr. Johnson's earlier years; it is one of the greatnesses of his *Life of Samuel Johnson LL.D.* (1791) that he succeeded, without inventing matter or deceiving the reader, in giving the sense of a life progressively unfolding. Another masterpiece of reconstruction in the face of little evidence is A.J.A. Symons' biography of the

Boswell's great life of Dr. Johnson

English author and eccentric Frederick William Rolfe, *The Quest for Corvo* (1934). A further difficulty is the unreliability of most collections of papers, letters, and other memorabilia edited before the 20th century. Not only did editors feel free to omit and transpose materials, but sometimes the authors of documents revised their personal writings for the benefit of posterity, often falsifying the record and presenting their biographers with a difficult situation when the originals were no longer extant.

The biographer writing the life of a person recently dead is often faced with the opposite problem: an abundance of living witnesses and a plethora of materials, which include the subject's papers and letters, sometimes reports of telephone conversations and conferences transcribed from tape, as well as the record of interviews granted the biographer by his subject's friends and associates. Frank Friedel, for example, in creating a biography of the United States president Franklin D. Roosevelt (1882–1945), has had to wrestle with something like 40 tons of paper. But finally, when writing the life of any man, whether long or recently dead, the biographer's chief responsibility is vigorously to test the authenticity of his materials by whatever rules and techniques are open to him.

Psychological aspects. Assembling a string of facts in chronological order does not constitute the life of a person, it only gives an outline of events. The biographer therefore seeks to elicit from his materials the motives for his subject's actions and to discover the shape of his personality. The biographer who has known his subject in life enjoys the advantage of his own direct impressions, often fortified by what the subject has himself revealed in conversations, and of his having lived in the same era (thus avoiding the pitfalls in depicting distant centuries). But on the debit side, such a biographer's view is coloured by the emotional factor almost inevitably present in a living association. Conversely, the biographer who knows his subject only from written evidence, and perhaps from the report of witnesses, lacks the insight generated by a personal relationship but can generally command a greater objectivity in his effort to probe his subject's inner life.

Biographers of the 20th century have had at their disposal the psychological theories and practice of Sigmund Freud and of his followers and rivals. The extent to which these new biographical tools for the unlocking of personality have been employed and the results of their use have varied greatly. On the one hand, some biographers have deployed upon their pages the apparatus of psychological revelation—analysis of behaviour symbols, interpretation based on the Oedipus complex, detection of Jungian archetypal patterns of behaviour, and the like. Other biographers, usually the authors of scholarly large-scale lives, have continued to ignore the psychological method; while still others, though avoiding explicit psychological analysis and terminology, have nonetheless presented aspects of their subjects' behaviours in such a way as to suggest psychological interpretations. In general, the movement, since World War I, has been toward a discreet use of the psychological method, from Katherine Anthony's *Margaret Fuller* (1920) and Joseph Wood Krutch's study of Edgar Allan Poe (1926), which enthusiastically embrace such techniques, through Erik Erikson's *Young Man Luther* (1958) and *Gandhi's Truth on the Origins of Militant Nonviolence* (1969), where they are adroitly and sagaciously used by a biographer who is himself a psychiatrist, to Leon Edel's vast biography of Henry James (vol. 1, 1953; 2 and 3, 1962; 4, 1969; 5, not yet published), where they are used with sophistication by a man of letters. The science of psychology has also begun to affect the biographer's very approach to his subject: a number of 20th-century authors seek to explore their own involvement with the person they are writing about before embarking upon the life itself.

Psychological interpretation of behaviour

Ethical aspects. The biographer, particularly the biographer of a contemporary, is often confronted with an ethical problem: how much of the truth, as he has been able to ascertain it, should be printed? Since the inception of biographical criticism in the later 18th century, this somewhat arid—because unanswerable—question has dominated both literary and popular discussion of biographical literature. Upon the publication of the *Life of Samuel Johnson*, James Boswell was bitterly accused of slandering his celebrated subject. More than a century and a half later, Lord Charles Moran's *Winston Churchill: The Struggle for Survival, 1940–1965* (1966), in which Lord Moran used the Boswellian techniques of reproducing conversations from his immediate notes and jottings, was attacked in much the same terms (though the question was complicated by Lord Moran's confidential position as Churchill's physician). In the United States, William Manchester's *Death of a President* (1967), on John F. Kennedy, created an even greater stir in the popular press. There the issue is usually presented as "the public's right to know"; but for the biographer it is a problem of his obligation to preserve historical truth as measured against the personal anguish he may inflict on others in doing so. Since no standard of "biographical morality" has ever been agreed upon—Boswell, Lord Moran, and Manchester have all, for example, had eloquent defenders—the individual biographer must steer his own course. That course in the 20th century is sometimes complicated by the refusal of the custodians of the papers of important persons, particularly national political figures, to provide access to all the documents.

The public's right to know

Aesthetic aspects. Biography, while related to history in its search for facts and its responsibility to truth, is truly a branch of literature because it seeks to elicit from facts, by selection and design, the illusion of a life actually being lived. Within the bounds of given data, the biographer seeks to transform plain information into illumination. If he invents or suppresses material in order to create an effect, he fails truth; if he is content to recount facts, he fails art. This tension, between the requirements of authenticity and the necessity for an imaginative ordering of materials to achieve lifelikeness, is perhaps best exemplified in the biographical problem of time. On the one hand, the biographer seeks to portray the unfolding of a life with all its cross-currents of interests, changing emotional states, events; yet in order to avoid reproducing the confusion and clutter of actual daily existence, he must interrupt the flow of diurnal time and group his materials so as to reveal traits of personality, grand themes of experience, and the actions and attitudes leading to moments of high decision. His achievement as a biographical artist will be measured, in great part, by his ability to suggest the sweep of chronology and yet to highlight the major patterns of behaviour that give a life its shape and meaning.

FORMS OF BIOGRAPHY

Biographies are difficult to classify. It is easily recognizable that there are many kinds of lifewriting, but one kind can easily shade into another; no standard basis for classification has yet been developed. A fundamental division offers, however, a useful preliminary view: biographies written from personal knowledge of the subject and those written from research.

Biographies written from firsthand knowledge. The biography that results from what might be called a vital relationship between the biographer and his subject often represents a conjunction of two main biographical forces: a desire on the part of the writer to preserve "the earthly pilgrimage of a man," as the 19th-century historian Thomas Carlyle calls it (*Critical and Miscellaneous Essays*, 1838), and an awareness that he has the special qualifications, because of direct observation and access to personal papers, to undertake such a task. This kind of biography is, in one form or another, to be found in most of the cultures that preserve any kind of written biographical tradition, and it is commonly to be found in all ages from the earliest literatures to the present. In its first manifestations, it was often produced by, or based upon the recollections of, the disciples of a religious figure—such as the biographical fragments concerning Buddha, portions of the Old Testament, and the Chris-

The relationship between writer and subject

tian gospels. It is sometimes called "source biography" because it preserves original materials, the testimony of the biographer, and often intimate papers of the subject (which have proved invaluable for later biographers and historians—as exemplified by Einhard's 9th-century *Vita Karoli imperatoris* ["Life of Charlemagne"] or Thomas Moore's *Letters and Journals of Lord Byron* [1830]). Biography based on a living relationship has produced a wealth of masterpieces: Tacitus' life of his father-in-law in the *Agricola*, William Roper's life of his father-in-law Sir Thomas More (1626), John Gibson Lockhart's biography (1837–38) of his father-in-law Sir Walter Scott, Johann Peter Eckermann's *Conversations with Goethe* (1836; trans. 1839), and Ernest Jones's *Life and Work of Sigmund Freud* (1953–57). Indeed, what is generally acknowledged as the greatest biography ever written belongs to this class: James Boswell's *Life of Samuel Johnson.*

Biographies compiled by research. Biographies that are the result of research rather than firsthand knowledge present a rather bewildering array of forms. First, however, there should be mentioned two special kinds of biographical activity.

Reference collections. Since the late 18th century, the Western world—and, in the 20th century, the rest of the world as well—has produced increasing numbers of compilations of biographical facts concerning both the living and the dead. These collections stand apart from literature. Many nations have multivolume biographical dictionaries such as the *Dictionary of National Biography* in Britain and the *Dictionary of American Biography* in the United States; general encyclopaedias contain extensive information about figures of world importance; classified collections such as *Lives of the Lord Chancellors* (Britain) and biographical manuals devoted to scholars, scientists, and other groups are available in growing numbers; information about living persons is gathered into such national collections as *Who's Who?* (Britain), *Qui è?* (Italy), and *Who's Who in America?*

Character sketches. The short life, however, is a genuine current in the mainstream of biographical literature and is represented in many ages and cultures. Excluding early quasi-biographical materials about religious or political figures, the short biography first appeared in China at about the end of the 2nd century BC, and two centuries later it was a fully developed literary form in the Roman Empire. The *Shih-chi* ("Historical Records"), by Ssu-ma Ch'ien (145?–*c.* 85 BC), include lively biographical sketches, very short and anecdotal with plentiful dialogue, grouped by character-occupation types such as "maligned statesmen," "rash generals," "assassins," a method that became established tradition with the *Han shu* (*History of the Former Han Dynasty*), by Ssu-ma Ch'ien's successor and imitator, Pan Ku (AD 32–92). Toward the end of the first century AD, in the Mediterranean world, Plutarch's *Lives of the Noble Grecians and Romans*, which are contrasting pairs of biographies, one Greek and one Roman, appeared; there followed within a brief span of years the *Lives of the Caesars*, by the Roman emperor Hadrian's librarian Suetonius. These works established a quite subtle mingling of character sketch with chronological narrative that has ever since been the dominant mark of this genre. Plutarch, from an ethical standpoint emphasizing the political virtues of man as governor, and Suetonius, from the promptings of sheer biographical curiosity, develop their subjects with telling details of speech and action; and though Plutarch, generally considered to be the superior artist, has greatly influenced other arts than biographical literature—witness Shakespeare's Roman plays, which are based on his *Lives*—Suetonius created in the *Life of Nero* one of the supreme examples of the form. Islāmic literature, from the 10th century, produced short "typed" biographies based on occupation—saints, scholars, and the like—or on arbitrarily chosen personal characteristics. The series of brief biographies has continued to the present day with such representative collections as, in the Renaissance, Giorgio Vasari's *Lives of the Most Eminent Italian Painters, Sculptors,* and *Architects*, Thomas Fuller's *History of the Worthies of England* in the 17th century, Samuel Johnson's *Lives of the English Poets* in the 18th, and, in more recent times, the "psychographs" of the American Gamaliel Bradford (*Damaged Souls*, 1923), Lytton Strachey's *Eminent Victorians* (1918) and the "profiles" that have become a hallmark of the weekly magazine *The New Yorker.*

The first short biography

Further classification of biographies compiled by research can be achieved by regarding the comparative objectivity of approach. For convenience, six categories, blending one into the other in infinite gradations and stretching from the most objective to the most subjective, can be employed.

Informative biography. This, the first category, is the most objective and is sometimes called "accumulative" biography. The author of such a work, avoiding all forms of interpretation except selection—for selection, even in the most comprehensive accumulation, is inevitable—seeks to unfold a life by presenting, usually in chronological order, the paper remains, the evidences, relating to that life. This biographer takes no risks but, in turn, seldom wins much critical acclaim: his work is likely to become a prime source for biographers who follow him. During the 19th century, the *Life of Milton: Narrated in Connection with the Political, Ecclesiastical, and Literary History of his Time* (7 vol., 1859–94), by David Masson, and *Abraham Lincoln: A History* (10 vol., 1890), by John G. Nicolay and John Hay, offer representative samples. In the 20th century such works as Edward Nehls's, *D.H. Lawrence: A Composite Biography* (1957–59) and David Alec Wilson's collection of the life records of Thomas Carlyle (1923–29), in six volumes, continue the traditions of this kind of life writing.

Inevitability of selection in biography

Critical biography. This second category, scholarly and critical, unlike the first, does offer a genuine presentation of a life. These works are very carefully researched; sources and "justifications" (as the French call them) are scrupulously set forth in notes, appendixes, bibliographies; inference and conjecture, when used, are duly labelled as such; no fictional devices or manipulations of material are permitted, and the life is generally developed in straight chronological order. Yet such biography, though not taking great risks, does employ the arts of selection and arrangement. The densest of these works, completely dominated by fact, have small appeal except to the specialist. Those written with the greatest skill and insight are in the first rank of modern life writing. In these scholarly biographies—the "life and times" or the minutely detailed life—the author is able to deploy an enormous weight of matter and yet convey the sense of a personality in action, as exemplified in Leslie Marchand's *Byron* (1957), with some 1,200 pages of text and 300 pages of notes, Dumas Malone's *Jefferson and his Time* (4 vol., 1948–70), Churchill's *Marlborough* (1933–38), Douglas S. Freeman's *George Washington* (1948–57). The critical biography aims at evaluating the works as well as unfolding the life of its subject, either by interweaving the life in its consideration of the works or else by devoting separate chapters to the works. Critical biography has had its share of failures: except in skillful hands, criticism clumsily intrudes upon the continuity of a life, or the works of the subject are made to yield doubtful interpretations of character, particularly in the case of literary figures. It has to its credit, however, such fine biographies as Arthur S. Link, *Wilson* (5 vol. 1947–65); Richard Ellmann, *James Joyce* (1959); Ernest Jones, *The Life and Works of Sigmund Freud;* Douglas S. Freeman, *Lee* (1934–35); and Edgar Johnson, *Charles Dickens* (1952).

The problems of critical biography

"Standard" biography. This third, and central, category of biography, balanced between the objective and the subjective, represents the mainstream of biographical literature, the practice of biography as an art. From antiquity until the present—within the limits of the psychological awareness of the particular age and the availability of materials—this kind of biographical literature has had as its objective what Sir Edmund Gosse called "the faithful portrait of a soul in its adventures through life." It seeks to transform, by literary methods that do not dis-

tort or falsify, the truthful record of fact into the truthful effect of a life being lived. Such biography ranges in style and method from George Cavendish's 16th-century life of Cardinal Wolsey, Roger North's late-17th-century lives of his three brothers, and Boswell's life of Johnson to modern works like Lord David Cecil's *Melbourne,* Garrett Mattingly's *Catherine of Aragon,* Andrew Turnbull's *Scott Fitzgerald,* and Leon Edel's *Henry James.*

Interpretative biography. This fourth category of life writing is subjective and has no standard identity. At its best it is represented by the earlier works of Catherine Drinker Bowen, particularly her lives of Tchaikovsky, *"Beloved Friend"* (1937), and Oliver Wendell Holmes, *Yankee from Olympus* (1944). She molds her sources into a vivid narrative, worked up into dramatic scenes that always have some warranty of documentation—the dialogue, for example, is sometimes devised from the indirect discourse of letter or diary. She does not invent materials; but she quite freely manipulates them—that is to say, interprets them—according to the promptings of insight, derived from arduous research, and with the aim of unfolding her subject's life as vividly as possible. (Mrs. Bowen, much more conservative in her later works, clearly demonstrates the essential distance between the third and fourth categories: her distinguished life of Sir Edward Coke, *The Lion and the Throne* [1957], foregoes manipulation and the "recreation" of dialogue and limits interpretation to the artful deployment of biographical resources.) Very many interpretative biographies stop just short of fictionalizing in the freedom with which they exploit materials. The works of Frank Harris (*Oscar Wilde,* 1916) and Hesketh Pearson (*Tom Paine, Friend of Mankind,* 1937; *Beerbohm Tree,* 1956) demonstrate this kind of biographical latitude.

Fictionalized biography. The books in this fifth category belong to biographical literature only by courtesy. Materials are freely invented, scenes and conversations are imagined; unlike the previous category, this class often depends almost entirely upon secondary sources and cursory research. Its authors, well represented on the paperback shelves, have created a hybrid form designed to mate the appeal of the novel with a vague claim to authenticity. This form is exemplified by writers such as Irving Stone, in his *Lust for Life* (on van Gogh) and *The Agony and the Ecstasy* (on Michelangelo). Whereas the compiler of biographical information (the first category) risks no involvement, the fictionalizer admits no limit to it.

Cursory research in fictionalized biography

Fiction presented as biography. The sixth and final category is outright fiction, the novel written as biography or autobiography. It has enjoyed brilliant successes. Such works do not masquerade as lives; rather, they imaginatively take the place of biography where perhaps there can be no genuine life writing for lack of materials. Among the most highly regarded examples of this genre are, in the guise of autobiography, Robert Graves's books on the Roman emperor Claudius, *I, Claudius* and *Claudius the God and His Wife Messalina;* Mary Renault's *The King Must Die* on the legendary hero Theseus; and Marguerite Yourcenar's *Memoirs of Hadrian.* The diary form of autobiography was amusingly used by George and Weedon Grossmith to tell the trials and tribulations of their fictional character, Charles Poster, in *The Diary of a Nobody* (1892). In the form of biography this category includes Graves's *Count Belisarius* and Hope Muntz's *Golden Warrior* (on Harold II, vanquished at the Battle of Hastings, 1066). Some novels-as-biography, using fictional names, are designed to evoke rather than re-create an actual life, such as W. Somerset Maugham's *Moon and Sixpence* (Gauguin) and *Cakes and Ale* (Thomas Hardy) and Robert Penn Warren's *All the King's Men* (Huey Long).

"Special-purpose" biography. In addition to these six main categories, there exists a large class of works that might be denominated "special-purpose" biography. In these works the art of biography has become the servant of other interests. They include potboilers (written as propaganda or as a scandalous exposé) and "as-told-to" narratives (often popular in newspapers) designed to publicize a celebrity. This category includes also "campaign biographies" aimed at forwarding the cause of a political candidate (Nathaniel Hawthorne's *Life of Franklin Pierce* [1852] being an early example); the weighty commemorative volume, not infrequently commissioned by the widow (which, particularly in Victorian times, has usually enshrouded the subject in monotonous eulogy); and pious works that are properly called hagiography, or lives of holy men, written to edify the reader.

FORMS OF AUTOBIOGRAPHY

Autobiography is a very close relative, or special form, of biographical literature: it is the life of a man that happens to have been written by himself and is therefore unfinished.

Informal autobiography. Autobiography, like biography, manifests a wide variety of forms, beginning with the intimate writings made during a life that were not intended (or apparently not intended) for publication.

Letters, diaries, and journals. Broadly speaking, the order of this category represents a scale of increasingly self-conscious revelation. Collected letters, especially in carefully edited modern editions such as W.S. Lewis' of the correspondences of the 18th-century man of letters Horace Walpole (34 vol., 1937–65), can offer a rewarding though not always predictable experience: some eminent people commit little of themselves to paper, while other lesser figures pungently re-create themselves and their world. The 15th-century *Paston Letters* constitute an invaluable chronicle of the web of daily life woven by a tough and vigorous English family among the East Anglian gentry during the Wars of the Roses; the composer Mozart and the poet Byron, in quite different ways, are among the most revealing of letter writers. Diarists have made great names for themselves out of what seems a humble branch of literature. To mention only two, in the 20th century the young Jewish girl Anne Frank created such an impact by her recording of narrow but intense experience that her words were translated to stage and screen; while a comparatively minor figure of 17th-century England, Samuel Pepys—he was secretary to the navy—has immortalized himself in a diary that exemplifies the chief qualifications for this kind of writing—candour, zest, and an unselfconscious enjoyment of self. The somewhat more formal journal is likewise represented by a variety of masterpieces, from the notebooks, which reveal the teeming, ardent brain of Leonardo da Vinci, and William Wordsworth's sister Dorothy's sensitive recording of experience in her *Journals* (1897), to French foreign minister Armand de Caulaincourt's recounting of his flight from Russia with Napoleon (translated as *With Napoleon in Russia,* 1935) and the *Journals* of the brothers Goncourt, which present a confidential history of the literary life of mid-19th-century Paris.

Memoirs and reminiscences. These are autobiographies that usually emphasize *what* is remembered rather than *who* is remembering; the author, instead of recounting his life, deals with those experiences of his life, people, and events that he considers most significant. (The extreme contrast to memoirs is the spiritual autobiography, so concentrated on the life of the soul that the author's outward life and its events remains a blur. The artless *res gestae,* a chronology of events, occupies the middle ground.)

The nature of memoirs

In the 15th century, Philippe de Commynes, modestly effacing himself except to authenticate a scene by his presence, presents in his *Mémoires* a life of Louis XI, master of statecraft, as witnessed by one of the most sagacious counsellors of the age. The memoirs of Giacomo Casanova boast of an 18th-century rake's adventures; those of Hector Berlioz explore with great brilliance the trials of a great composer, the reaches of an extraordinary personality, and the musical life of Europe in the first part of the 19th century. The memoir form is eminently represented in modern times by Sir Osbert Sitwell's polished volumes, presenting a tapestry of recollections that, as has been observed, "tells us little about what it feels like to be in Sir Osbert's skin"—a phrase

perfectly illustrating the difference between memoirs and formal autobiography.

Formal autobiography. This category offers a special kind of biographical truth: a life, reshaped by recollection, with all of recollection's conscious and unconscious omissions and distortions. The novelist Graham Greene says that, for this reason, an autobiography is only "a sort of life" and uses the phrase as the title for his own autobiography (1971). Any such work is a true picture of what, at one moment in a life, the subject wished—or is impelled—to reveal of that life. An event recorded in the autobiographer's youthful journal is likely to be somewhat different from that same event recollected in later years. Memory being plastic, the autobiographer regenerates his materials as he uses them. The advantage of possessing unique and private information, accessible to no researching biographer, is counterbalanced by the difficulty of establishing a stance that is neither overmodest nor aggressively self-assertive. The historian Edward Gibbon declares, ". . . I must be conscious that no one is so well qualified as myself to describe the service of my thoughts and actions." The 17th-century English poet Abraham Cowley provides a rejoinder: "It is a hard and nice subject for a man to write of himself; it grates his own heart to say anything of disparagement and the reader's ears to hear anything of praise from him."

There are but few and scattered examples of autobiographical literature in antiquity and the Middle Ages. In the 2nd century BC the Chinese classical historian Ssu-ma Ch'ien included a brief account of himself in the *Shih-chi*, "Historical Records." It is stretching a point to include, from the 1st century BC, the letters of Cicero (or, in the early Christian era, the letters of St. Paul); and Julius Caesar's *Commentaries* tell little about Caesar, though they present a masterly picture of the conquest of Gaul and the operations of the Roman military machine at its most efficient. The *Confessions* of St. Augustine, of the 5th century AD, belong to a special category of autobiography discussed below; the 14th-century *Letter to Posterity* of the Italian poet Petrarch is but a brief excursion in the field.

The first autobiography

Speaking generally, then, it can be said that autobiography begins with the Renaissance in the 15th century; and, surprisingly enough, the first example was written not in Italy but in England by a woman entirely untouched by the "new learning" or literature. In her old age Margery Kempe, the sobbing mystic, or hysteric, of Lynn in Norfolk, dictated an account of her bustling, far-faring life, which, however concerned with religious experience, racily reveals her somewhat abrasive personality and the impact she made upon her fellows. This is done in a series of scenes, mainly developed by dialogue. Though calling herself, in abject humility, "the creature," Margery knew, and has effectively transmitted the proof, that she was a remarkable person.

The first full-scale formal autobiography was written a generation later by a celebrated Humanist-publicist of the age, Enea Sylvio Piccolomini, after he was elevated to the papacy, in 1458, as Pius II—the result of an election that he recounts with astonishing frankness spiced with malice. In the first book of his autobiography—misleadingly named *Commentarii*, in evident imitation of Caesar—Pius II traces his career up to becoming pope; the succeeding 11 books (and a fragment of a 12th, which breaks off a few months before his death in 1464) present a panorama of the age, with its cruel and cultivated Italian tyrants, cynical *condottieri* (professional soldiers), recalcitrant kings, the politics and personalities behind the doors of the Vatican, and the urbane but exuberant character of the Pope himself. Pius II exploits the plasticity of biographical art by creating opportunities—especially when writing of himself as the connoisseur of natural beauties and antiquities—for effective autobiographical narration. His "Commentaries" show the art of formal autobiography in full bloom in its beginnings; they rank as one of its half dozen greatest exemplars.

The neglected autobiography of the Italian physician and astrologer Gironimo Cardano, a work of great charm, and the celebrated adventures of the goldsmith and sculptor Benvenuto Cellini in Italy of the 16th century; the uninhibited autobiography of the English historian and diplomat Lord Herbert of Cherbury, in the early 17th; and Colley Cibber's *Apology for The Life of Colley Cibber, Comedian* in the early 18th—these are representative examples of biographical literature from the Renaissance to the Age of Enlightenment. The latter period itself produced three works that are especially notable for their very different reflections of the spirit of the times as well as of the personalities of their authors: the urbane autobiography of Edward Gibbon, the great historian; the plainspoken, vigorous success story of an American who possessed all the talents, Benjamin Franklin; and the somewhat morbid introspection of a revolutionary Swiss-French political and social theorist, the *Confessions* of J.-J. Rousseau—the latter leading to two autobiographical explorations in poetry during the Romantic Movement (flourished 1798–1837) in England, Wordsworth's *Prelude* and Byron's *Childe Harold*, cantos III and IV. Significantly, it is at the end of the 18th century that the word autobiography apparently first appears in print, in *The Monthly Review*, 1797.

Autobiography in the Age of Enlightenment

Specialized forms. These might roughly be grouped under four heads: thematic, religious, intellectual, and fictionalized. The first grouping includes books with such diverse purposes as Adolf Hitler's *Mein Kampf* (1924), *The Americanization of Edward Bok* (1920), and Richard Wright's *Native Son* (1940). Religious autobiography claims a number of great works, ranging from the *Confessions* of St. Augustine and Peter Abelard's *Historia Calamitatum* (*The Story of My Misfortunes*) in the Middle Ages to the autobiographical chapters of Thomas Carlyle's *Sartor Resartus* ("The Everlasting No," "Centre of Indifference," "The Everlasting Yea") and Cardinal John Newman's beautifully wrought *Apologia* in the 19th century. That century and the early 20th saw the creation of several intellectual autobiographies. The *Autobiography* of the philosopher John S. Mill, severely analytical, concentrates upon "an education which was unusual and remarkable." It is paralleled, across the Atlantic, in the bleak but astringent quest of *The Education of Henry Adams* (1907). Edmund Gosse's sensitive study of the difficult relationship between himself and his Victorian father, *Father and Son* (1907), and George Moore's quasi-novelized crusade in favour of Irish art, *Hail and Farewell* (1911–14), illustrate the variations of intellectual autobiography. Finally, somewhat analogous to the novel as biography (for example, Graves's *I, Claudius*) is the autobiography thinly disguised as, or transformed into, the novel. This group includes such works as Samuel Butler's *Way of All Flesh* (1903), James Joyce's *Portrait of the Artist as a Young Man* (1916), George Santayana's *Last Puritan* (1935), and the gargantuan novels of Thomas Wolfe (*Look Homeward, Angel* [1929], *Of Time and the River* [1935]).

HISTORICAL DEVELOPMENT

Western literature. *Antiquity.* In the Western world, biographical literature can be said to begin in the 5th century BC with the poet Ion of Chios, who wrote brief sketches of such famous contemporaries as Pericles and Sophocles. It continued throughout the classical period for a thousand years, until the dissolution of the Roman Empire in the 5th century AD. Broadly speaking, the first half of this period exhibits a considerable amount of biographical activity, of which much has been lost; such fragments as remain of the rest—largely funeral elegies and rhetorical exercises depicting ideal types of character or behaviour—suggest that from a literary point of view the loss is not grievous. (An exception is the life of the Roman art patron Pomponius Atticus, written in the 1st century BC by Cornelius Nepos.) Biographical works of the last centuries in the classical period, characterized by numerous sycophantic accounts of emperors, share the declining energies of the other literary arts. But although there are few genuine examples of life writing, in the modern sense of the term, those few are masterpieces. The two greatest teachers of the classical Mediter-

ranean world, Socrates and Jesus Christ, both prompted the creation of magnificent biographies written by their followers. To what extent Plato's life of Socrates keeps to strict biographical truth cannot now be ascertained (though the account of Socrates given by Plato's contemporary the soldier Xenophon, in his *Memorabilia*, suggests a reasonable faithfulness) and he does not offer a full-scale biography. Yet in his two consummate biographical dialogues—*The Apology* (recounting the trial and condemnation of Socrates) and the *Phaedo* (a portrayal of Socrates' last hours and death)—he brilliantly re-creates the response of an extraordinary character to the crisis of existence. Some 400 years later there came into being four lives of Jesus, the profound religious significance of which has inevitably obscured their originality—their homely detail, anecdotes, and dialogue that, though didactic in purpose, also evoke a time and a personality. The same century, the first of the Christian era, gave birth to the three first truly "professional" biographers—Plutarch and Suetonius (discussed above) and the historian Tacitus, whose finely wrought biography of his father-in-law, *Agricola*, concentrating on the administration rather than the man, has something of the monumental quality of Roman architecture. The revolution in thought and attitude brought about by the growth of Christianity is signalled in a specialized autobiography, the *Confessions* of St. Augustine; but the biographical opportunity suggested by Christian emphasis on the individual soul was, oddly, not to be realized. If the blood of the martyrs fertilized the seed of the new faith, it did not promote the art of biography. The demands of the church and the spiritual needs of men, in a twilight world of superstition and violence, transformed biography into hagiography. There followed a thousand years of saints' lives: the art of biography forced to serve ends other than its own.

Plato's accounts of Socrates

Middle Ages. This was a period of biographical darkness, an age dominated by the priest and the knight. The priest shaped biography into an exemplum of otherworldliness, while the knight found escape from daily brutishness in allegory, chivalric romances, and broad satire (the fabliaux). Nevertheless, glimmerings can be seen. A few of the saints' lives, like Eadmer's *Life of Anselm*, contain anecdotal materials that give some human flavour to their subjects; the 13th-century French nobleman Jean, sire de Joinville's life of St. Louis (Louis IX of France), *Mémoires*, offers some lively scenes. The three most interesting biographical manifestations came early. Bishop Gregory of Tours' *History of the Franks* depicts artlessly but vividly, from firsthand observation, the lives and personalities of the four grandsons of Clovis and their fierce queens in Merovingian Gaul of the 6th century. Bede's *Ecclesiastical History of the English People*, of the 8th century, though lacking the immediacy and exuberance—and the violent protagonists—of Gregory, presents some valuable portraits, like those of "the little dark man," Paulinus, who converted the King of Northumbria to Christianity.

Einhard's "Life of Charlemagne"

Most remarkable, however, a self-consciously wrought work of biography came into being in the 9th century: this was the "Life of Charlemagne," written by a cleric at his court named Einhard. He is aware of his biographical obligations and sets forth his point of view and his motives:

> I have been careful not to omit any facts that could come to my knowledge, but at the same time not to offend by a prolix style those minds that despise everything modern. . . . No man can write with more accuracy than I of events that took place about me, and of facts concerning which I had personal knowledge. . . .

He composes the work in order to ensure that Charlemagne's life is not "wrapped in the darkness of oblivion" and out of gratitude for "the care that King Charles bestowed upon me in my childhood, and my constant friendship with himself and his children." Though Einhard's biography, by modern standards, lacks sustained development, it skillfully reveals the chief patterns of Charlemagne's character—his constancy of aims, powers of persuasion, passion for education. Einhard's work is far closer to modern biography than the rudimentary poetry and drama of his age are to their modern counterparts.

Renaissance. Like the other arts, biography stirs into fresh life with the Renaissance in the 15th century. Its most significant examples were autobiographical, as has already been mentioned. Biography was chiefly limited to uninspired panegyrics of Italian princes by their court Humanists, such as Simonetta's life of the great *condottiere*, Francesco Sforza, duke of Milan.

During the first part of the 16th century in England, now stimulated by the "new learning" of Erasmus, John Colet, Thomas More, and others, there were written three works that can be regarded as the initiators of modern biography: More's *History of Richard III*, William Roper's *Mirrour of Vertue in Worldly Greatness; or, the life of Syr Thomas More*, and George Cavendish's *Life of Cardinal Wolsey*. The *History of Richard III* (written about 1513 in both an English and a Latin version) unfortunately remains unfinished; and it cannot meet the strict standards of biographical truth since, under the influence of classical historians, a third of the book consists of dialogue that is not recorded from life. However, it is a brilliant work, exuberant of wit and irony, that not only constitutes a biographical landmark but is also the first piece of modern English prose. With relish, More thus sketches Richard's character:

> He was close and secret, a deep dissembler, lowly of countenance, arrogant of heart, outwardly companionable where he inwardly hated, not hesitating to kiss whom he thought to kill.

Worked up into dramatic scenes, this biography, as reproduced in the *Chronicles* of Edward Hall and Raphael Holinshed, later provided both source and inspiration for Shakespeare's rousing melodramatic tragedy, *Richard III*. The lives written by Roper and Cavendish display interesting links, though the two men were not acquainted: they deal with successive first ministers destroyed by that brutal master of politics, Henry VIII; they are written from first hand observation of their subjects by, respectively, a son-in-law and a household officer; and they exemplify, though never preach, a typically Renaissance theme: *Indignatio principis mors est*—"the Prince's anger is death." Roper's work is shorter, more intimate, and simpler; in a series of moving moments it unfolds the struggle within Sir Thomas More between his duty to conscience and his duty to his king. Cavendish offers a more artful and richly developed narrative, beautifully balanced between splendid scenes of Wolsey's glory and vanity and ironically contrasting scenes of disgrace, abasement, and painfully achieved self-knowledge.

Biography in the Renaissance

The remaining period of the Renaissance, however, is disappointingly barren. In Russia, which had also produced medieval saints' lives, there appears a modest biographical manifestation in the *Stepennaya Kniga* ("Book of Degrees," 1563), a collection of brief lives of princes and prelates. Somewhat similarly, in France, torn by religious strife, Pierre Brantôme wrote his *Lives of Famous Ladies* and *Lives of Famous Men*. The Elizabethan Age in England, for all its magnificent flowering of the drama, poetry, and prose, did not give birth to a single biography worthy of the name. Sir Fulke Greville's account of Sir Philip Sidney (1652) is marred by tedious moralizing; Francis Bacon's accomplished life of the first Tudor monarch, *The Historie of the Raigne of King Henry the Seventh* (1622), turns out to be mainly a history of the reign. But Sir Walter Raleigh suggests an explanation for this lack of biographical expression in the introduction to his *History of the World* (1614): "Whosoever, in writing a modern history, shall follow truth too near the heels, it may haply strike out his teeth"—as Sir John Hayward could testify, having been imprisoned in the Tower of London because his account (1599) of Richard II's deposition, two centuries earlier, had aroused Queen Elizabeth's anger.

17th and 18th centuries. In the 17th century the word biography was first employed to create a separate identity for this type of writing. That century and the first half of the 18th presents a busy and sometimes bizarre biograph-

ical landscape. It was an era of experimentation and preparation rather than of successful achievement. In the New World, the American Colonies began to develop a scattered biographical activity, none of it of lasting importance. France offers the celebrated *Letters* of the Marquise de Sévigné to her daughter, an intimate history of the Age of Louis XIV; numerous memoirs, such as those of Louis de Rouvroy, duc de Saint-Simon, and the acerbic ones of the Cardinal de Retz (1717); and the philosopher and critic Pierre Bayle's *Dictionnaire historique et critique* (1697), which was followed by specialized biographical collections and reference works. England saw an outpouring, beginning in the earlier 17th century, of Theophrastan "characters" (imaginary types imitated from the work of Theophrastus, a follower of Aristotle), journals, diaries, the disorganized but vivid jottings of John Aubrey (later published in 1898 as *Brief Lives*); and in the earlier 18th century there were printed all manner of sensational exposés, biographical sketches of famous criminals, and the like. In this era women appear for the first time as biographers. Lady Fanshawe wrote a life of her ambassador-husband (1829); Lucy Hutchinson, one of her Puritan warrior-husbands (written after 1664, published 1806); and Margaret Cavendish, duchess of Newcastle, produced a warm, bustling life—still good reading today—of her duke, an amiable mediocrity (*The Life of the thrice Noble Prince William Cavendishe, Duke Marquess, and Earl of Newcastle,* 1667). This age likewise witnessed the first approach to a professional biographer, the noted lover of angling, Izaak Walton, whose five lives (of the poets John Donne [1640] and George Herbert [1670], the diplomat Sir Henry Wotton [1651], and the ecclesiastics Richard Hooker [1665] and Robert Sanderson [1678]) tend to endow their diverse subjects with something of Walton's own genteel whimsicality but nonetheless create skillful biographical portraits. The masterpieces of the age are unquestionably Roger North's biographies (not published until 1742, 1744) of his three brothers: Francis, the lord chief justice, "my best brother"; the lively merchant-adventurer Sir Dudley, his favourite; and the neurotic scholar John. Also the author of an autobiography, Roger North likewise produced, as a preface to his life of Francis, the first extensive critical essay on biography, which anticipates some of the ideas of Samuel Johnson and James Boswell.

The first "professional" biographer

The last half of the 18th century witnessed the remarkable conjunction of these two remarkable men, from which sprang what is generally agreed to be the world's supreme biography, Boswell's *Life of Samuel Johnson LL.D.* (1791). Dr. Johnson, literary dictator of his age, critic and lexicographer who turned his hand to many kinds of literature, himself created the first English professional biographies in *The Lives of the English Poets.* In essays and in conversation, Johnson set forth principles for biographical composition: the writer must tell the truth—"the business of the biographer is often to . . . display the minute details of daily life," for it is these details that re-create a living character; and men need not be of exalted fame to provide worthy subjects.

For more than one reason the somewhat disreputable and incredibly diligent Scots lawyer James Boswell can be called the unique genius of biographical literature, bestriding both autobiography and biography. Early in his acquaintance with Johnson he was advised by the Doctor "to keep a journal of my [Boswell's] life, full and unreserved." Boswell followed this advice to the letter. His gigantic journals offer an unrivalled self-revelation of a fascinatingly checkered character and career—whether as a young rake in London or thrusting himself upon the aged Rousseau or making his way to Voltaire's seclusion at Ferney in Switzerland with the aim of converting that celebrated skeptic to Christianity. Boswell actively helped to stage the life of Johnson that he knew he was going to write—drawing out Johnson in conversation, setting up scenes he thought likely to yield rich returns—and thus, at moments, he achieved something like the novelist's power over his materials, being himself an active part of what he was to re-create. Finally, though he invented no new biographical techniques, in his *Life of Samuel Johnson* he interwove with consummate skill Johnson's letters and personal papers, Johnson's conversation as assiduously recorded by the biographer, material drawn from interviews with large numbers of people who knew Johnson, and his own observation of Johnson's behaviour, to elicit the living texture of a life and a personality. Boswell makes good his promise that Johnson "will be seen as he really was. . . ." The influence of Boswell's work penetrated throughout the world and, despite the development of new attitudes in biographical literature, has persisted to this day as a pervasive force. Perhaps equally important to life writers has been the inspiration provided by the recognition accorded Boswell's *Life* as a major work of literary art. Since World War II there have often been years, in the United States, when the annual bibliographies reveal that more books or articles were published about Johnson and Boswell than about all the rest of biographical literature together.

Boswell's influence on biography

19th century. The *Life of Johnson* may be regarded as a representative psychological expression of the Age of Enlightenment, and it certainly epitomizes several typical characteristics of that age: devotion to urban life, confidence in common sense, emphasis on man as a social being. Yet in its extravagant pursuit of the life of one individual, in its laying bare the eccentricities and suggesting the inner turmoil of personality, it may be thought of as part of that revolution in self-awareness, ideas, aspirations, exemplified in Rousseau's *Confessions,* the French Revolution, the philosophical writings of the German philosopher Immanuel Kant, the political tracts of Thomas Paine, and the works of such early Romantic poets as Robert Burns, William Blake, Wordsworth—a revolution that in its concern with the individual psyche and the freedom of man seemed to augur well for biographical literature. This promise, however, was not fulfilled in the 19th century.

That new nation, the United States of America, despite the stimulus of a robust and optimistic society, flamboyant personalities on the frontier, a generous share of genius, and the writing of lives by eminent authors such as Washington Irving and Henry James, produced no biographies of real importance. One professional biographer, James Parton, published competent, well-researched narratives, such as his lives of Aaron Burr and Andrew Jackson, but they brought him thin rewards and are today outmoded. In France, biography was turned inward, to romantic introspection, a trend introduced by Étienne Pivert de Senancour's *Obermann* (1804). It was followed by autobiographies thinly disguised as novels such as Benjamin Constant's *Adolphe* (1816), *La Vie de Henri Brulard* of Stendhal (Marie Henri Beyle), and similar works by Alphonse de Lamartine and Alfred de Musset, in which the emotional malaise of the hero is subjected to painstaking analysis. In Great Britain the 19th century opened promisingly with an outburst of biographical–autobiographical production, much of which came from prominent figures of the Romantic Movement, including Samuel Taylor Coleridge, Robert Southey, William Hazlitt, and Thomas De Quincey. Thomas Moore's *Letters and Journals of Lord Byron* (1830), John Gibson Lockhart's elaborate life (1837–38) of his father-in-law, Sir Walter Scott, and, later, Elizabeth Cleghorn Gaskells' *Life of Charlotte Bronte* (1857), James Anthony Froude's study of Carlyle (2 vol. 1882; 2 vol. 1884), John Forster's *Life of Charles Dickens* (1872–74) all followed, to some degree, what may loosely be called the Boswell formula. Yet most of these major works are marred by evasions and omissions of truth—though Lockhart and Froude, for example, were attacked as conscienceless despoilers of the dead—and, before the middle of the century, biography was becoming stifled. As the 20th century biographer and critic Sir Harold Nicolson wrote in *The Development of English Biography*, "Then came earnestness, and with earnestness hagiography descended on us with its sullen cloud . . .," 1927. Insistence on respectability, at the expense of candour, had led Carlyle to observe acridly, "How delicate, how decent is English biography, bless

Biography during the Victorian age

its mealy mouth!" and to pillory its productions as "vacuum-biographies."

Modern. The period of modern biography was ushered in, generally speaking, by World War I. All the arts were in ferment, and biographical literature shared in the movement, partly as a reaction against 19th-century conventions, partly as a response to advances in psychology, and partly as a search for new means of expression. This revolution, unlike that at the end of the 18th century, was eventually destined to enlarge and enhance the stature of biography. The chief developments of modern life writing may be conveniently classified under five heads: (1) an increase in the numbers and general competence of biographies throughout the Western world; (2) the influence on biographical literature of the counterforces of science and fictional writing; (3) the decline of formal autobiography and of biographies springing from a personal relationship; (4) the range and variety of biographical expression; and (5) the steady, though moderate, growth of a literature of biographical criticism. Only the first three of these developments need much elaboration.

Little has been said about biography since the Renaissance in Germany, Spain, Italy, Scandinavia, and the Slavic countries because, as in the case of Russia, there had been comparatively little biographical literature and because biographical trends, particularly since the end of the 18th century, generally followed those of Britain and France. Russian literary genius in prose is best exemplified during both the 19th and 20th centuries in the novel. In the 19th century, however, Leo Tolstoy's numerous autobiographical writings, such as *Childhood* and *Boyhood*, and Sergey Aksakov's *Years of Childhood* and *A Russian Schoolboy*, and in the 20th century, Maksim Gorky's autobiographical trilogy (*Childhood; In the World;* and *My Universities*, 1913–23) represent, in specialized form, a limited biographical activity. The close control of literature exercised by the 20th-century Communist governments of eastern Europe has created a wintry climate for biography. The rest of Europe, outside the iron curtain, has manifested in varying degrees the fresh biographical energies and practices illustrated in British–American life writing: biography is now, as never before, an international art that shares a more or less common viewpoint.

The second characteristic of modern biography, its being subject to the opposing pressures of science and fictional writing, has a dark as well as a bright side.

Influence of fiction on biography

Twentieth-century fiction, boldly and restlessly experimental, has, on the one hand, influenced the biographer to aim at literary excellence, to employ devices of fiction suitable for biographical ends; but, on the other, fiction has also probably encouraged the production of popular pseudobiography, hybrids of fact and fancy, as well as of more subtle distortions of the art form. Science has exerted two quite different kinds of pressure: the prestige of the traditional sciences, in their emphasis on exactitude and rigorous method, has undoubtedly contributed to a greater diligence in biographical research and an uncompromising scrutiny of evidences; but science's vast accumulating of facts—sometimes breeding the worship of fact for its own sake—has helped to create an atmosphere in which today's massive, note-ridden and fact-encumbered lives proliferate and has probably contributed indirectly to a reluctance in the scholarly community to take the risks inevitable in true biographical composition.

The particular science of psychology, as earlier pointed out, has conferred great benefits upon the responsible practitioners of biography. It has also accounted in large part, it would appear, for the third characteristic of modern biography: the decline of formal autobiography and of the grand tradition of biography resulting from a personal relationship. For psychology has rendered the self more exposed but also more elusive, more fascinatingly complex and, in the darker reaches, somewhat unpalatable. Since honesty would force the autobiographer into a self-examination both formidable to undertake and uncomfortable to publish, instead he generally turns his attention to outward experiences and writes memoirs and reminiscences—though France offers something of an exception in the journals of such writers as André Gide (1947–51), Paul Valéry (1957), François Mauriac (1934–50), Julien Green (1938–58). Similarly, psychology, in revealing the fallacies of memory, the distorting power of an emotional relationship, the deceits of observation, has probably discouraged biography written by a friend of its subject. Moreover, so many personal papers are today preserved that a life-long friend of the subject scarcely has time to complete his biography.

The impact of Lytton Strachey

After World War I, the work of Lytton Strachey played a somewhat similar role to that of Boswell in heading a "revolution" in biography. *Eminent Victorians* and *Queen Victoria* (1921), followed by *Elizabeth and Essex* (1928), with their artful selection, lacquered style, and pervasive irony, exerted an almost intoxicating influence in the 1920s and '30s. Writers seeking to capitalize on Strachey's popularity and ape Strachey's manner, without possessing Strachey's talents, produced a spate of "debunking" biographies zestfully exposing the clay feet of famous historical figures. By World War II, however, this kind of biography had been discredited; Strachey's adroit detachment and literary skill were recognized to be his true value, not his dangerously interpretative method; and, since that time, biography has steadied into an established, if highly varied, form of literature.

Other literatures. Biography as an independent art form, with its concentration upon the individual life and its curiosity about the individual personality, is essentially a creation of Western man. In the Orient, for all its long literary heritage, and in Islām, too, biographical literature does not show the development, nor assume the importance, of Western life writing. In China, until comparatively recently, biography had been an appendage, or by-product, of historical writing and scholarly preoccupation with the art of government, in the continuing tradition of the "Historical Records" of Ssuma Ch'ien and Pan Ku. In India it has been the enduring concern for spiritual values and for contemplation or mystical modes of existence that have exerted the deepest influence on literature from the first millennium BC to the present, and this has not provided a milieu suitable to biographical composition. Generally speaking, the literary history of Japan, too, offers only fragmentary or limited examples of life writing.

It was not until the beginning of the 20th century in China that biography began to appear as an independent form (and this was evidently the result of western influence), when Liang Ch'i-ch'ao (1873–1929) wrote a number of lives, including one of Confucius, and was followed by Hu Shih (1891–1962), who, like his predecessor, worked to promote biographical composition as an art form. Except for China after the establishment of the Communist state in 1949, biography in the Orient—notably in India and Japan—has shared, to a limited extent, the developments in biographical literature demonstrated in the rest of the world.

BIOGRAPHICAL LITERATURE TODAY

In the United States, Great Britain, and the rest of the Western world generally, biography today enjoys a moderate popular and critical esteem. In the year 1929, at the height of the biographical "boom," there were published in the United States 667 new biographies; in 1962 exactly the same number appeared, the population in the meantime having increased by something like 50 percent. On the average, in the English-speaking world, biographical titles account for approximately 5 percent of the annual output of books. Yet they have won their share of literary prizes and for their authors a considerable degree of literary eminence; if few universally acclaimed masterpieces are being produced, it is probably true that the art of biography is seeing a higher general level of achievement than ever before. The recreation of a life is also now being attempted in other media than that of prose. Biographical drama has of course been staged from before the time of Shakespeare; it continues

to be popular, whether translated from narrative to the theatre (as the *Diary of Anne Frank*) or written specifically for the stage, like Jean Anouilh's *Becket* and Robert Bolt's study of Sir Thomas More, *A Man for All Seasons* (which nonetheless owes a great deal to William Roper). The cinema often follows with its versions of such plays; it likewise produces original biographical films, generally with indifferent success. Television, too, offers historical "recreations" of various sorts, and with varying degrees of responsibility, but has achieved only a few notable examples of biographical illumination, for the conflict between gripping visual presentation and the often undramatic, but important, biographical truth is difficult to resolve. Biography, indeed, seems less innovative, less rewarding of experiment, and less adaptable to new media, than does fiction or perhaps even history. Words are no longer the only way to tell a story and perhaps in time will not be regarded as the chief way; but so far they seem to offer the best way of unfolding the full course of a life and exploring the quirks and crannies of a personality. Anchored in the truth of fact, though seeking the truth of interpretation, biography tends to be more stable than other literary arts; and its future would appear to be a predictably steady evolution of its present trends.

BIBLIOGRAPHY

Critical and scholarly books: JAMES L. CLIFFORD, *From Puzzles to Portraits: Problems of a Literary Biographer* (1970), examples of the author's own research followed by an analysis of biographical problems; LEON EDEL, *Literary Biography* (1959), essentially an account of the methods, psychological and narrative, used by the author in his multivolume life of Henry James; JOHN A. GARRATY, *The Nature of Biography* (1957), an historical survey coupled with a study of biographical methods, with emphasis on aids offered by psychology; PAUL M. KENDALL, *The Art of Biography* (1965), an historical survey, with emphasis on contemporary biography, and a study of biographical problems from the viewpoint of a practicing biographer; ANDRE MAUROIS, *Aspects de la biographie* (1928; Eng. trans., 1930) and HAROLD NICOLSON, *The Development of English Biography* (1928), particularly interesting for complementary views of the "new" biography of the 1920s by two eminent biographers; ROY PASCAL, *Design and Truth in Autobiography* 1960), an historical survey and a study of the chief problems, aspects, and varieties of autobiography.

Anthologies: JAMES L. CLIFFORD (ed.), *Biography As an Art: Selected Criticism 1560–1960* (1962); WILLIAM H. DAVENPORT and BEN SIEGEL (eds.), *Biography Past and Present* (1965), contains a number of critical essays as well as biographical selections; EDGAR JOHNSON (ed.), *A Treasury of Biography* (1941); JOHN C. METCALFE (ed.), *The Stream of English Biography* (1930).

(P.M.K.)

Biological Sciences

Biology may be defined as an area of learning that deals with all of the physicochemical aspects of life. But as a result of the modern tendency to unify scientific knowledge and investigation, there has been an overlapping of the field of biology with other scientific disciplines. Modern principles of other sciences, chemistry and physics, for example, are integrated with those of biology in such areas as biochemistry and biophysics.

Because biology is such a broad subject, it is subdivided into separate branches for convenience of study. Despite apparent differences, however, all the subdivisions are interrelated by basic principles that underly all biological manifestations. Thus, although it was once the custom to separate the study of plants (botany) from that of animals (zoology), and the study of the structure of organisms (morphology) from that of function (physiology), the current practice is to investigate those biological phenomena that all living things have in common.

The main purpose of this article is to serve as an overview of the organization of the biological sciences; the reader will find in it references to other articles that deal with specific topics in greater detail. This article is divided into the following major sections:

I. The nature and scope of the biological sciences
General features
The scope of biology
II. The history of biology
The importance of experimentation and instrumentation
The origin and early development of biological ideas
Advances to the 20th century
Biology in the 20th century
III. The philosophy of biology
Methods of study
Schools of biophilosophical thought
Redefining the role of the biologist

I. The nature and scope of the biological sciences

GENERAL FEATURES

The contemporary approach to the study of living things is based on the levels of biological organization involved —whether molecules, cells, individuals, or populations— in the specific subject matter under investigation—structure, function, behaviour, etc.—and on the common principles that unify the apparently disparate manifestations of life—homeostasis, unity, evolution, etc.

Levels of biological organization. Biology is often approached today on the basis of levels that deal with fundamental units of life. At the level of molecular biology, for example, life is regarded as a manifestation of chemical and energy transformations that occur among the many chemical constituents that comprise an organism. As a result of the development of more powerful and precise laboratory instruments and techniques, it is now possible to understand and define more exactly not only the invisible ultimate physiochemical organization (ultrastructure) of the molecules in living matter but also how living matter reproduces at the molecular level.

Fundamental units of life

Cell biology, the study of the fundamental unit of structure and function in a living organism, may be said to have begun in the 17th century, with the invention of the compound microscope. Before that, the individual organism was studied as a whole (organismic biology), an area of research still regarded as an important level of biological organization. Population biology deals with groups or populations of organisms that inhabit a given area or region. Included at this level are studies of the roles that specific kinds of plants and animals play in the complex and self-perpetuating interrelationships that exist between the living and nonliving world, as well as studies of the built-in controls that maintain these relationships naturally.

Biological specialties. Because each of the above levels is in itself too broad to be grasped by any one individual, there are specialists whose fields of investigation may be one of the following major biological subdivisions:

1. Morphology, the study of the structure of plants and animals, is divided into anatomy, the study of structure that can be observed with the naked eye; histology, the study of microscopic structure; and cytology, the study of the particulate minutiae of cellular structure.
2. Physiology is the study of the functions of cells, tissues, organs, and systems in living organisms. Although modern physiology concentrates on events at the cellular level, it traditionally has included all types of organismal functioning, from the manner in which food is digested in man to the way a plant absorbs water from the soil.
3. Taxonomy deals with the classification of all living things; it attempts to group organisms according to observed natural or hypothetical relationships or both.
4. Embryology is concerned with the stages in the formation and development of the embryo in plants and animals. It is usually limited to the part of an organism's life history that extends from the union of the reproductive cells—egg and spermatozoon—to the completion of its body structure.
5. Genetics is the study of inheritance and variation in organisms and the mechanisms by which these processes operate and are controlled.
6. Evolution is the study of the possible origins of living things, how they have changed, and the possible methods by which these changes have occurred.
7. Paleontology, which is closely related to evolution, is the study of life as it existed in past geological times; it is based mainly on fossil records of prehistoric plant and animal life.

8. Ecology is the study of the relationships of plants and animals to each other and to their environment.

As pointed out earlier, other biological disciplines that constitute separate areas of study include biochemistry, biophysics, and molecular biology. Moreover, some areas of biology are especially concerned with the investigation of one kind of living thing—*e.g.*, botany, the study of plants; zoology, the study of animals; ornithology, the study of birds; ichthyology, the study of fishes; mycology, the study of fungi; microbiology, the study of microorganisms; protozoology, the study of one-celled animals; herpetology, the study of amphibians and reptiles; and entomology, the study of insects.

Biological principles. *Homeostasis.* The concept of homeostasis—*i.e.*, that all living things maintain a constant internal environment—was first suggested by Claude Bernard, a 19th-century French physiologist, who stated that "all the vital mechanisms, varied as they are, have only one object: that of preserving constant the conditions of life. . ." (see HOMEOSTASIS).

As originally conceived by Bernard, homeostasis applied to the struggle of a single organism to survive. The concept was later extended to include any biological system from the cell to the entire biosphere, all the areas of the Earth inhabited by living things (see BIOSPHERE).

Unity. All living organisms, regardless of their uniqueness, have certain biological, chemical, and physical characteristics in common. All, for example, are composed of the same basic units, or cells, and the same chemical substances, which, when analyzed, exhibit noteworthy similarities, even in such disparate organisms as bacteria and man. Furthermore, since the action of any organism is determined by the manner in which its cells interact and since all cells interact in much the same way, the basic functioning of all organisms is also similar.

There is not only unity of basic living substance and functioning but also unity of origin of all living things. According to a theory proposed in 1855 by Rudolf Virchow, a German pathologist, "all living cells arise from pre-existing living cells." This theory appears to be true for all living things at the present time under existing environmental conditions. If, however, life originated more than once in the past, the fact that all organisms have a sameness of basic structure, composition, and function would seem to indicate that only one original type succeeded.

Similarities between man and slime molds

A common origin of life would explain why in man or slime mold—and in all forms of life in between—the same chemical substance, deoxyribonucleic acid (DNA), in the form of genes accounts for the ability of all living matter to replicate itself exactly and to transmit genetic information from parent to offspring. Furthermore, the mechanisms for this transmittal follow a pattern that is the same in all organisms (see HEREDITY). (For more information concerning the composition and structure of DNA, see NUCLEIC ACID; for a description of the manner in which DNA operates, see GENE.)

Whenever a change in a gene (a mutation) occurs, there is a change of some kind in the organism that contains the gene. It is this universal phenomenon that gives rise to the differences (variations) in populations of organisms from which nature selects for survival those that are best able to cope with changing conditions in the environment.

Evolution. In his theory of natural selection, which is discussed in greater detail later, Charles Darwin suggested that "survival of the fittest" was the basis for organic evolution (the modification of living things with time). Evolution itself is a biological phenomenon common to all living things, even though it has led to their differences. Evidence to support the theory of evolution has come primarily from the fossil record, from comparative studies of structure and function, and from studies of embryological development (see EVOLUTION; FOSSIL RECORD; DEVELOPMENT, BIOLOGICAL).

Diversity. Despite the basic biological, chemical, and physical similarities found in all living things, a diversity of life exists not only among and between species but also within every natural population. The phenomenon of diversity has had a long history of study because so many of the variations that exist in nature are visible to the eye. The fact that organisms changed during prehistoric times and that new variations are constantly evolving can be verified by paleontological records as well as by breeding experiments in the laboratory. Long after Darwin had assumed that variations existed, biologists discovered that they are caused by a change in the genetic material (DNA). This change can be a slight alteration in the sequence of the constituents of DNA (nucleotides), a larger change such as a structural alteration of a chromosome, or a complete change in the number of chromosomes. In any case, a change in the genetic material in the reproductive cells manifests itself as some kind of structural or chemical change in the offspring. The consequence of such a mutation depends upon the interaction of the mutant offspring with its environment.

Adjusting to changing conditions

It has been suggested that sexual reproduction became the dominant type of reproduction among organisms because of its inherent advantage of variability, which is the mechanism that enables a species to adjust to changing conditions. New variations are potentially present in genetic differences, but how preponderant a variation becomes in a gene pool depends upon the number of offspring the mutants or variants produce (differential reproduction). It is possible for a genetic novelty (new variation) to spread in time to all members of a population, especially if the novelty enhances the population's chances for survival in the environment in which it exists. Thus, when a species is introduced into a new habitat, it either adapts to the change by natural selection or by some other evolutionary mechanism or else it eventually dies off. Because each new habitat means new adaptations, habitat changes have been responsible for the millions of different kinds of species and for the heterogeneity within each species.

The total number of animal and plant species is estimated at between 2,000,000 and 4,500,000; authoritative estimates of the number of extinct species range from 15,000,000 up to 16,000,000,000. Although the use of classification as a means of producing some kind of order out of this staggering number of different types of organisms appears as early as the book of Genesis—with references to cattle, beasts, fowl, creeping things, trees, etc.—the first scientific attempt at classification is attributed to the Greek philosopher Aristotle, who tried to establish a system that would indicate the relationship of all things to each other. He arranged everything along a scale, or "ladder of nature," with nonliving things at the bottom; plants were placed below animals, and man was at the top. Other schemes that have been used for grouping species include large anatomical similarities, such as wings or fins, which indicate a natural relationship, and also similarities in reproductive structures.

At the present time taxonomy is based on two major assumptions: one is that similar body construction can be used as a criterion for a classification grouping; the other that, in addition to structural similarities, evolutionary and molecular relationships between organisms can be used as a means for determining classification (see CLASSIFICATION, BIOLOGICAL).

Behaviour and interrelationships. As was mentioned earlier, the study of the relationships of living things to each other and to their environment is known as ecology. Because these interrelationships are so important to the welfare of Earth and because they can be seriously disrupted by man's activities, ecology is becoming one of the most important branches of biology (see ECOLOGY).

Continuity. Whether an organism is man or a bacterium, its ability to reproduce is one of the most important characteristics of life. Because life comes only from preexisting life, it is only through reproduction that successive generations can carry on the properties of a species.

THE SCOPE OF BIOLOGY

The study of structure. Living things are defined in terms of the activities or functions that are missing in nonliving things. The life processes of every organism are carried out by specific materials assembled in definite

structures. Thus, a living thing can be defined as a system, or structure, that reproduces, changes with its environment over a period of time, and maintains its individuality by constant and continuous metabolism. This pattern of action or function results from and occurs in a pattern of organization.

Cells and their constituents. Knowledge of the structure and function of the cell has resulted from technological developments and methods.

Biologists once depended on the light microscope to study the morphology of cells found in higher plants and animals. Based on descriptions of structure, biochemists and physiologists studied and postulated the functioning of cells in unicellular and in multicellular organisms; the discovery of the chloroplastids in the cell, for example, led to the investigation of the process of photosynthesis. With the discovery of the electron microscope, the fine organization of the plastids could be utilized for further quantitative studies of the different parts of this process.

Quantitative studies make use of histochemistry to identify proteins, carbohydrates, and other chemical constituents of cells. Histochemistry has also been used to identify RNA and DNA in various cell parts.

Tracing the movement of substances

A valuable method useful in tracing the movement of substances in living matter is radioautography: when radioactive nutrients, which can be incorporated into cells, are injected into animals, they give off detectable rays by which their presence and location can be determined. Thymidine, for example, can be made radioactive and, when injected, becomes part of the DNA being synthesized in the nucleus before cell division; the nuclei then can be identified by their radioactivity and the process of the origin of new DNA studied. Radioautography has been used to locate the site of protein synthesis and enzyme storage in cells.

Advanced technological developments—the microspectrophotometer, the X-ray probe, the laser beam, the computer, the stereoscopic microscope, the quartz-fibre microbalance, and television microscopy—are used to study the action of enzymes in living cells. The elucidation of such processes as lipid synthesis, active transport of large particles from the blood into cells, and the continuous formation of taste cells has been dependent on similar instrumentation.

Tissues and organs. Early biologists viewed their work as a study of the organism. The organism, then considered the fundamental unit of life, is still the prime concern of some modern biologists, and the maintenance of organisms is still an important part of biological research.

In 1912 an experiment showed that cells can be kept alive indefinitely if proper conditions are maintained. Utilizing stringent laboratory techniques, workers have kept bits of chicken heart tissue alive for over 30 years. Techniques for keeping organs alive in preparation for transplants stem from such experiments.

In modern biological research it is necessary to deal with the study of structure and function at all levels of biological organization from the molecule to the organism. Electronics, mathematics, and computers have become increasingly important in solving problems at all of these levels.

The study of function. To maintain life, an organism not only repairs or replaces (or both) its structures by a constant supply of the materials of which it is composed but also keeps its life processes in operation by a steady supply of energy. The initial source of this energy is the environment outside of the organism. The process by which the organism provides the necessary raw materials for the continuation of life is called nutrition (*q.v.*). Plants obtain their nutrients from water, from minerals, and from the carbohydrates they manufacture. Animals, which cannot manufacture their own food, need at least the following kinds of nutrients: water, minerals, organic carbon, organic nitrogen, vitamins, certain amino acids, and fatty acids.

Many experiments have been directed toward solving the problem of biological differentiation. It has been determined that, although all genes of an organism are present in every cell, they do not all act at the same time: some genes act only at certain times during development; others never act in some cells. Whether a gene is active is sometimes the result of an interaction between cells. Cells seem to develop differently in different locations. How this is controlled is not definitely known; one possibility is the presence of an electrical communication between cells or of a substance that diffuses out of the cell. The latter idea is suggested by experiments demonstrating that the formation of the tissues of organs such as the eye, kidney, and liver are directly influenced by the tissues bordering them. Many of these experiments make use of tissue culture techniques, which permit the growth of cells outside of the body. It is possible to grow a single embryonic muscle cell into a colony of differentiated muscle. It is through such experiments that the questions about development and its implications may eventually be answered. (E.R.G.)

II. The history of biology

THE IMPORTANCE OF EXPERIMENTATION AND INSTRUMENTATION

The importance of creative minds and technology

There are moments in the history of all sciences when remarkable progress is made in relatively short periods of time. Such leaps in knowledge result in great part from two factors: one is the presence of a creative mind —a mind sufficiently perceptive and original to discard hitherto accepted ideas and formulate new hypotheses; the second is the technological ability to test the hypotheses by appropriate experiments. The most original and inquiring mind is severely limited without the proper tools to conduct an investigation; conversely, the most sophisticated technological equipment cannot of itself yield insights into any scientific process.

An example of the relationship between these two factors was the discovery of the cell. For hundreds of years there had been speculation concerning the basic structure of both plants and animals. Not until optical instruments were sufficiently developed to reveal cells, however, was it possible to formulate a general hypothesis, the cell theory, that satisfactorily explained how plants and animals are organized. Similarly, the significance of Gregor Mendel's studies on the mode of inheritance in the garden pea remained neglected for many years, until technological advances made possible the discovery of the chromosomes and the part they play in cell division and heredity. Moreover, as a result of the relatively recent development of extremely sophisticated instruments, such as the electron microscope and the ultracentrifuge, biology has moved from being a largely descriptive science—one concerned with entire cells and organisms—to a discipline that increasingly emphasizes the subcellular and molecular aspects of organisms and attempts to equate structure with function at all levels of biological organization.

THE ORIGIN AND EARLY DEVELOPMENT OF BIOLOGICAL IDEAS

Although it is not known when the study of biology originated, early man must have had some knowledge of the animals and plants around him. His very survival depended upon the accurate recognition of nonpoisonous food plants and upon an understanding of the habits of dangerous predators. Archaeological records indicate that even before the development of civilization, man had domesticated virtually all the amenable animals available to him and had developed an agricultural system sufficiently stable and efficient to satisfy the needs of large numbers of people living together in communities. It is clear, therefore, that much of the history of biology predates the time at which man began to write and to keep records.

Earliest biological records. *Biological practices among Assyrians and Babylonians.* Much of the earliest recorded history of biology is derived from bas-reliefs the Assyrians and Babylonians made of their cultivated plants and from carvings depicting their veterinary medicine. Illustrations on certain seals reveal that the Babylonians had learned that the date palm reproduces sex-

ually and that pollen could be taken from the male plant and used to fertilize female plants. Although a precise dating of these early records is lacking, a Babylonian business contract of the Hammurabi period (*c.* 1800 BC) mentions the male flower of the date palm as an article of commerce, and descriptions of date harvesting date back to about 3500 BC.

Another source of information concerning the extent of biological knowledge of these early peoples was the discovery of several papyri that pertain to medical subjects; one, believed to date back to 1600 BC, contains anatomical descriptions; another (*c.* 1500 BC) indicates that the importance of the heart had been recognized. Because these ancient documents, which contained mixtures of fact and superstition, probably summarized then-current knowledge, it may be assumed that some of their contents had been known by earlier generations.

Biological knowledge of Egyptians, Chinese, and Hindus. Papyri and artifacts found in tombs and pyramids indicate that the Egyptians also possessed considerable medical knowledge. Their well-preserved mummies demonstrate that they had a thorough understanding of the preservative properties of herbs required for embalming; plant necklaces and bas-reliefs from various sources also reveal that the ancient Egyptians were well aware of the medicinal value of certain plants 2,000 years before Christ. Even earlier (*c.* 2800 BC), a work now ascribed to the Chinese emperor Shen Nung described the therapeutic powers of numerous medicinal plants and included descriptions of many important food plants, such as the soybean. Furthermore, the ancient Chinese not only utilized the silkworm *Bombyx mori* to produce silk for commerce but also understood the principle of biological control, employing one type of insect, an entomophagous (insect-eating) ant, to destroy insects that bored into trees.

Early use of medicinal plants

As early as 2500 BC the Hindus of India had a well-developed science of agriculture. The ruins at Mohenjodaro have yielded seeds of wheat and barley that were cultivated at this time. Millet, dates, melons, and other fruits and vegetables, as well as cotton, were known to this civilization. Plants were not only a source of food, however. A Hindu document, believed to date back to the 6th century BC, described the use of about 960 medicinal plants and included information on such topics as anatomy, physiology, pathology, and obstetrics.

Biology in the Greco-Roman world. Although the Babylonians, Assyrians, Egyptians, Chinese, and Indians amassed much biological information, they lived in a world believed to be dominated by unpredictable demons and spirits. Hence, learned men in these early cultures directed their studies toward an understanding of the supernatural, rather than the natural, world. Anatomists, for example, dissected animals not to gain an understanding of their structure but to study their organs in order to predict the future. With the emergence of the Greek civilization, however, these mystical attitudes began to change. Around 600 BC there arose a school of Greek philosophers who believed that every event has a cause and that a particular cause produces a particular effect. This concept, known as causality, had a profound effect on subsequent scientific investigation. Furthermore, these philosophers assumed the existence of a "natural law" that governs the universe and can be comprehended by men through the use of his powers of observation and deduction. Although they established the science of biology, the greatest contribution the Greeks made to science was the idea of rational thought.

Theories about man and the origin of life. One of the earliest Greek philosophers, Thales of Miletus (*c.* 7th century BC), maintained that the universe contained a creative force that he called physis, an early progenitor of the term physics; he also postulated that the world and all living things in it were made from water. Anaximander, a student of Thales, did not accept water as the only substance from which living things were derived; he believed that in addition to water, living things consisted of earth and a gaslike substance called *apeiron*, which could be divided into hot and cold. Various mixtures of these materials gave rise to the four elements: earth, air, fire, and water. Although he was one of the first to describe the Earth as a sphere rather than as a flat plane, Anaximander proposed that life arose spontaneously in mud and that the first animals to emerge had been fishes covered with a spiny skin. The descendants of these fishes eventually left water and moved to dry land, where they gave rise to other animals by transmutation (the conversion of one form into another). Thus, an early evolutionary theory was formulated.

Early evolutionary theory

At Crotone in southern Italy, where an important school of natural philosophy was established by Pythagoras about 500 BC, one of his students, Alcmaeon, investigated animal structure and described the difference between arteries and veins, discovered the optic nerve, and recognized the brain as the seat of the intellect. As a result of his studies of the development of the embryo, Alcmaeon may be considered the founder of embryology.

Although the Greek physician Hippocrates, who established a school of medicine on the Aegean island of Cos around 400 BC, was not an investigator in the sense of Alcmaeon, he did recognize through observations of patients the complex interrelationships involved in the human body. He also understood how the environment can influence human nature and suggested that sharply contrasting climates tend to produce a powerful type of inhabitant, while an even, temperate climate is conducive to indolence.

Hippocrates and his predecessors were all concerned with the central philosophical question of how the cosmos and its inhabitants were created. Although they accepted the physis as the creative force, they differed with regard to the importance of the roles played by earth, air, fire, water, and other elements. Although Anaximenes, for example, who may have been a student of Anaximander, adhered to the then-popular precept that life originated in a mass of mud, he postulated that the actual creative force was to be found in the air and that it was influenced by the heat of the Sun. Members of the Hippocratic school also believed that all living bodies were made up of four humours—blood, black bile, phlegm, and yellow bile—which supposedly originated in the heart, spleen, brain, and liver, respectively. An imbalance of the humours was thought to cause an individual to be sanguine, melancholy, phlegmatic, or choleric. The persistence of these words in current vocabulary attests to the lengthy popularity of the idea of humoral influences. For centuries it was also believed that an imbalance in the humours was the cause of disease, a belief that resulted in the common practice of bloodletting to get rid of excessive humours.

Aristotelian concepts. Around the middle of the 4th century BC, ancient Greek science reached a climax with Aristotle, who was interested in all branches of knowledge, including biology. Using his own observations and theories, Aristotle was the first to attempt a system of animal classification, in which he contrasted animals containing blood with those that were bloodless. The animals with blood included those now grouped as mammals (except the whales, which he placed in a separate group), birds, amphibians, reptiles, and fishes. The bloodless animals were divided into the cephalopods, the higher crustaceans, the insects, and the testaceans, the last group being a collection of all the lower animals. His careful examination of animals led to the understanding that mammals have lungs, breathe air, are warm-blooded, and suckle their young. Aristotle was the first to show any understanding of an overall systematic taxonomy and to recognize units of different degrees within the system.

Aristotle's classification of animals

The most important part of Aristotle's work was that devoted to reproduction and the related subjects of heredity and descent. He identified four means of reproduction, including the abiogenetic origin of life from nonliving mud, a belief held by Greeks of that time. Other modes of reproduction recognized by him included budding (asexual reproduction), sexual reproduction without copulation, and sexual reproduction with copu-

lation. Aristotle described sperm and ova and believed that the menstrual blood of viviparous organisms (those that give birth to living young) was the actual generative substance.

Biological principles formulated by Aristotle

Although Aristotle recognized that species are not stable and unalterable and although he attempted to classify the animals he observed, he was far from developing any pre-Darwinian ideas concerning evolution. In fact, he rejected any suggestion of natural selection and sought teleological explanations (*i.e.*, all phenomena in nature are shaped by a purpose) for any given observation. Nevertheless, many important scientific principles, some of which are often thought of as 20th-century concepts, can be ascribed to Aristotle. The following are a few such: (1) Using birds as an example, he formulated the principle that all organisms are structurally and functionally adapted to their habits and habitats. (2) Nature is parsimonious; it does not expend unnecessary energy. (3) In classifying animals, Aristotle rejected the idea of dividing them solely by their external structures (*e.g.*, animals with wings and those without wings). He recognized instead a basic unity of plan among diverse organisms, a principle that is still conceptually and scientifically sound. Further, Aristotle also believed that the entire living world could be described as a unified organization rather than as a collection of diverse groups. (4) By his observations, Aristotle realized the importance of structural homology, basically similar organs in different animals, and functional analogy, different structures that serve somewhat the same function—*e.g.*, the hand, claw, and hoof are analogous structures. These principles constitute the basis for the biological field of study known as comparative anatomy. (5) Aristotle's observations also led to the formulation of the principle that general structures appear before specialized ones and that tissues differentiate before organs.

Botanical investigations. Of all the works of Aristotle that have survived, none deals with what was later differentiated as botany, although it is believed that he wrote at least two treatises on plants. Fortunately, however, the work of Theophrastus, one of Aristotle's students, has been preserved to represent plant science of the Greek period. Like Aristotle, Theophrastus was a keen observer, although his works do not express the depth of original thought exemplified by his teacher. In his great work, *De historia et causis plantarum* (*The Calendar of Flora*, 1761), in which the morphology, natural history, and therapeutic use of plants are described, Theophrastus distinguished between the external parts, which he called organs, and the internal parts, which he called tissues. This was an important achievement because Greek scientists of this period had no established scientific terminology by which a specific structure could be referred to with a scientific term. For this reason, both Aristotle and Theophrastus were obliged to write very long descriptions of structures that can be described rapidly and simply today. Because of this difficulty, Theophrastus sought to develop a scientific nomenclature by giving special meaning to words that were then in more or less current use; for example, *karpos* for fruit and *perikarpion* for seed vessel.

Development of scientific terminology

Although he did not propose an overall classification system for plants, over 500 of which are mentioned in his writings, Theophrastus did unite many species into what are now considered genera. In addition to writing the earliest detailed description of how to pollinate the date palm by hand and the first unambiguous account of sexual reproduction in flowering plants, he also recorded observations on seed germination and development.

Post-Grecian biological studies. With Aristotle and Theophrastus, the great Greek period of scientific investigation came to an end. The most famous of the new centres of learning were the library and museum in Alexandria. From 300 BC until around the time of Christ all significant biological advances were made by physicians at Alexandria. One of the most outstanding of these men was Herophilus, who dissected human bodies and compared their structures to those of other large mammals. He recognized the brain, which he described in detail, as the centre of the nervous system and the seat of intelligence. Based on his knowledge, he wrote a general anatomical treatise, a special one on the eyes, and a handbook for midwives.

Erasistratus, a younger contemporary and reputed rival of Herophilus who also worked at the museum in Alexandria, studied the valves of the heart and the circulation of blood. Although he was wrong in supposing that blood flows from the veins into the arteries, he was correct in assuming that small interconnecting vessels exist. He thus suspected (but did not see) the presence of capillaries; he thought, however, that the blood changed into air, or *pneuma*, when it reached the arteries, to be pumped throughout the body.

Perhaps the last of the ancient biological scientists of note was Galen, a Greek physician born in Asia Minor who practiced in Rome around the middle of the 2nd century AD. His early years were spent as a surgeon at the gladiatorial arena, which gave him the opportunity to observe details of human anatomy. But this was an age when it was considered improper to dissect human bodies, and, as a result, detailed study was not possible. Thus, although Galen's research on animals was thorough, his knowledge of human anatomy was faulty. Because his work was extensive and clearly written, Galen's writings, nevertheless, dominated medicine for centuries to come.

The Middle Ages. After Galen there were no further biological investigations for many centuries. It is sometimes claimed that the rise of Christianity was the cause of the decline in science; this, however, is not a tenable viewpoint, for science was already virtually dead by the end of the 2nd century AD, a time when Christianity was still an obscure sect. It is true, however, that the rise of Christianity did not favour the questioning attitude of the Greeks.

The decline of science

Arab domination of biology. During the almost 1,000 years that science was dormant in Europe, the Arabs, who by the 9th century had extended their sphere of influence as far as Spain, became the custodians of science and dominated biology, as they did other disciplines. At the same time, as the result of a revival of learning in China, new technical inventions flowed from there to the West. The Chinese had discovered how to make paper and how to print from movable type, two achievements that were to have an inestimable effect upon learning. Another important advance that also occurred during this time was the introduction into Europe from India of the so-called Arabic numerals.

From the 3rd until the 11th centuries biology was essentially an Arab science. Although they themselves were not great innovators, they discovered the works of such men as Aristotle and Galen, translated them into Arabic, studied them, and wrote commentaries about them. Of the Arab biologists, al-Jāḥiẓ, who died about 868, is particularly noteworthy. Among his biological writings is *Kitāb al-ḥayawān* ("Book of Animals"), which, although revealing some Greek influence, is primarily an Arabic work. In it, the author emphasized the unity of nature and recognized relationships between different groups of organisms. Because al-Jāḥiẓ believed that the Earth contained both male and female elements, he found the Greek doctrine of spontaneous generation (life emerging from mud) to be quite reasonable.

Ibn Sīnā, or Avicenna as he is better known, was an outstanding Arab scientist around the beginning of the 11th century; he was the true successor to Aristotle. His writings on medicine and drugs, which were particularly authoritative and remained so until the Renaissance, did much to bring the works of Aristotle back to Europe, where they were translated into Latin from Arabic.

Development of botany and zoology. During the 12th century the growth of biology was sporadic. Nevertheless, it was during this time that botany was developed from the study of plants with healing properties; similarly, from veterinary medicine and the pleasures of the hunt came zoology. Because of the interest in medicinal plants, herbs in general began to be described and illustrated in a realistic manner. Although Arabic science was

well developed during this period and was far in advance of Latin, Byzantine, and Chinese cultures, it began to show signs of decline. Latin learning, on the other hand, rapidly increasing, was best exemplified perhaps by a mid-13th-century German scholar, St. Albertus Magnus (Albert the Great), who was probably the greatest naturalist of the Middle Ages. His biological writings (*De Vegetabilibus*, seven books, and *De Animalibus*, 26 books) were based on the classical Greek authorities, predominantly Aristotle. But in spite of this classical basis, a significant amount of his work contained new observations and facts; for example, he described with great accuracy the leaf anatomy and venation of the plants he studied.

Albert was particularly interested in plant propagation and reproduction and discussed in some detail the sexuality of plants and animals. Like his Greek predecessors, he believed in spontaneous generation; he also believed that animals were more perfect than plants because they required two individuals for the sexual act. Perhaps one of Albert's greatest contributions to medieval biology was the denial of many superstitions believed by his contemporaries, a skepticism that, together with the reintroduction of Aristotelian biology, was to have profound effects on subsequent European science.

One of Albert's pupils was St. Thomas Aquinas, who endeavoured to reconcile Aristotelian philosophy and the teachings of the church. Because Aquinas was a rationalist, he declared that God created the reasoning mind; hence, by true intellectual processes of reasoning, man could not arrive at a conclusion that was in opposition to Christian thought. Acceptance of this philosophy made possible a revival of rational learning that was constant with Christian belief.

Revival of rational learning

Revitalization of anatomy. Italy, during the Middle Ages, became the most active scientific centre, although its major interests were concentrated on agriculture and medicine. A development of particular significance at this time was the introduction of dissection into medical schools, a step that revitalized the study of anatomy. Because of what it reveals about medieval anatomy in general, the work of Mondino dei Liucci, the most famous of the Italian anatomists at the beginning of the 14th century, is particularly important. First, because there was no way of preserving cadavers, organs that spoiled quickly had to be dissected rapidly. Furthermore, it was the custom for the teacher to leave the actual dissection to an underling, who, not wishing to offend the teacher, agreed with all of his statements. Thus, although Mondino performed all of his own dissections and, from his observations, could have corrected the errors of the Greeks and Arabs, he did not choose to contradict any of the authorities. Even when the authorities contradicted themselves, Mondino sought to harmonize their views. Perhaps Mondino exemplifies the difficulty that was so characteristic of the era; namely, the problem of breaking away from established authority.

The Renaissance. *Resurgence of biology.* Beginning in Italy during the 14th century there was a general ferment within the culture itself, which, together with the rebirth of learning (partly as a result of the rediscovery of Greek work), is referred to as the Renaissance. Interestingly, it was the artists, rather than the professional anatomists, who were intent upon a true rendering of the bodies of animals and men and thus were motivated to gain their knowledge firsthand by dissection. No individual better exemplifies the Renaissance than Leonardo da Vinci, whose anatomical studies of the human form during the late 1400s and early 1500s were so far in advance of the age that they included details not recognized until a century later. Furthermore, while dissecting animals and examining their structure, Leonardo compared them to the structure of man. In doing so he was the first to indicate the homology between the arrangements of bones and joints in the leg of the human and that of the horse, despite the superficial differences. Homology was to become an important concept in uniting outwardly diverse groups of animals into distinct units, a factor that is of great significance in the study of evolution.

Da Vinci's discovery of homologous structures

Other factors had a profound effect upon the course of biology in the 1500s, particularly the introduction of printing around the middle of the century, the increasing availability of paper, and the perfected art of the wood engraver, all of which meant that illustrations as well as letters could be transferred to paper. In addition, after the Turks had conquered Byzantium in 1453, many Greek scholars took refuge in the West; the scholars of the West thus had direct access to the scientific works of antiquity, rather than indirect access through Arabic translations.

Advances in botany. Otto Brunfels, the German theologian and botanist, published in 1530 a book about medicinal herbs, *Herbarum vivae eicones*, which, with its fresh and vigorous illustrations, contrasted sharply with earlier texts, whose authors had been content merely to copy from old manuscripts. In addition to books on the same subject, Hieronymus Bock (latinized to Tragus) and Leonhard Fuchs also published around the mid-1500s descriptive, well-illustrated texts about common wild flowers. The books published by the three men, who are often referred to as the German fathers of botany, may be considered the forerunners of modern botanical floras (treatises on or lists of the plants of an area or period).

Throughout the 16th century, interest in botanical study also existed in such other countries as the Netherlands, Switzerland, Italy, and France. During this time there was a great improvement in the classification of plants, which had been described in ancient herbals merely as trees, shrubs, or plants and, in later books, were either listed alphabetically or arranged in some arbitrary grouping. Now it was realized that there had to be some systematic method to designate the increasing number of plants being described. Accordingly, using a binomial system very similar to modern biological nomenclature, Gaspard Bauhin, a Swiss botanist of the late 16th and early 17th centuries, designated plants by a generic and a specific name. Although affinities between plants were indicated by the use of common generic names, Bauhin did not speculate on their common kinship.

Bauhin's botanical nomenclature

Pierre Belon, a French naturalist who travelled extensively in the Middle East, where he studied the flora, illustrates the wide interest of the 16th-century biologists. Although his botanical work was limited to two volumes, one on trees and one on horticulture, his books on travel included numerous biological entries, and his two books on fishes reveal much about the current state of systematics, including not only fishes but also such other aquatic creatures as mammals, crustaceans, mollusks, and worms. In his *L'Histoire de la nature des oyseaux* ("Natural History of Birds"), however, in which Belon's taxonomy was remarkably similar to that being used today, he showed a clear grasp of comparative anatomy, particularly of the skeleton, publishing the first picture of a bird skeleton beside a human skeleton to point out the homologies. Numerous other European naturalists who travelled extensively also brought back accounts of exotic animals and plants, and most of them wrote voluminous records of their excursions. Two other factors contributed significantly to the development of botany at this time: first was the establishment of botanical gardens by the universities, as distinct from the earlier gardens that had been established for medicinal plants; second was the collection of dried botanical specimens, or herbaria.

It is perhaps surprising that the great developments in botany during the 16th century had no parallel in zoology. Instead, there arose a group of biologists known as the Encyclopedists, best represented by Conrad Gesner, a 16th-century Swiss naturalist, who compiled books on animals that were illustrated by some of the finest artists of the day (Albrecht Dürer, for example). But because the descriptions of many of the animals were grossly inaccurate, in many cases continuing the legends of the Greeks, apart from their aesthetic value the books did little to advance zoological knowledge.

Advances in anatomy. Like that of botany, the beginning of the scientific study of anatomy can be traced to a

combination of humanistic learning, Renaissance art, and the craft of printing. Although Leonardo da Vinci initiated anatomical studies of human cadavers, his work was not known to his contemporaries. Rather, the appellation father of anatomy must be accorded to the Belgian anatomist Andreas Vesalius, who studied at the rather conservative schools in Louvain and Paris, where he became a successful teacher very familiar with Galen's work. As a result of disagreements with his superiors, however, Vesalius moved at the end of 1537 to Padua, where he became noted for far-reaching teaching reforms. Most important, Vesalius abolished the practice of having someone else do the actual dissection; instead, he dissected his own cadavers and lectured to students from his findings. His text, *De humani corporis fabrica libri septem* (1543; "Seven Books on the Structure of the Human Body"), was the first modern book on the subject of anatomy and, as such, constituted a foundation of great importance for biology. Perhaps Vesalius' greatest contribution, however, was that he inspired a group of younger scientists to be critical and to accept a description only after they had verified it. Thus, as anatomists became more questioning and critical of the works of others, the stranglehold of Galen was finally broken. Of Vesalius' successors, Michael Servetus, a Spanish theologian and physician, discovered the pulmonary circulation of the blood from the right chamber of the heart to the lungs and stated that the blood did not pass through the central septum (wall) of the heart, as had previously been believed.

ADVANCES TO THE 20TH CENTURY

Seventeenth century advances in biology included the establishment of scientific societies for the dissemination of ideas and progress in the development of the microscopy, through which man discovered a hitherto invisible world that had far-reaching effects on biology. Systematizing and classifying, however, dominated biology throughout much of the 17th and 18th centuries, and it was during this time that the importance of the comparative study of living organisms, including man, was realized. During the 18th century the longheld idea that living organisms could originate from nonliving matter (spontaneous generation) began to crumble, but it was not until after the mid-19th century that it was finally disproved by Pasteur. Biological expeditions added to the growing body of knowledge of plant and animal forms and led to the 19th-century development of the theory of evolution. The 19th century was one of great progress in biology: in addition to the formulation of the theory of evolution, the cell theory was established, the foundations for modern embryology were laid, and the laws of heredity were discovered.

The circulation of the blood. William Harvey, an Englishman who studied at Padua with one of Vesalius' students, is credited with the discovery of the circulation of the blood. Prior to Harvey, the Aristotelian–Galenistic theory of circulation supposed that the blood sucked up by the heart during its expansion ebbed away during contraction; further, the theory also suggested that the blood flowed through pores between the two halves of the heart and that the heart produced a vital heat, which was tempered by the air from the lungs. In his own work, however, Harvey demonstrated that the heart expands passively and contracts actively. Also, by measuring the amount of blood flowing from the heart, he concluded that the body could not continuously produce that amount. Finally, he was able to show that blood was returned to the heart through the veins, postulating a connection (the capillaries) between the arteries and veins that was not to be discovered for another century. Harvey was also interested in embryology, to which he made a significant contribution by suggesting that there is a stage (the egg) in the development of all animals during which they are undifferentiated living masses. A biological dictum, *ex ovo omnia* ("everything comes from the egg"), is a summation of this concept.

Postulation of the existence of blood capillaries

The establishment of scientific societies. A development of great importance to science was the establishment in Europe of academies or societies; they consisted of small groups of men who met to discuss subjects of mutual interest. Although some of the groups enjoyed the financial patronage of princes and other wealthy members of society, the members' interest in science was the sole sustaining force. The academies also provided freedom of expression, which, together with the stimulus of exchanging ideas, contributed greatly to the development of scientific thought. One of the earliest of these organizations was the Italian Academy of the Lynx, founded in Rome around 1603. Galileo Galilei made a microscope for the society; another of its members, Johannes Faber, an entomologist, gave the instrument its name. Other academies in Europe included the French Academy of Science (founded in 1666), a German Academy in Leipzig, and a number of small academies in England that in 1662 became incorporated under royal charter as the Royal Society of London, an organization that was to have considerable influence on scientific developments in England.

In addition to providing a forum for the discussion of scientific matters, another important aspect of these societies was their publications. Before the advent of printing there were no convenient means for the wide dissemination of scientific knowledge and ideas; hence, scientists were not well informed about the works of others. To correct this deficiency in communications, the early academies initiated several publications, the first of which, *Journal des Savants*, was published in 1665 in France. Three months later, the Royal Society of London originated its *Philosophical Transactions*. At first this publication was devoted to reviews of work completed and in progress; later, however, the emphasis gradually changed to accounts of original investigations that maintained a high level of scientific quality. Gradually, specialized journals of science made their appearance, although not until at least another century had passed.

Early publications

The development of the microscope. The magnifying power of segments of glass spheres was known to the Assyrians before the time of Christ; during the 2nd century AD, Claudius Ptolemy, an astronomer, mathematician, and geographer at Alexandria, wrote a treatise on optics in which he discussed the phenomena of magnification and refraction as related to such lenses and to glass spheres filled with water. Despite this knowledge, however, glass lenses were not used extensively until around 1300, when some anonymous person invented spectacles for the improvement of vision. This invention aroused curiosity concerning the property of lenses to magnify, and in the 16th century several papers were written about such devices. Then, around the end of the 16th century, it was discovered that if certain lenses are mounted together in a tube, they form what physicists now call a Galilean telescope when viewed through one end, and a Galilean microscope when viewed through the other. When, in the early 1600s, Galileo used this instrument to examine the stars and planets, he was able to record such new discoveries as the rings of Saturn and the four satellites of Jupiter. Although Galileo is often credited with making the first biological observations with the microscope, he did not make any further contributions to its development.

Following subsequent technological improvements in the instrument and the development of a more liberal attitude toward scientific research, five microscopists emerged who were to have a profound affect on biology: Marcello Malpighi, Antonie van Leeuwenhoek, Jan Swammerdam, Nehemiah Grew, and Robert Hooke.

The classical microscopists

Malpighi's animal and plant studies. Marcello Malpighi, an Italian biologist and physician, conducted extensive studies in animal anatomy and histology (the microscopic study of the structure, composition, and function of tissues). He was the first to describe the inner (malpighian) layer of the skin, the papillae of the tongue, the outer part (cortex) of the cerebral area of the brain, and the red blood cells. He wrote a detailed monograph on the silkworm; a further major contribution was a description of the development of the chick, beginning with the 24-hour stage. In addition to these and other animal

studies, Malpighi made detailed investigations in plant anatomy. He systematically described the various parts of plants, such as bark, stem, roots, and seeds, and discussed such processes as germination and gall formation; he may even have suspected that plants were made up of cells, a concept that had not yet been introduced. Many of Malpighi's drawings of plant anatomy remained unintelligible to botanists until the structures were rediscovered in the 19th century. Although Malpighi was not a technical innovator, he does exemplify the functioning of the educated 17th-century mind, which, together with curiosity and patience, resulted in many advances in biology.

The discovery of "animalcules." Antonie van Leeuwenhoek, a Dutchman who spent most of his life in Delft, sold cloth for a living. As a young man, however, he became interested in grinding lenses, which he mounted in gold, silver, or copper plates. Indeed, he became so obsessed with the idea of making perfect lenses that he neglected his business and was ridiculed by his family and neighbours. Using single lenses rather than compound ones (a system of two or more), Leeuwenhoek achieved magnifications from 40 to 270 diameters, a remarkable feat for hand-ground lenses. Among his most conspicuous observations was the discovery in 1675 of the existence in stagnant water and prepared infusions of many protozoans, which he called animalcules. He observed the connections between the arteries and veins; gave particularly fine accounts of the microscopic structure of muscle, the lens of the eye, the teeth, and other structures; and recognized bacteria of different shapes, postulating that they must be on the order of 25 times as small as the red blood cell. Because this is the approximate size of bacteria, it indicates that his observations were correct. Leeuwenhoek's fame was consolidated when he confirmed the observations of a student that male seminal fluid contains spermatozoa. Furthermore, he discovered spermatozoa in other animals as well as in the female tract following copulation; the latter destroyed the idea held by others that the entire future development of an animal is centred in the egg, and that sperm merely induce a "vapour," which penetrates the womb and effects fertilization. Although this theory of preformation, as it is called, continued to survive for some time longer, Leeuwenhoek initiated its eventual demise.

The discovery of sperm

Leeuwenhoek's animalcules raised some disquieting thoughts in the minds of his contemporaries. The theory of spontaneous generation, held by the ancient world and passed down unquestioned, was now being criticized. Christiaan Huygens, a scientific friend of Leeuwenhoek, hypothesized that these little animals might be small enough to float in the air and, on reaching water, reproduce themselves. At this time, however, criticism of spontaneous generation went no further.

Swammerdam's innovative techniques. In contrast to Leeuwenhoek, who was virtually unschooled, his contemporary fellow countryman Jan Swammerdam was an educated and highly systematic worker who confined his attention to studying relatively few organisms in great detail. He employed highly innovative techniques; for example, he injected wax into the circulatory system to hold the blood vessels firm, he dissected fragile structures under water to avoid destroying them, and he used micropipettes to inject and inflate organisms under the microscope. In 1669 Swammerdam published *Algemeene Verhandeling van bloedeloose diertjens* (*The Natural History of Insects*, 1972), in which he described the structure of a large number of insects as well as spiders, snails, scorpions, fishes, and worms. He regarded all of these animals as insects, distinguishing between them according to their mode of development. Although this classification was erroneous, Swammerdam did discover a great deal of information concerning insect development.

Unfortunately, Swammerdam was subject to fits of mental instability, which, combined with financial difficulties, led to periods of depression. It was while in a state of mental disturbance that he produced his classic *Ephemeri vita* ("Life of the Ephemera") in 1675, a book about the life of the mayfly noteworthy for its extremely detailed illustrations. Sometime after his death at the age of 43, Swammerdam's works were published collectively as the *Bijbel der Natuure* (1737; "Bible of Nature"), which is considered by many authorities to be the finest collection of microscopic observations ever produced by one man.

Grew's anatomical studies of plants. Nehemiah Grew was educated at Cambridge and is regarded by some as one of the founders of plant anatomy. In 1672 he published the first of his great books, *An Idea of a Philosophical History of Plants*, followed in 1682 by *The Anatomy of Plants*. Although Grew clearly recognized cells in plants, referring to them as vesicles, or bladders, their biological significance evaded him. He is best known for his recognition of flowers as the sexual organs of plants and for his description of their parts. He also described the individual pollen grains and observed that they are transported by bees, but he did not realize the significance of this observation. Twelve years after the publication of *TheAnatomy of Plants*, a German physician utilized Grew's anatomical studies to verify experimentally sexual reproduction in plants.

The discovery of cells. Of the five microscopists, Robert Hooke was perhaps the most intellectually pre-eminent. As curator of instruments at the Royal Society of London, he was in touch with all new scientific developments and exhibited interest in such disparate subjects as flying and the construction of clocks. In 1665 Hooke published his *Micrographia*, which was primarily a review of a series of observations that he had made while following the development and improvement of the microscope. Hooke described in detail the structure of feathers, the stinger of a bee, the radula, or "tongue," of mollusks, and the foot of the fly. It is Hooke who coined the word cell; in a drawing of the microscopic structure of cork, he showed walls surrounding empty spaces and refers to these structures as cells. He described similar structures in the tissue of other trees and plants and discerned that in some tissues the cells were filled with a liquid while in others they were empty. He therefore supposed that the function of the cells was to transport substances through the plant.

Contributions of the classical microscopists

Although the work of any of the classical microscopists seems to lack a definite objective, it should be remembered that these men embodied the concept that observation and experiment were of prime importance, that mere hypothetical, philosophical speculations were not sufficient. It is remarkable that so few men, working as individuals totally isolated from each other, should have recorded so many observations of such fundamental importance. The great significance of their work was that it revealed, for the first time, a world in which living organisms display an almost incredible complexity.

Unfortunately, work with the compound microscope languished for nearly 200 years, mainly because the early lenses tended to break up white light into its constituent parts. This technical problem was not solved until the invention of achromatic lenses, which were introduced about 1830. In 1878 a modern achromatic compound microscope was produced from the design of the German physicist Ernst Abbe. Abbe subsequently designed a substage illumination system, which, together with the introduction of a new substage condenser, paved the way for the biological discoveries of that era.

The development of taxonomic principles. In 1687 in England Isaac Newton, mathematician, physicist, and astronomer, published his great work *Principia*, in which he described the universe as fixed, with the Earth and other heavenly bodies moving harmoniously in accordance with mathematical laws. This approach of systematizing and classifying was to dominate biology in the 17th and 18th centuries. One reason was that the 16th-century "fathers of botany" had been content merely to describe and draw plants, assembling an enormous and diverse number that continued to increase as explorations of foreign countries made it evident that every country had its own native plants and animals.

Aristotle began the process of classification when he

used mode of reproduction and habitat to distinguish groups of animals. Indeed, the words genus and species are translations of the Greek *genos* and *eidos* used by Aristotle. As mentioned earlier, it was the Swiss botanist Bauhin who introduced a binomial system of classification, using a generic name and a specific name. Most classification schemes proposed before the 17th century were confused and unsatisfactory, however.

The use of structure for classifying organisms. Two systematists of the 17th and 18th centuries were John Ray and Carolus Linnaeus, also known as Carl von Linné. Ray, an English naturalist who studied at Cambridge, was particularly interested in the work of the ancient compilers of herbals, especially those who had attempted to formulate some means of classification. Recognizing the need for a classification system that would apply to both plants and animals, Ray employed in his classification schemes extremely precise descriptions for genera and species. By basing his system on structures, such as the arrangement of toes and teeth in animals, rather than colour or habitat, Ray introduced a new and very important concept to taxonomic biology.

The need for a common classification system

Reorganization of groups of organisms. Prior to Linnaeus, a Swedish botanist and taxonomist, most taxonomists started their classification systems by dividing all the known organisms into large groups and then subdividing these into progressively smaller groups. Unlike his predecessors, Linnaeus began with the species, organizing them into larger groups or genera, then arranging analogous genera to form families and related families to form orders and classes. Probably utilizing the earlier work of Grew and others, Linnaeus chose the structure of the reproductive organs of the flower as a basis for grouping the higher plants. Thus he distinguished between plants with real flowers and seeds (phanerogams) and those lacking real flowers and seeds (cryptogams), subdividing the former into hermaphroditic (bisexual) and unisexual forms. For animals, following Ray's work, Linnaeus relied upon teeth and toes as the basic characteristics of mammals; he used the shape of the beak as the basis for bird classification. Having demonstrated that a binomial classification system based on concise and accurate descriptions could be used for the grouping of organisms, Linnaeus established taxonomic biology as a discipline.

Later developments in classification were initiated by three French biologists, the Comte de Buffon, Jean-Baptiste Lamarck, and Georges Cuvier, all of whom made lasting contributions to biological science, particularly in comparative studies. Subsequent systematists have been chiefly interested in the relationships between animals and have endeavoured to explain not only their similarities but also their differences in broad terms that encompass, in addition to structure, composition, function, genetics, evolution, and ecology.

The development of comparative biological studies. Once the opprobrium attached to the dissection of human bodies had been dispelled in the 16th century, anatomists directed their efforts toward a better understanding of man's structure. In doing so they generally ignored other animals, at least until the latter part of the 17th century, when biologists began to realize that important insights could be gained by comparative studies of all animals, including man. One of the first of such anatomists was Edward Tyson, an English physician who studied the anatomy of an immature chimpanzee in detail and compared it with that of man. In making further comparisons between the chimpanzee and other primates, Tyson clearly recognized points of similarity between these animals and man. Not only was this a major contribution to physical anthropology but also an indication—nearly two centuries before Darwin—of the existence of relationships between man and other primates.

Among those who gave comparative studies their greatest impetus was Georges Cuvier, a French naturalist who utilized large collections of biological specimens sent to him from all over the world to work out a systematic organization of the animal kingdom. In addition to establishing a connection between systematic and comparative anatomy, he believed that there was a "correlation of parts" according to which a given type of structure (*e.g.*, feathers) is related to a certain anatomical formation (*e.g.*, a wing), which in turn is related to other specific formations (*e.g.*, the collarbone), and so on. In other words, he felt that a great deal of anatomical information could be deduced about an organism even if the whole specimen were not available. This was to be of great practical importance in the study of fossils, in which Cuvier played a leading role. Indeed, the 1812 publication of Cuvier's *Recherches sur les ossemens fossiles de quadrupèdes* (translated as *Research on Fossil Bones* in 1835) laid the foundation for the science of paleontology. But in order to reconcile his scientific findings with his personal religious beliefs, Cuvier postulated a series of catastrophic events that could account for both the presence of fossils and the immutability of existing species.

The founding of paleontology

The origin of life. *Spontaneous generation.* If a species can develop only from a pre-existing species, then how did life originate? Among the many philosophical and religious ideas advanced to answer this question, one of the most popular was the theory of spontaneous generation, according to which, as already mentioned, living organisms could originate from nonliving matter. With the increasing tempo of discovery during the 17th and 18th centuries, however, investigators began to examine more critically the Greek belief that flies and other small animals arose from the mud at the bottom of streams and ponds by spontaneous generation. Then, when Harvey announced his biological dictum *ex ovo omnia* ("everything comes from the egg"), it appeared that he had solved the problem, at least insofar as it pertained to flowering plants and the higher animals, all of which develop from an egg. But Leeuwenhoek's subsequent disquieting discovery of animalcules demonstrated the existence of a densely populated but previously invisible world of organisms that had to be explained.

A 17th-century Italian physician and poet, Francesco Redi, was one of the first to question the spontaneous origin of living things. Having observed the development of maggots and flies on decaying meat, Redi in 1668 devised a number of experiments, all pointing to the same conclusion: if flies are excluded from rotten meat, maggots do not develop. On meat exposed to air, however, eggs laid by flies develop into maggots. But renewed support for spontaneous generation came from the publication in 1745 of a book, *An Account of Some New Microscopical Discoveries*, by John Turberville Needham, an English Catholic priest; he found that large numbers of organisms subsequently developed in prepared infusions of many different substances that had been exposed to intense heat in sealed tubes for 30 minutes. Assuming that such heat treatment must have killed any previous organisms, Needham explained the presence of the new population on the grounds of spontaneous generation. The experiments appeared irrefutable until Lazzaro Spallanzani, an Italian biologist, repeated them and obtained conflicting results. He published his findings around 1775, claiming that Needham had not heated his tubes long enough nor had he sealed them in a satisfactory manner. Although Spallanzani's results should have been convincing, Needham had the support of the influential French naturalist Buffon; hence the matter of spontaneous generation remained unresolved.

Conflicting experimental results

The death of spontaneous generation. After a number of further investigations had failed to solve the problem, the French Academy of Sciences, in January 1860, offered a prize for contributions that would "attempt, by means of well-devised experiments, to throw new light on the question of spontaneous generation." In response to this challenge, Louis Pasteur, who at that time was a chemist, subjected flasks containing a sugared yeast solution to a variety of conditions. Pasteur was able to demonstrate conclusively that any micro-organisms that developed in suitable media came from micro-organisms in the air, not from the air itself, as Needham had suggested. Support for Pasteur's findings came in 1876 from

an English physicist, John Tyndall, who devised an apparatus to demonstrate that air had the ability to carry particulate matter. Because such matter in air reflects light when the air is illuminated under special conditions, Tyndall's apparatus could be used to indicate when air was pure. Tyndall found that no organisms were produced when pure air was introduced into media capable of supporting the growth of micro-organisms. It was these results, together with Pasteur's findings, that put an end to the doctrine of spontaneous generation.

When Pasteur later showed that parent micro-organisms generate only their own kind, he thereby established the study of microbiology. Moreover, he not only succeeded in convincing the scientific world that microbes are living creatures, which come from pre-existing forms, but also showed them to be an immense and varied component of the organic world, a concept that was to have important implications for the science of ecology. Further, by isolating various species of bacteria and yeasts in different chemical media, Pasteur was able to demonstrate that they brought about chemical change in a characteristic and predictable way, thus making a unique contribution to the study of fermentation and to biochemistry.

The origin of primordial life. In recent years a Soviet biochemist, A.I. Oparin, and other scientists have suggested that life may have come from nonliving matter under conditions that existed on the primitive Earth, when the atmosphere consisted of the gases methane, ammonia, water vapour, and hydrogen. According to this concept, energy supplied by electrical storms and ultraviolet light may have broken down the atmospheric gases into their constituent elements, and organic molecules may have been formed when the elements recombined.

The notion of life from atmospheric gases

Some of these ideas have been verified by recent advances in geochemistry and molecular genetics; experimental efforts have succeeded in producing amino acids and proteinoids (primitive protein compounds) from gases that may have been present on the Earth at its inception, and amino acids have been detected in rocks that are over 3,000,000,000 years old. With improved techniques it may be possible to produce precursors of or actual self-replicating living matter from nonliving substances. But whether it is possible to create the actual living heterotrophic forms from which autotrophs supposedly developed remains to be seen.

Although it may never be possible to determine experimentally how life originated or whether it originated only once or more than once, it would now seem—on the basis of the ubiquitous genetic code found in all living organisms on Earth—that life appeared only once and that all the diverse forms of plants and animals evolved from this primitive creation.

Biological expeditions. Although a number of 16th- and 17th-century travellers provided much valuable information about the plants and animals in the Orient, America, and Africa, most of this information was collected by curious individuals rather than trained observers. A development that occurred during the 18th and 19th centuries was the organization of scientific expeditions, usually under the auspices of a particular government. The most notable of these efforts were the voyages of the "Endeavour," the "Investigator," the "Beagle," and the "Challenger," all sponsored by the English government.

Captain James Cook sailed the "Endeavour" to the South Sea islands, New Zealand, New Guinea, and Australia in 1768; the voyage provided Joseph Banks, a young naturalist, with the opportunity to make a very extensive collection of plants and notes, which helped establish him as a leading biologist. Another expedition to the same area in the "Investigator" in 1801 included a botanist, Robert Brown, whose work on the plants of Australia and New Zealand became a classic; especially important were his descriptions of how certain plants adapt to different environmental conditions. Brown is also credited with discovering the cell nucleus and analyzing sexual processes in higher plants

Without doubt, one of the most famous biological expeditions of all time was that of the "Beagle" in 1831, the members of which included the naturalist Charles Darwin. Although Darwin's primary interest at the time was geology, his visit to the Galápagos Islands aroused his interest in biology and caused him to speculate about their curious insular animal life and the significance of isolation in space and time for the formation of species. During the "Beagle" voyage, Darwin collected specimens of and accumulated copious notes on the plants and animals of South America and Australia, for which he received great acclaim on his return to England.

Darwin's voyage on the HMS "Beagle"

The voyage of the "Challenger" from 1872 to 1876 was organized by the British Admiralty to study oceanography, meteorology, and natural history. Under the leadership of Charles Wyville Thomson, the chief naturalist, vast collections of plants and animals were made, the importance of plankton (minute free-floating aquatic plants and animals) as a source of food for larger marine organisms was recognized, and many new planktonic species were discovered. A particularly significant aspect of the "Challenger" voyage was the interest it stimulated in the new science of marine biology.

In spite of these expeditions, the contributions made by individuals were still very important. Such an individual was the English naturalist Alfred Russel Wallace, who undertook explorations of the Malay Peninsula from 1854 to 1862. In 1876 he published his book *The Geographical Distribution of Animals*, in which he divided the landmasses into six zoogeographical regions and described their characteristic fauna. Wallace also contributed to the theory of evolution, publishing in 1870 a book expressing his views, *Contributions to the Theory of Natural Selection.*

The cell theory. Although the microscopists of the 17th century had made detailed descriptions of plant and animal structure and although Hooke had coined the term cell for the compartments he had observed in cork tissue, their observations lacked an underlying theoretical unity. It was not until 1838 that Matthias J. Schleiden, a German botanist interested in plant anatomy, stated, "the lower plants all consist of one cell, while the higher ones are composed of (many) individual cells." When Schleiden's friend, the German physiologist Theodor Schwann, extended the cellular theory to include animals, he thereby brought about a rapprochement between botany and zoology. The formation of the cell theory—all plants and animals are made up of cells—marked a great conceptual advance in biology, and it resulted in renewed attention to the living processes that go on in cells (see CELL THEORY AND CLASSIFICATION).

In 1846, after several investigators had described the streaming movement of the cytoplasm in plant cells, Hugo von Mohl, a German botanist, coined the word protoplasm to designate the living substance of the cell. The concept of protoplasm as the physical basis of life led to the development of cell physiology.

A further extension of the cell theory was the development of cellular pathology by Rudolf Virchow, who established the relationship between abnormal events in the body and unusual cellular activities. This gave a new direction to the study of pathology and resulted in advances in medicine.

The detailed description of cell division was contributed by Eduard Strasburger, a German botanist, who observed the mitotic process in plant cells and further demonstrated that nuclei arise only from pre-existing nuclei. The parallel work in mammals was done by the German anatomist Walther Flemming, who published his most important findings in *Zellsubstanz, Kern und Zelltheilung* ("Cell Substance, Nucleus and Cell Division") in 1882.

The theory of evolution. As knowledge of plant and animal forms accumulated during the 16th, 17th, and 18th centuries, a few biologists began to speculate about the ancestry of these organisms, although the prevailing view was that promulgated by Linnaeus—namely, the immutability of the species. Among the early speculations voiced during the 18th century, Erasmus Darwin,

an English physician and the grandfather of Charles Darwin, concluded that species descend from common ancestors and that there is a struggle for existence among animals. A French naturalist, Jean Baptiste Lamarck, who was probably the most important of the 18th-century evolutionists, recognized the role of isolation in species formation; he also saw the unity in nature and conceived the idea of the evolutionary tree.

A complete theory of evolution was not announced, however, until the publication in 1859 of Charles Darwin's *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life.* In his book Darwin stated that all living creatures multiply so rapidly that, if left unchecked, they would soon overpopulate the world. The struggle for existence

According to Darwin, the checks on population size are maintained by competition for the means of life. Hence, if any member of a species differs in some way that makes it better fitted to survive, then it will have an advantage that its offspring would be likely to perpetuate. Darwin's work reflects the influence of Thomas Robert Malthus, the English economist who in 1838 published an essay on population in which he warned that if man multiplies more rapidly than his food supply, competition for existence would result. Darwin was also influenced by an English geologist, Charles Lyell, who realized from his studies of geological formations that the relative ages of deposits could be estimated by means of the proportion of living and extinct mollusks. But it was not until after his travels in the "Beagle" in 1831, during which he observed a great richness and diversity of island fauna, that Darwin began to develop his theory of evolution. Alfred Russel Wallace reached conclusions similar to those of Darwin following his studies of plants and animals in the Malay Peninsula. A short paper dealing with this subject sent by Wallace to Darwin finally resulted in the publication of Darwin's own theories.

Conceptually, the theory was of the utmost significance, accounting as it did for the formation of new species. Following the subsequent discovery of the chromosomal basis of inheritance and the laws of heredity, it could be seen that natural selection does not involve the sharp alternatives of life or death but results from the differential survival of variants. Today, the universal principle of natural selection, which is the central concept of Darwin's theory, is firmly established.

The reproduction and development of organisms. *Preformation versus epigenesis.* A question posed by Aristotle was whether the embryo is preformed and therefore only enlarges during development or whether it differentiates from an amorphous beginning. Two conflicting schools of thought were based on this question: the preformation school maintained that the egg contains a miniature individual that develops into the adult stage in the proper environment; the epigenesis school believed that the egg is initially undifferentiated and that development occurs as a series of steps. Prominent supporters of the preformation doctrine, which was widely held until the 18th century, included Malpighi, Swammerdam, and Leeuwenhoek. In the 19th century, as criticism of preformation mounted, Karl Ernst von Baer, an Estonian embryologist, provided the final evidence against the theory. His discovery of the mammalian egg and his recognition of the formation of the germ layers out of which the embryonic organs develop laid the foundations of modern embryology.

The fertilization process. Despite the many early descriptions of spermatozoa, their essential role in fertilization was not proven until 1879, when Hermann Fol, a Swiss physician and zoologist, observed the penetration of a spermatozoon into an ovum. Prior to this discovery, during the period from 1823 to 1830, the existence of the sexual process in flowering plants had been demonstrated by Giovanni Battista Amici, an Italian astronomer and botanist, and confirmed by others. The discovery of fertilization in plants was of great importance to the development of plant hybrids, which are produced by cross-pollination between different species; it was also of great significance to the studies of genetics and evolution.

The universal occurrence and remarkable similarity of the fertilization process, regardless of the organism in which it occurs, provoked many of the leading investigators of the time to search for the underlying mechanism. It was realized that there must be some way by which the number of chromosomes is reduced before fertilization; otherwise the chromosome number would double every time a spermatozoon fused with an egg. In 1883 Edouard Van Beneden, a Belgian cytologist, showed that the eggs and spermatozoa in the worm *Ascaris* contain half the number of chromosomes found in the body cells. To account for the halving of the chromosomes in the sex cells, a process that is called meiosis, in 1887 August Weismann, a German biologist, suggested that there must be two different types of cell division, and by 1900 the details of meiosis had been elucidated. Van Beneden's work on chromosomes

The study of heredity. *Pre-Mendelian theories of heredity.* The fundamental laws of heredity were discovered in 1865 by Gregor Mendel, a Bavarian monk and biologist, but his work was ignored until its rediscovery in 1900. There were, however, a number of views on the subject that had been expressed long before Mendel. The Greek philosophers, for example, believed that the traits of individuals were acquired from contact with the environment and that such acquired characteristics could be inherited by offspring. Because Lamarck was the most famous proponent of the inheritance of acquired characteristics, the theory is called Lamarckism. This concept, which emphasized the use and disuse of organs as the significant factor in determining the characteristics of an individual, postulated that any alternations in the individual could be transmitted to the offspring through the gametes. Yet the inheritance of acquired characteristics has never been experimentally verified, despite many attempts. Furthermore, many of Lamarck's examples, such as the long neck of the giraffe, can be more satisfactorily explained by means of natural selection.

In 1885 Weismann suggested that hereditary characteristics were transmitted by what he called germ plasm—as distinguished from the somatoplasm (body cells)—which linked the generations by a continuous stream of dividing germ cells. In stating definitely seven years later that the material of heredity was in the chromosomes, Weismann anticipated the chromosomal basis of inheritance.

Francis Galton, a 19th-century English anthropologist, made a number of important contributions to genetics, one of which was a study of the hereditary nature of ability, from which he developed the concept that judicious breeding could improve the human race (eugenics). Galton's most significant work was the demonstration that each generation of ancestors makes a proportionate contribution to the total makeup of the individual. Thus, he suggested that if a tall man marries a short woman, each should contribute half of the total heritage, and the resultant offspring should be intermediate between the two parents.

Mendelian laws of heredity. The fame of Gregor Mendel, the father of genetics, rests on experiments he did with garden peas, which possess sharply contrasting characteristics—*e.g.*, tall versus short; round seed versus wrinkled seed. When Mendel fertilized short plants with pollen from tall plants, he found the offspring (first filial generation) to be uniformly tall. But if he allowed the plants of this generation to self-pollinate (fertilize themselves), their offspring (the second filial generation) exhibited the characters of the grandparents in a rather consistent ratio of three tall to one short. Furthermore, if allowed to self-pollinate, the short plants always bred true—*i.e.*, never produced anything but short plants. From these results Mendel developed the concept of dominance, based on the supposition that each plant carried two trait units, one of which dominated the other. Nothing was known at that time about chromosomes or meiosis, yet Mendel deduced from his results that the trait units, later called genes, could be a kind of physical particle that was transmitted from one generation to another through the reproductive mechanism. Discovery of dominance

Mendel's most important concept was the idea that the paired genes present in the parent separate or segregate during the formation of the gametes. Moreover, in later experiments in which he studied the inheritance of two pairs of traits, Mendel showed that one pair of genes is independent of another. Thus, the principles of segregation and of independent assortment were established.

Mendel's findings were ignored for 35 years, probably for two reasons. Because the distinguished Swiss botanist Karl Wilhelm von Nägeli failed to recognize the significance of the work after Mendel had sent him the results, he did nothing to encourage Mendel. Nägeli's great prestige and the lack of his endorsement indirectly weighed against widespread recognition of Mendel's work. Moreover, when the work was published, little was known about the cell, and the processes of mitosis and meiosis were completely unknown. Mendel's work was finally rediscovered in 1900, when three botanists independently recognized the worth of his studies from their own research and cited his publication in their work.

Elucidation of the hereditary mechanism. By 1901 it was understood how the hereditary units postulated by Mendel are distributed; it was also known that the somatic (body) cells have a double, or diploid, complement of chromosomes, while the reproductive cells have a single, or haploid, chromosome number. The experimental demonstration of the chromosomal basis for heredity had been firmly established by the German biologist Theodor Boveri soon after the turn of the century and subsequently confirmed by others. To account for the large number of observed hereditary characters, Boveri suggested that each chromosome in a pair can exchange the hereditary factors it carries with those of the other chromosome. At first the U.S. geneticist Thomas Hunt Morgan dismissed this concept, but later, when he found that it agreed with his own laboratory findings, Morgan and his collaborators assigned the hereditary units (genes) specific positions, or loci, within the chromosomes. With the genes established as the carriers of hereditary traits, William Bateson, an English biologist, coined the name genetics for the experimental study of heredity and evolution.

BIOLOGY IN THE 20TH CENTURY

The rise of molecular biology

Just as the 19th century can be considered the age of cellular biology, the 20th century has been characterized by developments in molecular biology.

Important conceptual developments. By utilizing modern methods of investigation, such as X-ray diffraction and electron microscopy, to explore levels of cellular organization beyond that visible with light microscope—*i.e.*, the ultrastructure of the cell—new concepts of cellular function have been produced. Not only has the study of the molecular organization of the cell probably had the greatest impact upon biology during the 20th century but it has also led directly to the convergence of many different scientific disciplines in order to acquire a better understanding of life processes.

Another 20th-century development has been the realization that man is as dependent upon the Earth's natural resources as are other animals. The progressive destruction of the environment can be attributed, in part, to an increase in population pressure as well as to certain technological advances. Thus, although lifesaving advances in medicine have resulted in a dramatic drop in the death rate, they have also been a factor contributing to the explosive increase in the human population. Moreover, chemical contaminants being introduced into the environment by manufacturing processes, pesticides, automobile emissions, and other means are seriously endangering all forms of life. It is for these reasons that biologists are beginning to pay much greater attention to the relationships of living things to each other as well as to their biotic and abiotic environments.

Intradisciplinary nature. There are many important categories in the biological sciences (see the Table). Botany, zoology, and microbiology deal with types of organisms and their relationships with each other. Such disciplines are subdivided into more specialized categories; for example, ichthyology is the study of fishes, algology the study of algae. All of them draw upon paleontology, taxonomy, morphology, and evolution.

Important Categories in the Biological Sciences

botany	embryology	ecology	space biology
zoology	physiology	behaviour	parasitology
microbiology	biochemistry		
paleontology	biophysics		
morphology	cell biology		
taxonomy			
evolution			

Disciplines such as embryology and physiology, which deal with the development and function of an organism, may be divided further according to the kind of organism studied; for example, invertebrate embryology and mammalian physiology. In the past few decades, many developments in physiology and embryology have resulted from studies in cell biology, biophysics, and biochemistry. This has given rise to cell physiology, cytochemistry, and ultrastructural studies, which aim at correlating structure with function. Ecology, the study of the relations of a group of organisms to its environment, includes both the physical features of the environment and other organisms that may compete for food and shelter. Ecology may be subdivided according to the environment—for example, freshwater ecology and marine ecology—and draws upon animal behaviour. One aspect of cell biology, formerly called cytology, is the investigation of the structure, composition, and function of cells; biochemistry and biophysics provide important information.

Thus, biology encompasses a number of disciplines; in fact, it has become common to divide biology into its several levels of organization, rather than to try separating the disciplines. It is useful, for example, to differentiate between organismic biology, meaning the study of the whole organism, and cell biology. Similarly the technological advances of the 20th century have allowed increased understanding of the molecules comprising living things and their aggregation and organization into such structures as chromosomes and membranes. Knowledge of this aspect of biology, called molecular biology, represents the molecular level of organization. The fourth level of organization in biology, population biology, involves the complex interaction of population of animals and plants with the environment.

Importance of technological advances

The interdisciplinary nature of modern biology. In the 17th century, with the invention of the microscope, which made possible study of the cellular level of organization, biology began to receive the benefits of scientific developments in physics. In the 18th century such developments in chemistry as a better understanding of the nature of oxygen, carbon dioxide, and water began to have important implications for biology. Today, through the disciplines of biochemistry and biophysics, both chemistry and physics have continued to make significant contributions to biology, particularly in the area of molecular biology.

Biology is also very closely related to the disciplines medicine and agriculture, out of which it developed as an independent discipline. In a sense, the roles have been reversed in the 20th century, for it is basic research being conducted in biology that is contributing to major advances currently being made in medicine and agriculture. It was biological research in the structure and function of viruses, for example, that led directly to the development of a vaccine against poliomyelitis.

Another scientific discipline, that of geology, is closely related to the biological study of paleontology. The technique of radiocarbon dating, which was developed by chemists to determine the age of biological remains, has been of great use in the fields of archaeology and anthropology as well as biology. A new discipline, space biology, has arisen through the activities of the scientists and engineers concerned with the exploration of space. The conceptual framework of biology has had to be al-

tered to accommodate newly discovered facts. In the process biology has received contributions from and made contributions to many other disciplines, in the humanities as well as in the sciences. (S.H.J./E.R.G.)

III. The philosophy of biology

An important aspect of biology, as well as of every other science, is its underlying philosophy. The accumulation of facts alone does not constitute a science; the facts have to be related and organized into some understandable framework. The philosophy of biology is constantly reassessed as new facts become known. Because it was found, for example, that certain early-19th-century hypotheses were no longer adequate to explain later data, new hypotheses had to be formulated. In the subject of evolution these new hypotheses brought biology into conflict with religious and political opinions of the day, necessitating a reassessment of the underlying philosophy of biology at that time.

METHODS OF STUDY

Like their colleagues in all other sciences, biologists observe a phenomenon and then make statements about their observations. Next, assuming that what is observed exists in reality, biologists employ two methods in an attempt to put these statements into some kind of systematic order. In the order-analytical method, existing structures or organisms are compared with each other and with fossil remains so that similarities and differences can be noted and an order established. In the causal-analytical method, experiments are performed and observations made to establish a causal connection (a cause–effect relationship) that may be of help in the search for order.

Systematizing observations

Used by biologists since the 19th century, the order-analytical method of comparison, which is fundamental in taxonomy, comparative physiology, and biochemistry, is the older of the two methods. During the 20th century, however, the experimental method of inquiry has been used more frequently. Introduced by Galileo in the physical sciences, it permits the development of a hypothesis on the basis of limited experimental results. This can then lead to further experiments to verify the original hypothesis.

There is disagreement among philosophers as to which method is more powerful. Some say that the experimental (causal-analytical) method is the only one that can lead to the formulation of new laws; others think the order-analytical method is just as valid because it can lead to such generalizations as the taxonomic system, which suggests the existence of past evolution.

SCHOOLS OF BIOPHILOSOPHICAL THOUGHT

The questions of purpose and function in the biological sciences have changed considerably since the scientific revolution of the 16th and 17th centuries. Before that time Aristotle's scientific method included the idea that in nature there is a "final cause"—a purpose—for everything that happens or exists. Most modern biologists, however, consider the idea of predetermination or purpose behind any natural relationship to be unscientific and totally lacking in proof. This has led to a question concerning the use of the concept of "function" by biologists.

The notion of predetermination

The utilization of the word "function," and the idea that it is an activity of a part of a living thing and is in some way necessary or useful for the organism, led to a dispute in the second half of the 19th century between two schools of philosophical thought, the vitalists and the mechanists. Vitalism argues that life has some characteristic that is unique to itself and that is not present in any inorganic system. Mechanism, on the other hand, holds that all life processes are basically physical and chemical processes. If biology cannot express its concepts in physical and chemical terms, according to the mechanists, it is merely the result of present scientific ignorance, which will be corrected in time with more information.

Mechanism. One of the early proponents of mechanism was René Descartes, a 17th-century French mathematician and philosopher. Although many of his biological views were not accurate, Descartes was among the first to show an interest in the activity of the organism as a whole, then a new way of thinking about the organism. He also suggested that man himself could be understood in terms of natural laws. Descartes's major contribution to biology was the concept that life can be understood through a careful application of mathematics. Because Descartes was not an experimental biologist, his theories lacked a firm factual basis. Yet by attempting to explain the workings of the body in mechanical terms, he stimulated the study of physiology. To Descartes, the bodies of animals were only machines, whereas those of humans possessed an immortal soul, a dualistic hypothesis that persists.

Mechanism in animals. Following the approach of Descartes, G.A. Borelli, an Italian student of Galileo's at Pisa, suggested that muscular contraction could be explained in physiological and chemical terms. Toward the end of the 18th century, Luigi Galvani, an Italian physiologist, demonstrated that nerves could be stimulated by electrical currents and that an electrical current was associated with the transmission of nerve impulses. In the latter half of the 19th century a German physiologist, Hermann Helmholtz, measured the speed of a nerve impulse and showed that chemical activity was involved in the process. With his famous generalization that "the invariability of the internal environment is the essential condition of free independent life," Claude Bernard, another 19th-century French physiologist, reflected the belief that the constancy of the internal environment in living systems is maintained by physiological regulation. In his studies of protozoans during the first half of the 20th century, the U.S. naturalist Herbert Spencer Jennings showed that they also responded to such stimuli as light and chemicals in a predictable way that followed the laws of physics and chemistry.

Mechanism in plants. Although the mechanist viewpoint was supported by extensive physiological studies on animals, relatively little had been done in plant physiology. In a book published in 1727, however, a plant physiologist, Stephen Hales, in England, attempted to explain certain processes in living plants in terms of physical laws. He related the amount of water absorbed by the roots of a plant to the amount of water given off by the leaves (transpiration); he also realized plant nutrition involved some necessary constituents that the plant probably obtained from the air. In France in the late 1700s Antoine-Laurent Lavoisier, a chemist, discovered that air consists of separate gases and demonstrated that plants immersed in water give off oxygen, which is utilized by animals. Lavoisier was also able to show the significance of water, oxygen, and carbon dioxide in respiration. Late in the 18th century a Dutch scientist, Jan Ingenhousz, discovered that plants take in carbon dioxide from the air and use sunlight as the energy source for nutrition, and in 1837 it was shown that only cells containing the green substance chlorophyll can combine carbon dioxide with other substances to form nutrients. When, in the late 1800s, Julius von Sachs and Nathanael Pringsheim, both German biologists, described the plant plastid as the carrier of chlorophyll, the major steps in plant nutrition became established according to the mechanistic concept.

Mechanistic concept of plant nutrition

Vitalism. Although the dispute still exists between the vitalists and mechanists, most modern biologists are mechanists who, by employing chemistry and physics in such areas as genetics, biochemistry, and molecular biology, have achieved phenomenal success in discovering and analyzing the substructure of fundamental biological processes. The vitalists, however, still claim that, even at a molecular level, the problem of function remains. They say that two things are necessary to describe completely a biological process: the analysis of a structure down to its physical and chemical (molecular) components and the function of this structure in the whole of which it is a part. For example, in the analysis of the hemoglobin molecule, the molecular structure of which is now well-known, the question of the function

of a part of it, the heme, in relation to the entire molecule becomes meaningful only when considered in relation to the function of the hemoglobin in carrying oxygen from the lungs to the organs of a vertebrate. Thus, a complete biological analysis of the molecule cannot be made without asking the right questions concerning its place in the total system.

REDEFINING THE ROLE OF THE BIOLOGIST

Changing social and scientific values. Regardless of the philosophical school with which a biologist may be associated, his role in society as well as his moral and ethical responsibility in the discovery and development of new ideas had led to a reassessment of his social and scientific value systems. A scientist can no longer ignore the consequences of his discoveries; he is as concerned with the possible misuses of his findings as he is with the basic research in which he is involved. This emerging social and political role of biologists and all other scientists requires a weighing of values that cannot be done with the accuracy or the objectivity of a laboratory balance. As a member of society, it is necessary for a biologist now to redefine his social obligations and his functions, particularly in the realm of making judgments about such ethical problems as man's control of his environment or his manipulation of genes to direct further evolutionary development.

Coping with problems of the future. As a result of recent discoveries concerning hereditary mechanisms, genetic engineering, by which human traits are made to order, may soon be a reality. As desirable as it might be, such an accomplishment would entail many value judgments. Who would decide, for example, which traits should be selected for change? In cases of genetic deficiencies and disease, the desirability of the change is obvious, but the possibilities for social misuse are so numerous that they may far outweigh the benefits.

Genetic engineering

Probably the greatest biological problem of the future, as it is of the present, will be to find ways to curb environmental pollution without interfering with man's constant effort to improve the quality of his life. Many scientists believe that underlying the spectre of pollution is the problem of surplus human population. A rise in population necessitates an increase in the operations of modern industry, the waste products of which increase the pollution of air, water, and soil. With predictions that, at the present rate of reproduction, the Earth's population will be approximately 7,000,000,000 by the year 2000, the question of how many people the resources of the Earth can support is one of critical importance.

Although the solutions to these and many other problems are yet to be found, they do indicate the need for biologists to work with social scientists and other members of society in order to determine the requirements necessary for maintaining a healthy and productive planet. For although many of man's present and future problems may seem to be essentially social, political, or economic in nature, they have biological ramifications that could affect the very existence of life itself. (E.R.G.)

BIBLIOGRAPHY. The following publications deal with fundamental biological concepts as well as with the structure and functions of living things: JEFFREY J.W. BAKER, *Cell* (1966), a paperback that presents the structure of the cell as seen by the electron microscope; and with GARLAND E. ALLEN, *Matter, Energy, and Life*, 2nd ed. (1970), a fundamental and easily understood basic chemistry of living systems, and *The Study of Biology*, 2nd ed. (1971), a general college-level survey of biological principles; GEORGE W. and MURIEL BEADLE, *The Language of Life* (1966), the chemistry of the gene explained in an easily understood manner; PETER R. BELL and C.L.F. WOODCOCK, *The Diversity of Green Plants* (1968), a compact survey of the multiformity of green plants; ERNEST BOREK, *The Atoms Within Us* (1961), a nontechnical paperback explaining the chemical processes involved in living matter; A.J. CARLSON, V. JOHNSON, and H.M. CAVERT, *The Machinery of the Body*, 5th ed. rev. (1961), metabolism, digestion, and other human physiological processes clearly explained in this excellent text; RACHEL CARSON, *The Sea Around Us*, rev. ed. (1966), a beautifully written book that discusses food chains and ecological problems of the sea in an almost poetic fashion; T. DOBZHANSKY, *Evolution, Genetics, and Man* (1955), an excellent textbook on the relationship of the evolutionary theory to man; L.C. DUNN, *Heredity and Evolution in Human Populations*, rev. ed. (1967), a nontechnical and interesting explanation of genetics and the evolution of man; JAMES D. EBERT and IAN M. SUSSEX, *Interacting Systems in Development*, 2nd ed. (1970), one of an excellent series of books, each developed around a different biological concept; PAUL R. and ANNE H. EHRLICH, *Population, Resources, Environment: Issues in Human Ecology* (1970), a general textbook of ecological problems written for the layman; LOREN EISELEY, *The Immense Journey* (1957), a well-written and most interesting story of man's development as an organism; DONALD KENNEDY, *The Living Cell* (1965), a collection of articles from *Scientific American* with good photographs of cell structure; H.R. MAHLER and E.H. CORDES, *Biological Chemistry*, 2nd ed. (1971), a standard college-level text of biochemistry; A.I. OPARIN, *The Chemical Origin of Life* (1964; orig. pub. in Russian, 1936), an updated version of Oparin's theory of the origin of life; E.P. ODUM, *Fundamentals of Ecology*, 3rd ed. (1971), a standard college-level text on ecology with an excellent discussion of the cycles in nature; JAMES A. PETERS (ed.), *Classic Papers in Genetics* (1959), original papers by outstanding geneticists; A.E. ROMER, *Man and the Vertebrates* (1941; rev. as *The Vertebrate Story*, 1959), a comparative study of the anatomy of man and other vertebrates with emphasis on evolutionary significance; M.W. STRICKBERGER, *Genetics* (1968), a standard college text with a clearly written section on population genetics; N. TINBERGEN, *Social Behavior in Animals*, 2nd ed. (1965), an introductory, well-written book about animal behaviour; FRITS WENT *et al.*, *The Plants* (1963), presents the structure, evolution, and function of plants in a nontechnical and well-illustrated fashion.

Publications concerned with the history and philosophy of biology include: ISAAC ASIMOV, *A Short History of Biology* (1964), a book written for the layman, emphasizing accomplishments of the 19th and 20th centuries; SIR GAVIN DE BEER, *Charles Darwin* (1963), a biography of Darwin by an eminent biologist; M. BERGER, *Famous Men of Modern Biology* (1968), brief biographies of 19th- and 20th-century biologists; F.S. BODENHEIMER, *The History of Biology: An Introduction* (1958), contains a short history, but also a source reader from 130 authors, beginning with the Egyptians; MORDECAI L. GABRIEL and SEYMOUR FOGEL (eds.), *Great Experiments in Biology* (1955), a presentation of scientific writings in the original, from Robert Hooke to the 20th century; ELDON J. GARDNER, *History of Biology*, 2nd ed. (1965), a college-level text, and *History of Life Science* (1960), an outline of biological history written for the biology student; PHILIP GOLDSTEIN, *Triumphs of Biology* (1965), a survey of selected aspects of biology, written for college level; URLESS LANHAM, *Origins of Modern Biology* (1968), a survey of biology from Greece to the present; RUTH MOORE, *The Coil of Life* (1961), a clearly written description of the development of molecular biology for the general reader; CHARLES SINGER, *A History of Biology*, rev. ed. (1950), a highly readable classic that surveys the historical development of biological problems; M.J. SIRKS and CONWAY ZIRKLE, *The Evolution of Biology* (1964), a college-level survey of the history of biology from prehistory to the present; HENRY OSBORN TAYLOR, *Greek Biology and Medicine* (1922, reprinted 1963), a sketch for the layman of the scientific method of Greece and Rome.

Philosophical viewpoints in biology are dealt with in the following: HANS DRIESCH, *The Science and Philosophy of the Organism*, 2nd ed., 2 vol. (1929), a vitalist's view of biological philosophy; PHILIP HANDLER (ed.), *Biology and the Future of Man* (1970), a comprehensive survey of the present status of and problems in the biological sciences; JACQUES LOEB, *The Mechanistic Conception of Life* (1964), a mechanist explains his biological philosophy; JACQUES MONOD, *Le Hasard et la nécessité: essai sur la philosophie naturelle de la biologie moderne* (1970; Eng. trans., *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*, 1971), a concise and well-written modern approach to the understanding of biology and its significance in today's society.

(E.R.G.)

Bioluminescence

Bioluminescence is the emission of light by an organism or by a test-tube biochemical system derived from an organism. It could be the ghostly glow of bacteria on decaying meat or fish, the shimmering phosphorescence of protozoans in tropical seas, or the flickering signals of fireflies. The phenomenon occurs sporadically in a wide range of protists and animals, from bacteria and fungi to insects, marine invertebrates, and fish; but it is not known

to exist naturally in true plants or in amphibians, reptiles, birds, or mammals. Bioluminescence is due to a chemical reaction (chemiluminescence) in which the conversion of chemical energy to radiant energy is direct and virtually 100 percent efficient; that is, very little heat is given off in the process. For this reason the emission is called cold light, or luminescence (*q.v.*; compare with incandescence and fluorescence). Certain light producers emit radiant energy they have stored for some time as electronic energy from prior exposure to a light source; they are said to phosphoresce.

THE SIGNIFICANCE OF BIOLUMINESCENCE

Its role in behaviour. Light production appears to be associated with the protection and survival of a species. This is quite clear in certain squids, who secrete a luminous cloud to confuse an enemy and make an escape, and in many deep-sea fishes who dangle luminous lures to attract food or show light organs that disguise their form from enemies, frighten predators, or simply light the way in the perpetual darkness of the ocean deeps.

Survival value

The survival value of bioluminescence is indisputable for many organisms who use their flashes as species-recognition and mating signals. In *Photinus pyralis*, a common North American firefly, the male flashes spontaneously while in flight, emitting on the average a 0.3-second flash every 5.5 seconds if the temperature is 25° C. The females watch from the ground and wait for a male to flash. Upon seeing a flash, a female flashes a response after an interval of about two seconds; it is this response that attracts the male. The female is unable to identify a male by his flashing; thus it is the male that recognizes the correct signal—*i.e.*, interval between flashes—and seeks out the female. The interval between the male's signal and the female's response, therefore, is crucial. Similar specific recognition codes are used by many species of fireflies. Other fireflies possibly rely on colour differences in the light signals between sexes: in *Pteroptyx* fireflies, from Rabaul, New Britain, the males glow bluish and the females yellow.

Lantern fishes and hatchet fishes, along with many other deep-sea organisms, possess distinct arrangements of light organs on the body, which may serve as species and sex recognition patterns. The light organs, or photophores, of many deep-sea fishes are placed on the ventral (lower) and lateral (side) surfaces of the body, and the light is emitted downward and outward. Such an arrangement is believed to allow the light of the photophores to be used to match the intensity of sunlight penetrating from above, thus concealing the fish's own shadow from a predator below. Some lantern fishes possess, in addition, a large nasal organ; others have a patch of luminous tissue in the tail region. In deep-sea angler fishes, the first dorsal spine is turned forward into an elongated rod, from the end of which dangles a luminous organ. When an unsuspecting prey approaches the luminous lure, it is engulfed in the large jaw. Many deep-sea fishes and organisms undertake diurnal vertical migration; *i.e.*, they rise to the surface with the setting of the sun and return to the twilight zone at sunrise. Such migrations are believed to take place for the purpose of feeding on planktonic organisms, which grow in abundance in the photosynthetic zone of the ocean's surface. During the migrations, the concealment mechanism is probably used for protection.

Its role in metabolism. The functional role of bioluminescence in lower organisms such as bacteria, dinoflagellates, and fungi is difficult to discern. Partly because the glow of luminous bacteria is extinguished when oxygen is removed, it has been suggested that the bioluminescent reaction was originally used to remove oxygen toxic to primitive types of bacteria that developed during a time when oxygen was absent or very rare in the Earth's atmosphere. The metabolic reaction that combines the oxygen with a reducing substance (luciferin) liberates sufficient energy to excite a molecule in the organism to emit visible radiation. Most of those luminous primitive organisms have subsequently developed systems of utilizing oxygen, but they have retained the luminescent capability as parts of related metabolic pathways or for some survival value that luminescence may confer on the organism.

Its role in research. The luminescent reaction of the firefly has been used as an assay method for the determination of adenosine triphosphate (ATP), an important metabolic substance used by all living cells in numerous reactions in which energy is either stored or expended. (Muscle contraction is made possible by energy released in the breakdown of ATP, and photosynthesis involves the storage of the energy of sunlight through the synthesis of ATP.) The glow of a specially blended extract of firefly lanterns eventually dims and disappears as ATP is broken down. The addition of fresh ATP, either as a pure chemical or as a constituent of a tissue extract, immediately restores the luminescence. The intensity of the glow is a direct measure of the amount of ATP present in the extract. This assay method has been widely used in medical and biological research to determine the amount of ATP present in extracts of cells and tissues. The study of reactions involving ATP has led to a detailed understanding of the mechanisms of energy conversion in cells. The firefly reaction is one of the few reactions in which ATP is directly involved with light emission. All other bioluminescent reactions involve compounds that are chemically distinct from ATP (see below).

THE RANGE AND VARIETY OF BIOLUMINESCENT ORGANISMS

Taxonomic pattern

Luminous species are widely scattered taxonomically, with no clear-cut pattern discernible. Many luminous shrimps are known but no luminous crab. Many luminous squids are known but only a single luminous octopus (*Callistoctopus arakawai* of Japan). Again, luminous centipedes and millipedes are not uncommon, but luminous scorpions and spiders are apparently nonexistent. Many plantlike protists (bacteria and fungi) exhibit bioluminescence, but no luminous true plant is known.

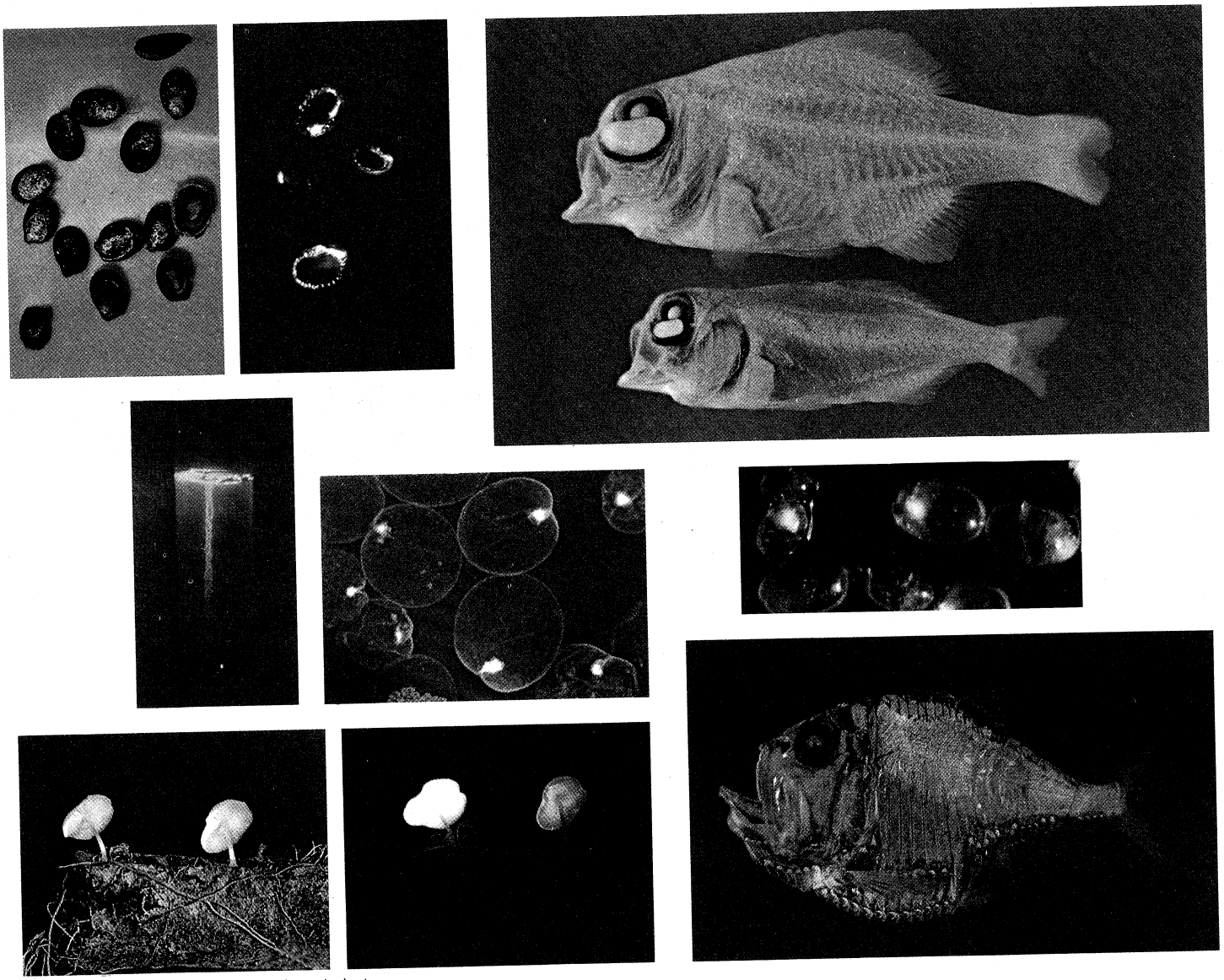
Almost half the animal phyla contain luminous forms, but the number of representatives is very small compared to the total number of known animal species. The protists are not so rich in luminous species but are greatest in sheer abundance, especially in tropical seas. In fact, the majority of luminous organisms are marine.

Marine organisms. The ocean surface in many parts of the tropics is dense with single-celled luminous planktonic organisms, primarily dinoflagellates, that glow when stimulated mechanically, as by the churning of the waves, or, when washed ashore, by the pressure of a foot. Some organisms exhibit a 24-hour rhythm of light intensity, highest at night and lowest during the day.

Among crustaceans, luminous species are especially remarkable in the copepods, shrimps, and ostracods. Luminous copepods are widely distributed throughout the waters of the world. Some are surface dwellers, while others live in the deep sea. Two famous groups of luminous copepods are *Pleuromma* and *Metridia*. Some shrimps (*Hoplophorus*) emit a luminous secretion from luminous organs, while others possess true light organs (photophores), which consist of a lens, reflector, and light-emitting photogenic cells. Of the three or four species of the ostracod genus *Cypridina* known to be luminous, the most famous is *Cypridina hilgendorfii*, found in the coastal waters and sands of Japan, running as far southward as Okinawa. This tiny, shelled organism, which ejects a blue luminous secretion into the water when disturbed, may be collected and dried for the light-emitting components, which are active indefinitely. At Palau Island in the Pacific, *Cypridina noctiluca* occurs in great numbers in surface waters. When a strong beam of light is directed into the water, the organism responds by ejecting a cloud of luminous secretion three to 15 centimetres (about one to six inches) in diameter. Luminous spots, each representing the secretion of an individual organism, may be seen when the light is turned off.

Natural displays

Natural displays of luminescence have been known from early times. Three famous displays are those at Phosphorescent Bay, Puerto Rico (*Pyrodinium bahamense*); Oyster Bay, Jamaica (*Pyrodinium bahamense*); and Sandakan Bay, North Borneo (from *Noctiluca*



Bioluminescent animals and plants.
(Top left) New Zealand freshwater limpet (*Latia neritoides*) dorsal view photographed in daylight, (centre) ventral view photographed by its own light. (Top right) Luminous fish *Photoblepharon palpebratus* (above) and *Anomalops katopteron* (below) showing large light organs beneath the eye. (Centre left) Stab culture of symbiotic luminous bacteria cultivated from a luminous organ of Australian pinecone fish (*Cleidopus gloria-maris*). (Centre) Microscopic dinoflagellate protozoan *Noctiluca*. (Centre right) Luminous ostracod crustacean *Cypridina hilgendorfii* photographed by its own light. (Bottom left) *Mycena chlorophos*, a luminous fungus from New Guinea photographed in daylight (left) and by its own light (right). (Bottom right) Deep-sea hatchet fish (*Argyropelecus* species) showing rows of ventral photophores.
Yata Haneda

miliaris with green symbionts; *i.e.*, organisms living with *Noctiluca* in a mutually beneficial association). The slightest disturbance of the water causes the organisms to flash brightly. In the bays just mentioned, luminescence may be observed throughout the year. Elsewhere, however, luminescence of the sea is seasonal. In Japan, for example, the oceans become more luminous between April and May, probably because of an increase in the number of *Noctiluca*.

Other organisms responsible for large patches of light in the ocean are jellyfish and other coelenterates and comb jellies (ctenophores). A large proportion of the floating, transparent siphonophores and the feathery, bottom-dwelling sea pens are luminous. Many of the hydroids and jellyfish are also luminous, but none of the sea anemones or corals are. Many ctenophores (*e.g.*, *Mnemiopsis*) are luminous. Sea pens (*Pennatula*), sea cactus (*Cavernularia*), and sea pansy (*Renilla*) are colonies, which upon stimulation generate a wave of luminous light that travels down the organism. The luminescence in these organisms appears to be under nervous control.

Among annelids, marine worms and earthworms both contain luminous forms. *Odontosyllis*, the fire worm of Bermuda, swarms in great numbers a few days after the full moon. Female worms, about two centimetres (almost an inch) in length, rise to the surface shortly after sunset and swim in circles while ejecting a luminous secretion. Smaller male worms, one centimetre in length, swim to where the females are circling and mate. The male is also luminous, but the light is intermittent and of intracellular character. It is not certain whether luminescence has any relationship to mating since nonluminous *Odontosyllis* exhibits similar courtship behaviour. *Chaetopterus* spends its life in a tube of parchment membrane, with openings at both ends. It luminesces when disturbed, but it is doubtful whether the luminescence has any special purpose. *Polynoe* and *Polycirrus* are luminous annelids that usually live in sand or rock. But, when occasionally attached to a sponge, they lend luminosity to the sponge itself. Luminous mollusks include *Pholas* (a bivalve), *Rocellaria* (a bivalve), *Phyllirrhoe* (a floating nudibranch), *Plocamopherus* (a nudibranch), *Planaxis* (a

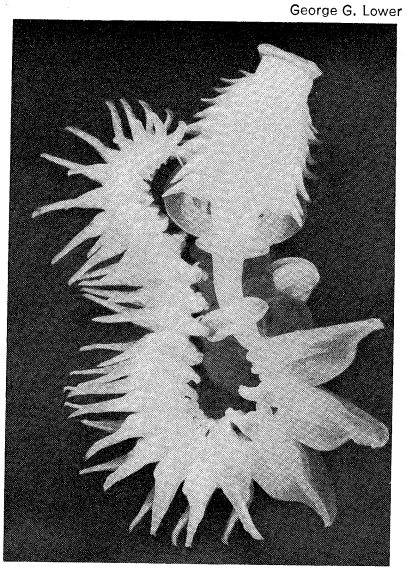
marine gastropod), *Latia* (a freshwater limpet), and squids (cephalopods).

The luminous squids and deep-sea fishes possess the most complicated light organs; they consist of photogenic cells, reflector, lens body, and, in certain cases, colour filters. Of the open-ocean forms (oegopsids) such as *Lycoteuthis*, *Histioteuthis*, and *Enoploteuthis*, as many as 75 percent are self-luminous; that is, light results not from symbiotic luminous bacteria but from an internal biochemical reaction. In the shore forms (myopsids) such as *Euprymna*, *Uroteuthis*, and *Sepiola*, of which only a few percent are luminous, a pair of luminous organs is present on the ink sac, and the light can be obscured by a flow of ink into a space between the light organ and its lens. In deep-sea squids, the light organ is often found on the eyelid or on the eyeball itself. In others—*e.g.*, *Watasenia scintillans*—light organs are present also at the end of the tentacles and over other surfaces of the body. Differences in the colour of the light emitted may result from filters or even from the colours of the reflectors. Some squids and copepods have asymmetric distribution of luminous organs. Generally speaking, the luminous organ or region is glandular, and the luminous material is ejected as an extracellular secretion. This is the typical pattern for *Heteroteuthis dispar* from the Mediterranean area and *Sepiolina nipponensis* from Suruga Bay, Japan.

Typical luminous organ

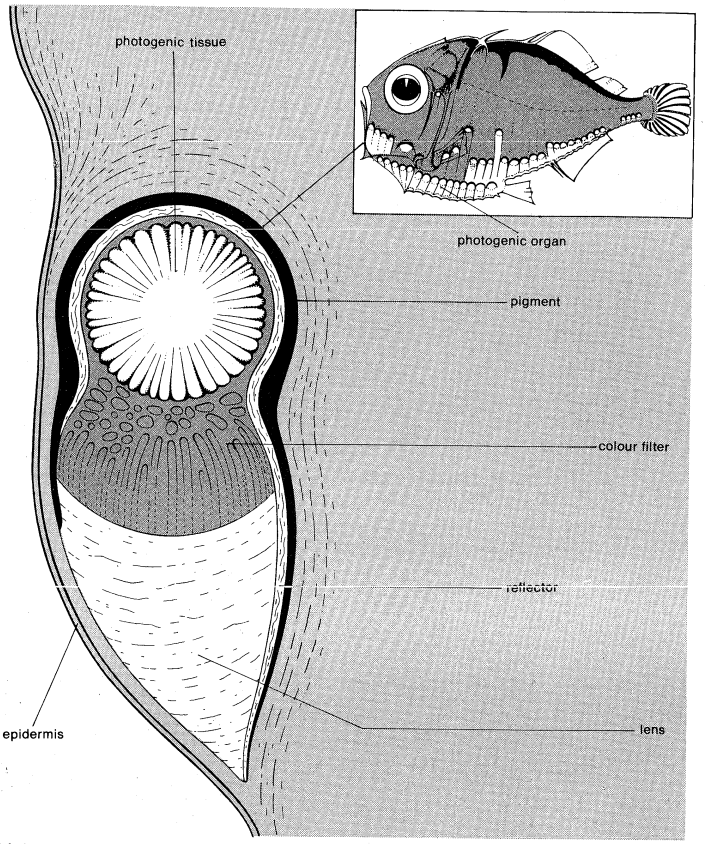
The anatomical structure of the luminous organs of many fishes is similar to that of squids. Deep-sea fishes have photophores along the body, under the eyes, and often on barbels or antennae. The typical luminous organ consists of a lens, luminous body, colour filter, and reflector (see the Figure). Luminous organs of most deep-sea fishes have no screen of melanophores over them. The light is frequently under nervous control, and emission usually takes place when the fish is alive. After death, the ability to luminesce disappears more or less rapidly. Whether the light-producing components are developed by the fish or ingested by the fish is not clear. A distinct possibility exists that a fish feeds on crustaceans such as *Cypridina* organisms and utilizes their light-emitting components for its own light production. A few genera of deep-sea fishes and several families of shallow-water fishes produce light by virtue of harbouring symbiotic luminous bacteria within light organs. This type of organ

George G. Lower



Luminescent parchment worm (*Chaetopterus*).

is endowed with a rich blood supply that nourishes and maintains the luminous bacteria. It appears that each fish species becomes infected with a specific bacterial type. The bacteria-filled organ is continuously luminous, but the light can be controlled either by melanophores scattered over the surface of the organ or by a black membrane that may be mechanically drawn over the



Light organs of the hatchet fish.

organ. Control is brought about by the contraction and expansion of melanophores, or pigment granules. Expansion of the melanophores cuts off the light, whereas contraction allows light to pass through.

The famous Indonesian fishes *Anomalops* and *Photoblepharon*, from Banda Island, possess large light organs beneath the eyes. The light in these fishes is due to symbiotic luminous bacteria (luminous bacteroids). The light is extinguished in *Anomalops* by rotation of the entire organ so that the luminous face is turned down and toward the body, thereby presenting the black-pigmented opposite face. In *Photoblepharon* the light is extinguished when a fold of black skin is drawn upward over the organ.

An indirect-emission type of luminous organ is present in some fish. The luminous organ is connected to the gut via a short duct and is often embedded in tissue. The light passes to the outside through the translucent keel and ventral muscles, as in *Leiognathus*, *Acropoma*, *Paratrachichthys*, *Parapriacanthus*, *Apogon*, *Siphamia*, *Rhabdamia*, and *Archamia*. In fish other than those mentioned above, luminescence is produced intracellularly. The light is emitted by special light-producing cells (photocytes). In some fishes (*e.g.*, *Searsia*) a luminous secretion is produced from a shoulder organ. The deep-sea angler fish *Himantolophus* has a main secretory organ on the antenna.

Among other higher animals, the chordate subphylum Tunicata contains luminous forms. The genus *Pyrosoma* includes several species that account for the brilliant luminescence among macroplanktons of the seas, giving rise to the name "fire body." *Pyrosoma* is a floating colonial form, pelagic and translucent. The colonies usually reach a length of three to ten centimetres (about one to four inches); each individual, generally about five millimetres (0.2 inch) long, possesses four or two luminous organs.

Luminous bacteria and fungi

Luminous plantlike protists are represented by only two groups, bacteria and fungi. Luminous bacteria are all marine forms, requiring salt for growth and luminescence, and are widely distributed throughout the oceans of the world. The most common are *Vibrio* and *Photo-*

bacterium species. According to a recent taxonomic study, the luminous bacteria can be divided into three major groups, which span three genera, namely *Vibrio*, *Photobacterium*, and a suggested new genus, *Lucibacterium*, to accommodate the "*Photobacterium harveyi*" organisms. While luminous bacteria come in various shapes, they do not form clusters or chains, as do many other bacteria. The light of an individual bacterium, of course, cannot be seen with the naked eye, but the light from a liquid or agar culture containing billions of bacteria is readily visible. The light is bluish and continuous. Although these bacteria are not pathogenic to man, some of them do infect sand fleas, shrimps, and insects. Many luminous bacteria live in the light organs of fish and squids, without adversely affecting their hosts.

Land and freshwater organisms. Small, whitish, luminous fungi commonly grow on dead wood of forests, particularly where the ground is moist and wet; these forms predominate in the tropics. The light of fungi ranges from blue to green and yellow, depending on the species. Among the large luminous forms are *Pleurotus lampas* of Australia and the jack-o'-lantern (*Clitocybe illudens*) of the U.S., which reach approximately 13 centimetres (about five inches) in diameter. The poisonous moonlight mushroom (*Lampteromyces japonicus*), or *tsukiyo-dake*, of Japan may reach 15 centimetres (about six inches) in diameter. In *Mycena rorida* var. *lamprospora*, from Rabaul, the spores are luminous, giving the ground beneath the mushroom a luminous glow.

Luminosity among land animals is not associated with any particular habitat, but almost all these forms are nocturnal. The centipede *Orphaneus*, widely distributed in tropical Asia, gives off luminous secretions from each segment. The entire body of *Luminodesmus sequoiae*, a millipede found in the Sierra Nevada Mountains of California, glows with a diffuse light.

Luminous insects include some true flies (order Diptera), notably *Arachnocampa luminosa*, the larva of which luminesces a greenish blue from a knob at the end of its body. The larvae dangle at the ends of filaments that hang from the ceilings of caves in New Zealand. The cave at Waitomo, 200 miles (320 kilometres) north of Wellington, is visited by many tourists each year for its spectacular display. Luminous beetles include the fireflies and the elaterid *Pyrophorus* (the click beetle, or *cucujo* in South America). The luminescent larvae of fireflies and some luminescent wingless adults are known as glowworms. The female *Diplocladon hasseltii*, called starworm, or diamond worm, gives off a continuous greenish-blue luminescence from three spots on each segment of the body, forming three longitudinal rows of light, the appearance of which inspired the common name night train. *Phrixothrix*, the railroad worm, possesses two longitudinal rows, with a red luminous spot on the head.

Synchronous displays

One of the most fascinating bioluminescent displays is the synchronous flashing of innumerable fireflies gathered in trees. The phenomenon is common in parts of tropical Asia. A single firefly probably serves as a pacemaker at the outset, but subsequent flashings are more difficult to explain. Waves of light appear to run from tree to tree when horizontal rows of trees are involved or from top to bottom or bottom to top if only a single tree is involved. Only one species congregates in a tree, and the synchronous flashing is commonly done by the males.

The limpet *Latia neritoides*, found in streams around Auckland, N.Z., is the only strictly freshwater luminous form known. A so-called luminous shrimp (*hotaru ebi*) is found in Lake Suwa, Japan, but the light is from luminous bacteria that infect the shrimp and kill it in about 24 hours.

The rarity of freshwater luminous organisms is a puzzle. No luminous fauna is known from the two deepest lakes in the world, Tanganyika in Africa and Baikal in Siberia. While Tanganyika contains no oxygen below about 200 metres (650 feet), Baikal is oxygenated at all depths yet still has no luminous forms in the dark depths, where it would seem an advantage for at least some organisms to be luminescent.

THE BIOCHEMICAL EVENTS OF LIGHT EMISSION

Enzymatic systems. The essential light-emitting components are the oxidizable organic molecule luciferin and the enzyme luciferase, which are specific for different organisms. The present custom is to use generic names according to origin; *e.g.*, firefly luciferin and luciferase, *Cypridina* luciferin and luciferase. The luciferin–luciferase reaction is actually an enzyme–substrate reaction in which luciferin, the substrate, is oxidized by molecular oxygen, the reaction being catalyzed by the enzyme luciferase, with the consequent emission of light. The light emission continues until all the luciferin is oxidized. This type of reaction is found in *Cypridina*, *Latia*, and many fish (see below).

Luciferin–luciferase reaction

The chemistry of many bioluminescent reactions has been elucidated. Most of the early chemical studies dealt with the *Cypridina* system, largely because the organism could be collected in large quantities, and the light-emitting components were fairly stable. Systematic studies led to the isolation of luciferin and luciferase in a highly pure state and finally to the determination of the chemical structure of firefly luciferin, *Cypridina* luciferin, and *Latia* luciferin. The three types of luciferin have been synthesized in the laboratory. *Cypridina* luciferin is an orange-coloured indole compound ($C_{22}H_{27}N_7O$, molecular weight 405); firefly luciferin is a thiazole compound ($C_{11}H_8N_2O_3S_2$, molecular weight 280), and *Latia* luciferin ($C_{15}H_{24}O_2$) has a molecular weight of 236. The chemical and conformational structures of the respective luciferases are unknown. Bacterial luciferin appears to be the reduced form of flavin mononucleotide, which is a highly important compound in cellular respiration.

In firefly luminescence, the substance adenosine triphosphate (ATP) initially reacts with firefly luciferase, magnesium ion, and firefly luciferin to form a complex (luciferase-luciferyl-adenylate) and pyrophosphate. This complex then reacts with molecular oxygen to emit light. Enough energy is liberated in the last step to convert the electronic configuration of the luciferase-luciferyl-adenylate complex from a low-energy ground state to a high-energy excited state. The high-energy complex then loses energy by radiating a photon of visible light and returns to the ground state. The colour of light emitted by different species of fireflies ranges from intense green to bright yellow, probably influenced by the enzyme luciferase. The common eastern firefly of the United States, *Photinus pyralis*, emits light in wavelengths ranging from 500 to 660 nanometres, for example, with an emission maximum at 562 nanometres, which expresses itself as a cool yellow green. The energy required for emission of light in this range is at least 60,000 calories.

Among fireflies of the genera *Photinus* and *Photuris*, which emit short, sharp flashes, the light organ has a regularly arranged air-supply system, consisting of tubules called tracheae, whose branches run into tracheal end organs (tracheal end cells and tracheolar cells). *Diphotus* and *Pyrophorus* lack the end cells, and they produce long, lingering glows. In *Photuris*, control of flashing appears to involve a three-element neuroeffector system, consisting of a peripheral nerve, a tracheal end organ, and a light-emitting cell, or photocyte. The nerve terminates in the tracheal end organ. It appears that in the transfer of excitation from the peripheral nerve to photocyte, via the tracheal end organ, a chemical mediator secreted by the nerve ending acts on the photocyte.

The *Renilla* system, like that of the firefly, requires an activation reaction. Luciferyl sulfate reacts with diphosphoadenosine to form luciferin. The luciferin then reacts with molecular oxygen in the presence of *Renilla* luciferase to give light.

Luminescent bacteria employ the enzymatic oxidation of reduced flavin mononucleotide ($FMNH_2$). In the complete reaction, bacterial luciferase reacts with $FMNH_2$ and oxygen to form a long-lived intermediate complex, which then reacts with a long-chain aliphatic aldehyde molecule (*e.g.*, decanal) to emit light. Aldehydes of eight to 15 carbon atoms in chain length are very effective in increasing luminescence intensity by a factor of 100 or more.

The marine worm *Balanoglossus biminiensis* possesses a still different type of light-emitting system. When disturbed, the worm secretes a greenish-blue luminous slime from its surface epithelium. Balanoglossid luciferase, luciferin, and hydrogen peroxide are all that is required for luminescence, no oxygen being needed.

When the luciferin and luciferase of certain different species are mixed, a light-emitting cross-reaction occurs. Reciprocal light-emitting cross-reactions exist among apogonid fishes, the pempherid fish, *Parapriacanthus ransonneti*, the batrachoid fishes *Porichthys notatus* and *P. porosissimus*, and the crustacean *Cypridina hilgendorfii*. In fact, the *Cypridina* system represents the only luminescent cross-reaction known among such disparate organisms as fish and crustaceans. The question as to whether the apogonids, *Parapriacanthus*, and *Porichthys* thus obtain their luciferin and luciferase by ingesting crustaceans is under study.

Other fishes that emit light via a luciferin–luciferase mechanism are the mesopelagic (200–1,000 metres) lantern and hatchet fishes. In *Porichthys notatus* the luminescent system is characterized by rows of more than 700 small, white photophores scattered over the body surface. The photophores light up when a small amount of epinephrine is injected beneath the skin or when electrical stimulation is given. The former result suggests that light emission may be partially under hormonal control. Light production in these fishes may be associated with mating. The reactions in lantern and hatchet fishes are very similar to the *Cypridina* reaction.

Nonenzymatic systems. A bioluminescent reaction that does not employ an enzyme and is not analogous to the luciferin–luciferase system involves a luminescent protein, or photoprotein, with varying molecular weights in different organisms. This system is found in two genera of jellyfishes (*Aequorea* and *Halistaura*), in the marine worm *Chaetopterus*, and in the euphausiid shrimp *Meganyctiphanes norvegica*. Light is emitted when the photoprotein of either *Aequorea* or *Halistaura* is mixed with calcium ion. Oxygen is not required in the reaction. In the *Chaetopterus* system light emission occurs when the photoprotein is mixed with molecular oxygen; a hydroperoxide; iron (ferrous ion); and two cofactors, one resembling a nucleoprotein, the other a lipid. In the *Meganyctiphanes* system light is produced when the photoprotein is mixed with an unidentified heat-stable fluorescent compound (molecular weight less than 1,000) and molecular oxygen. In the boring clam (*Pholas dactylus*) luminescence results from the reaction of a photoprotein, another protein, and molecular oxygen.

Photoprotein

An additional system is found in the marine dinoflagellate *Gonyaulax polyedra*. A crystalline-like particle called scintillon, which has been isolated from that unicellular organism, emits a flash of light when the surrounding fluid is acidified by lowering the pH from about 8 to 5.7. Soluble luciferin and luciferase, which emit light when mixed in the presence of oxygen and salt, can also be dissociated from the scintillons.

Summary. The biochemical reactions discussed above may be classified into the following five reaction types: (1) substrate oxidation (*Cypridina*, *Apogon*, *Parapriacanthus*, *Porichthys*, lantern and hatchet fishes, and *Latia*); (2) substrate activation followed by oxidation (firefly and *Renilla*); (3) reduction followed by oxidation (bacteria); (4) peroxidation (*Balanoglossus*); and (5) "precharged" systems, the photoprotein and scintillon reactions, which possess the unique feature of being readily triggered to emit light.

BIBLIOGRAPHY. For a popular account, see W. BEEBE, *Half Mile Down* (1934), a view through the window of a bathysphere; E.N. HARVEY, *Living Light* (1940), biology and chemistry of light production in animals; N.B. MARSHALL, *Aspects of Deep Sea Biology* (1954), on deep-sea, luminous organisms; for a scientific monograph, see E.N. HARVEY, *Bioluminescence* (1952), comprehensive coverage of the biology and chemistry of bioluminescence; for recent work, see P. BUCHNER, *Endosymbiosis of Animals with Plant Microorganisms*, rev. Eng. ed. (1965; orig. pub. in German, 1953), on symbiosis of luminous bacteria; H.H. SELIGER and W.D. MCELROY, *Light: Physical and Biological Action* (1965), chemistry and biology of light absorption and emission; F.H. JOHNSON, H. EYRING, and M.J. POLISSAR, *Kinetic Basis of Molecular Biology* (1954), on chemistry and biology of light production; F.H. JOHNSON (ed.), *Luminescence of Biological Systems* (1955), symposium volume on luminescence; W.D. MCELROY and B. GLASS (eds.), *A Symposium on Light and Life* (1961), symposium on photochemistry; F.H. JOHNSON and Y. HANEDA (eds.), *Bioluminescence in Progress* (1966), recent symposium on bioluminescence; for history, see E.N. HARVEY, *History of Luminescence: From the Earliest Times Until 1900* (1957), scholarly historical review; for classic account, see C.G. EHRENBERG, *Das Leuchten des Meeres* (1834), bioluminescence of the ocean surface; H. MOLISCH, *Leuchtende Pflanzen* (1904), physiology of luminous bacteria and fungi; E. MANGOLD, *Die Produktion von Licht* (1910), luminous organisms; U. DAHLGREN, "The Production of Light by Animals," *J. Frank. Inst.* vol. 180–183 (1915–17), histology and morphology of light organs; E.N. HARVEY, *The Nature of Animal Light* (1920), chemistry of light emission in animals; S. KANDA, *Hotaru* (1935), study of luminescence of Japanese fireflies (in Japanese).

(Y.Ha./F.I.T.)

Bionics

The word bionics was coined in 1958 by Maj. Jack E. Steele, of the Aerospace Division of the United States Air Force, to describe a new science of constructing artificial systems that resemble or have the characteristics of living systems. Bionics is not a specialized science but an interscience discipline.

Bionics may be compared with cybernetics, another interscience discipline. Bionics and cybernetics have been called the two sides of the same coin. Both use models of living systems, bionics in order to find new ideas for useful artificial machines and systems, cybernetics to seek the explanation of living beings' behaviour.

Relationship of bionics and cybernetics

Bionics is thus entirely distinct from bioengineering (or biotechnology), which is the science of using living beings to perform certain industrial work, such as the culture of yeasts on petroleum to furnish food proteins, the use of micro-organisms capable of concentrating metals from low-grade ores, and the digesting of wastes by bacteria in biochemical batteries to supply electrical energy.

Mimicry as the basis for bionics. Mimicry of nature is an old idea. Many inventors have modelled machines after animals or flying birds. An 18th-century French mechanical engineer, Jacques Vaucanson, for example, constructed, among other automata, a duck that swam, flapped its wings, and quacked. Direct imitation of nature is a dead end, however. Much more promising is indirect imitation. An example is the medieval whirligig. When maple-tree fruit falls from the tree, the winged membrane of the fruit causes it to spin around its axis; it can fly long distances in favourable wind. From observation of this fruit arose the idea of making an artificial screw or helix of this kind and of making it fly by giving it a rapid movement around itself. Leonardo da Vinci sketched this type of helicopter, which in his day was a popular children's toy. In the same notebook, Leonardo drew a flying machine based on the bat's wing. Four hundred years later the first man-made machine reputed to have flown was constructed by a French engineer, Clement Ader, in the shape of bat's wings. Ader did not try to move the wings as in the natural model, however: he used propellers made of quill feathers.

Nevertheless, direct inspiration from nature and close imitation of it are still useful today, as the following examples will show. The dolphin moves at great speed with a minimal muscular effort. One reason for this lies in a peculiar characteristic of the dolphin's skin, which damps out turbulence of the water. Artificial skins similar to that of the dolphin have been devised for torpedoes to reduce turbulence and achieve greater speed with the same engine power. In 1964 a new type of ship propeller modelled on a fish tail was invented by a French engineer; this design has distinct advantages over conventional propellers, in particular a high thrust at low speed. It can also function in very shallow water and is not fouled by vegetation. Vehicles have been designed by the

United States Army that, in place of wheels, have legs directly copied from the articulated legs of certain insects (arthropods) to permit movement in swamps or over very rough terrain (see illustration).

Copying from nature has distinct advantages. Most living creatures now on the earth are the end product of 2,000,000,000 years of evolution, and the construction of machines to work in an environment resembling that of living creatures can profit from this enormous experience. Although the easiest way may be thought to be direct imitation of nature, this is often difficult if not impossible, among other reasons because of the difference in scale. Though Ader's flying machine was inspired by the bat's wing, Ader realized that there was no point in trying to make the wings move, as others had already done without success: the increase in dimensions from those of a bat's wing to a 14-metre (46-foot) wingspan required a different method of propulsion. Bionics researchers since have followed the same path in solving engineering problems. The essential is not to copy in detail but to understand the principles of why things work in nature.

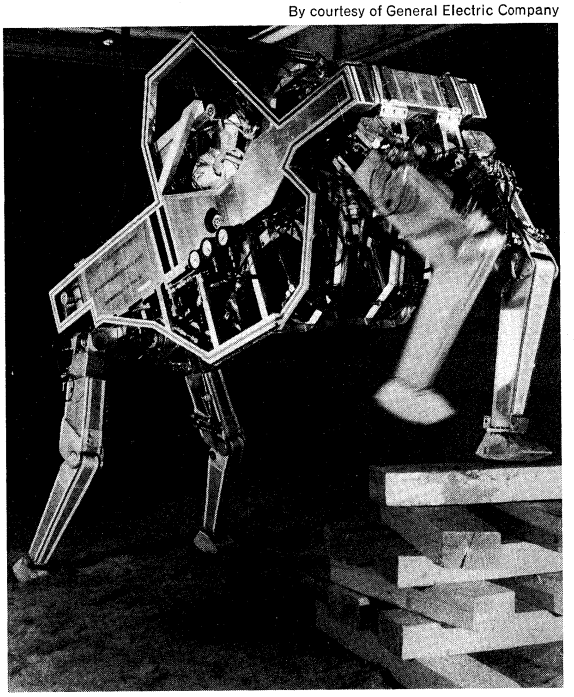
The use of natural models. The next step is the generalized search for inspiration from nature. Living beings can be studied from several points of view. Animal muscle is an efficient mechanical motor; storage of solar energy in a chemical form is performed by plants with almost 100 percent efficiency; transmission of information within the nervous system is more complex than the largest telephone exchanges; problem solving by a human brain exceeds by far the capacity of the most powerful computers. The field of action opening to bionics appears nearly unlimited in scope.

Efficiency of natural processes

Because a living organism is a system that either transforms energy or deals with information or both, two main fields—information processing and energy transformation and storage—for bionics activities exist.

Sense organs of living beings. The general pattern of the information network of living organisms is the following: environmental sensations are received by the organs of sense and then coded onto nervous-impulse signals that are transmitted by nerves to the centres of processing and memorization of the brain. Pit vipers of the subfamily Crotalinae (which includes the rattlesnakes), for example, have a highly sensitive infrared heat-sensing mechanism located in a pit between nostrils and eyes. This organ is so sensitive that it can detect a mouse at a few metres' distance. Though much more sensitive man-made infrared detectors exist, bionics can still profit from study of these sensing mechanisms. First, it would be interesting and of potential value to understand the principle of energy transformation (thermal to chemical) involved in the rattlesnake's infrared pit. Second, unlike man-made detectors, the rattlesnake's has no amplifying mechanism between the heat detector and the neural network; therefore, its sensitivity must be quite high. Another striking example is the odour-sensing organ of the *Bombyx mori*, silk moth. The male can detect the chemical secreted by the female in a quantity as small as a few molecules.

By courtesy of General Electric Company



Research prototype of a four-legged walking vehicle developed for the U.S. Army. The 11-foot, 3,000-pound machine, whose movements correspond to those of its operator, was designed to increase the capabilities of the foot soldier in modern combat.

Neural information network. Bionics must be prepared to consider any natural model as potentially useful for providing new ideas. It can also learn from the natural transmission of signals by the nerves. In a conductor such as a long-distance telephone wire, the signal is attenuated as it travels along the wire, and amplifiers must be placed at intervals to reinforce it. This is not the case for the animal nerve axon: the neural impulse issued from sense organs does not weaken in travelling along the axon. This impulse can travel in only one direction. These properties make the nerve axon a very special kind of cable capable of logic operations. In 1960 a new device called a neuristor was devised, capable of propagating a signal in one direction without attenuation. It is possible to build such neuristors with very small pieces of semiconductor material similar to that used in transistors and to arrange these devices in such a manner that they can perform numerical and logical operations. The neuristor computer is an example of bionics in two senses: the components are inspired by a natural model, and the behaviour of the components when arranged as computing circuits imitates the dynamic behaviour of natural neural information networks in contrast to conventional computers. The properties of neuristor computers are thus entirely new: each circuit can serve sequentially for different operations in a manner similar to that of the nervous system.

Pattern-recognition and learning machines. Bionics collaboration between the data-processing specialist and the specialist in nerve physiology produced the neuristor computer. This idea is far from being alone in this important field.

Another question of interest to bionics is how a living system makes use of information. Faced with a given environment, man begins by considering what would be the results of alternative courses of action. No situation is entirely new; there is always some resemblance to a situation experienced before. "Pattern recognition" is obviously an important element in human action, and as such it has implications for bionics.

One way to design an artificial machine capable of pattern-recognition properties is to use learning processes. In 1957 such a machine, called Perceptron, was built. Perceptron learns by establishing and modifying connections among a large number of possible alternative routes in a net of pathways. This learning, however, is still rudimentary and far from human.

Perceptron learning machine

The neural nets in the human brain are extremely complex, the interconnections growing during the first months of life. Science and technology do not yet know how to realize such complexity with man-made components. Before constructing a truly intelligent machine it is necessary to understand better the faculty in man.

Man's associative memory. Man is an astonishing information-processing system, using an extremely small amount of energy with an indefinitely small failure rate and a virtually unlimited memory. To produce a machine capable of translating one natural language into another natural language, a machine that will have the volume of one litre (about one quart) and the weight of one kilogram (about 2.2 pounds) might well be considered impossible; yet such a machine already exists: the human brain. The problem of automatic language translation, quite apart from the size specification, remains

intractable, although this does not mean that the problem eventually cannot be solved but rather that it is necessary to find new approaches to its solution.

Computer and human memories

The first essential difference between existing electronic computers and the human brain lies in the way their memories are organized. In either the memory of a living being or that of a machine the main problem lies in retrieving information once it has been stored. The method computers use is called "addressing." A computer memory can be compared to a large rack of pigeonholes, each having a particular number or address (location). It is possible to find a certain piece of information if the address—that is, the number of the pigeonhole—is known. The human memory works in a very different way, using association of data. The retrieval is done according to the content of the information, not according to an external address artificially added to the useful content. That difference is qualitative as well as quantitative. Man-made memory devices are now constructed using associative principles, and there is a great potential in this field with such new technologies as holography and optical storage of data.

The second main difference between the existing electronic computers and the human brain resides in the manner of dealing with the information. A computer processes rigorously precise data. Man accepts fuzzy data and carries out operations that are not strictly rigorous. Also, computers perform only very simple elementary operations, producing complex results by performing a vast number of such simple operations at very high speed. In contrast, the human brain performs at low speed but in parallel rather than in sequence, producing several simultaneous results that can be compared.

In a computer, memory and treatment are distinct; in man, they are mixed. With the technology of the 1970s it is possible to imagine devices that could both store and process large amounts of data at the same time. The combination of new man-made components with new ideas coming from bionics could result in entirely new computers in the near future.

Thus, it is not in the field of arithmetical calculation that bionics can bring in new ideas, because in that area the human brain has limited capacity. But in the realm of decision making and problem solving it can serve as an invaluable model for creating artificial intelligence.

The association of the two words artificial and intelligence is shocking. Computers and computer programming are man-made, and therefore the machine cannot do more than the man. The machine can do more in another sense, however, because of its higher speed and its low failure rate, properties that can be exploited.

Energy transformation and storage. As noted before, every living system can be studied from two main points of view: energetic and informational. Bionics can supply fresh ideas in the information field. It can perhaps be even more productive in the field of energy.

The world energy consumption doubles regularly every 20 years, with increasing environmental degradation among the consequences. Though the most convenient energy form for most purposes is electrical, such energy cannot be stored and must be constantly supplied. In the living world, energy is stored in the form of chemical composites; its use always brings about chemical reactions. The use of solar energy by plants involves complex chemical processes. The muscular motor is a chemical motor. The light produced by such living organisms as mushrooms, glowworms, and certain fishes is of chemical origin. In every case the energy transformation is remarkably efficient compared with thermal engines.

A beginning is just being made in understanding how these transformations take place in living material and the nature of the complex role played by living membranes. Living molecules are extremely complex and relatively fragile. If a biochemical reaction is too strong or violent, the heat exchanged might destroy life instead of maintaining it. Perhaps some of these limitations could be overcome in man-made artificial-energy machines and better results achieved than in natural membranes. Studies are now in progress in this field.

BIBLIOGRAPHY. LUCIEN A. GERARDIN, *La Bionique* (1968; Eng. trans., *Bionics*, 1968), presents a general view; while VINCENT J. MARTEKA, *Bionics* (1965), deals mainly with bionics as mimicry of nature. The proceedings of various bionics symposia are reprinted in the following publications: *Bionics Symposium: Living Prototypes, the Key to New Technology*, 32 reports, WADD Technical Report 60–600 (1961); E.E. BERNARD and M.R. KARE (eds.), *Biological Prototypes and Synthetic Systems: Proceedings of the Second Annual Bionic Symposium*, 48 reports (1962); bionics issue of the *IEEE Transactions on Military Electronics* (April–July 1963), an account of the third symposium of bionics, including 29 reports; *Bionics Symposium, Dayton, Ohio 1966* (1968), 61 reports published by the Aerospace Medical Division, Wright–Patterson Air Force Base; and the report of the colloquium on *Les Modèles bioniques des systèmes sonars animaux* held in Frascati in September 1966 (1967).

(L.A.G.)

Biophysics

In its broadest sense, biophysics is concerned with the solution of biological problems in terms of the concepts of physics and other physical sciences. The relatively recent emergence of biophysics as a scientific discipline may be attributed, in particular, to the spectacular success of biophysical tools in unravelling the molecular structure of deoxyribonucleic acid (DNA), the fundamental hereditary material, and in establishing the precisely detailed structure of proteins such as hemoglobin in order that the position of each atom may be known. Biophysics and the intimately related subject molecular biology now are firmly established as cornerstones of modern biology.

HISTORICAL ASPECTS

Origins. The origin of biophysics antedates the division of natural sciences into separate disciplines. Bioluminescence must be considered among the most ancient objects of biophysical exploration, because the emission of light by living organisms has long stimulated the curiosity of natural philosophers. Perhaps the first scientific investigation of animal luminescence was that of Athanasius Kircher, a 17th-century German Jesuit priest, who devoted two chapters of his book *Ars Magna Lucis et Umbrae* to bioluminescence. In the midst of his more scientific observations, Kircher found time to expose as a fallacy the notion that an extract made from fireflies could be used to light houses.

The relation between electricity and biology became a subject of speculation in the 17th century and one of intense exploration in the 18th and 19th. Sir Isaac Newton in the *Principia* (1687) wrote of "a certain most subtle spirit which pervades and lies hid in all gross bodies," and that "all sensation is excited, and the members of animal bodies move at the command of the will, namely, by the vibrations of this spirit, mutually propagated along the solid filaments of the nerves, from the outward organs of sense to the brain, and from the brain into the muscles." Man's fascination with animal electricity is illustrated in a letter written by John Walsh in 1773 to the American inventor and statesman Benjamin Franklin; Walsh wrote the details of his discovery of the electrical nature of the discharge from the torpedo or electric ray:

> I am concerned that other engagements have prevented me from giving to the Royal Society, before their recess, a complete account of my experiments on the electricity of the torpedo; a subject not only serious in itself, but opening a large field of interesting inquiry, both to the electrician in his walk of physics, and to all who consider, particularly or generally, the animal oeconomy.

The role of physical and chemical studies. Typical of the unity of science that then prevailed were the advances sometimes made either by professors of physics who were interested in biological phenomena or professors of anatomy, a subject that at that time included physiology. Thus Abbé Giovanni Beccaria, professor of physics in Turin and Italy's leading student of electricity in the mid-18th century, carried out experiments on the electrical stimulation of muscles. Albrecht von Haller, professor of anatomy and surgery at Göttingen, discussed "the nervous fluid" and conjectured as to whether "electrical matter"

Electrical stimulation of muscles

and "animal spirits" were the same. In 1786 Luigi Galvani, a physician in Bologna, made the crucial experiment that helped end this controversy. Galvani supposedly was performing experiments with a machine in the company of friends, when, by chance, one member of the party idly probed with a knife the nerves of the thigh of a skinned frog to be used for soup. As the muscles of the frog leg suddenly and unexpectedly contracted, Galvani's wife noted that a spark had been produced by the electrical machine and "fancied that there was an agreement in point of time." Although Galvani's own account of the occurrence differed somewhat in detail from the preceding, it is certain that the experiment was repeated and verified, setting the stage for a long controversy between the advocates of Galvani's view that current generated by an animal can cause contraction and those of Alessandro Volta, who claimed that the frog leg served only as a detector of minute differences in electrical potential external to it. The Galvani partisans performed an experiment in which no external sources of electricity were present, thus proving that current generated by an animal could cause the muscle contraction. But it was also possible to cause contraction by contact with metals; Volta performed such investigations, and they culminated in his invention of the electrical battery, which was so important that it overshadowed Galvani's research. As a result, the study of electrical potential in animals disappeared from scientific consideration until 1827.

Because for many years the frog leg was the most sensitive detector of differences in electrical potential, final acceptance of the view that currents can be generated by living tissues had to await the construction of galvanometers sensitive enough to measure the minute currents generated in muscles and the small potential differences across nerve membranes. Galvanometers were built by the great German 19th-century electrophysiologist Du Bois-Reymond, professor of physiology in Berlin. His investigations of muscular current and electrical potential of nerves depended upon a galvanometer of his own devising that required 3.17 miles (5.10 kilometres) of wire wound in 24,000 turns. Research in this subject, called neurophysiology, grew in stature with increased understanding of both electrical phenomena and cellular physiology; it served as one point of origin for biophysics.

Biophysics also grew out of investigations on diffusion gradients and osmotic pressure—two forces responsible for the passive flow of matter in living organisms. Osmotic pressure, the pressure that develops in a solution separated from a solvent by a membrane permeable only to solvent, was first described by Abbé J.A. Nollet, who became professor of experimental physics at the College of Navarre. The semipermeable membranes required to produce the fluid flow that characterizes osmotic phenomena initially came from biological sources; French scientist René Dutrochet wrote in 1828, "it appears from these new studies that the endosmotic and exosmotic phenomena, which I discovered, belong to a new class of physical phenomena, whose powerful intervention in the vital phenomenon is no longer doubtful." Following the first quantitative measurements by the botanist W.F.P. Pfeffer, the fundamental laws governing diffusion were enunciated by Adolph Fick, who in 1856 published what is probably the first biophysics text, *Die medizinische Physik* ("Medical Physics"). Fick developed the laws of diffusion not from experiment but by analogy with the laws governing the flow of heat; subsequent laboratory experiments proved the analogy to be quantitatively exact.

The development of physical chemistry

Physical and chemical investigation coalesced in physical chemistry, a subject that began to develop with the emergence of the *Zeitschrift für Physikalische Chemie* in 1887, a journal founded by Dutch chemist Jacobus van't Hoff and German chemist Wilhelm Ostwald. The first volume contains contributions from the most noted physical chemists of the time, including van't Hoff, Ostwald, François Raoult, and Svante Arrhenius. They were concerned with reactions in solution, a central topic in biology because the interior milieu of all living cells is aqueous, and the chemical reactions that sustain life take place in water. The scientific interests of van't Hoff in particular transcended the boundaries between disciplines. He stressed the importance of the laws of osmosis, which he had clearly delineated, to the economy of all living processes.

SCOPE OF BIOPHYSICS

Interdisciplinary nature. The biophysical approach is unified by a consideration of biological problems in the light of physical concepts, so that biophysics is, perforce, interdisciplinary. Biophysics may be thought of as the central circle in a two-dimensional array of overlapping circles, which include physics, chemistry, physiology, and general biology. Relations with chemistry are mediated through biochemistry and chemistry; those with physiology, through neurophysiology and sensory physiology. Biology, which may be viewed as a general subject pervading biophysical study, is evolving from a purely descriptive science into a discipline increasingly devoted to understanding the nature of the prime movers of biological events. The evolution of biology in these directions has received great impetus from the biophysical and biochemical discoveries of the 20th century. An understanding of the physical principles governing biological effects is the proper end of biophysics.

The content of biophysics. The content and methods of biophysics are illustrated by examining several notable contributions to science.

Protein structure. Within two days after the initial publication of the discovery of X-rays in 1895 by Wilhelm Röntgen, a surgeon in Scotland used X-rays to observe a needle as he extracted it from the palm of an unfortunate seamstress. Although this medical application resulted in the development of radiological diagnosis and treatment of disease by radiation, physical aspects of Röntgen's discovery also provided the means for elucidating the structure of proteins and other large molecules. The laws governing the diffraction of X-rays were discovered by the two Braggs, Sir William and Sir Lawrence, who were father and son. At the Cavendish Laboratory at the University of Cambridge, where Sir Lawrence was professor, the scientist J.D. Bernal was studying the use of X-ray diffraction for the determination of the structure of large biological molecules. He had already used X-rays to define the size and shape of the tobacco mosaic virus and showed it to have a regular internal structure. At the Cavendish Laboratory the group that formed around Bernal, a man of wide public and scientific interests, included the Nobel Prize winners Max Perutz and John Kendrew, who in 1937 began to use X-rays to analyze two proteins fundamental to life, myoglobin and hemoglobin, both of which function in the transport of gases in the blood. Twenty-two years passed before the structures of these proteins were established; the significance of the work is that it provided the basis for an understanding of the mechanism of the action of enzymes and other proteins, now an active and fruitful subject of investigation.

The significance of X-rays

Deoxyribonucleic acid. Interest in biophysics at the Cavendish Laboratory resulted in another important discovery, the structure of deoxyribonucleic acid (DNA), the genetic material. This achievement by a British biophysicist, Francis H.C. Crick, and by a U.S. biochemist, James Watson, was based on X-ray data obtained by Maurice Wilkins at King's College, London. When Crick first went to the Cavendish Laboratory for education in biophysics, he worked under Perutz's direction; when Watson came to the Cavendish, he and Crick began the collaboration that led to the establishment of the structure of DNA, for which Watson, Crick, and Wilkins later were awarded a Nobel Prize.

Much impetus for biophysical investigation following World War II came from the desire of physicists to move away from physics and into biology; this drive was strengthened by the publication in 1944 of Erwin Schrödinger's book *What is Life?* Schrödinger, the Austrian physicist who contributed substantially to the development of wave mechanics, was anxious to determine whether biological events could be accounted for in

terms of known laws of physics and chemistry, or whether a full explanation would require the formulation of physical laws not yet known to exist. Because biological reproduction seemed to pose intractable problems, he devoted a chapter of his book to a consideration of the gene. The discussion was based on the model put forward by Max Delbrück, a physicist who had for some years been studying the genetics of viruses that infect bacteria (bacteriophages). Delbrück's summer course on bacteriophages in 1945 at Cold Spring Harbour in New York set in motion the chain of events that led to understanding the genetic code by which the sequence of the nucleotides in DNA is translated into the sequence of amino acids in a protein. The use of bacteriophage also provided an opportunity for experiments with a primitive living organism that could be studied without anatomic complexities. This aspect of biophysics has become more biochemically oriented as it has developed and is now known as molecular biology; sometimes it is considered a distinct discipline, and other times it is subsumed under the biophysical sciences.

The nerve impulse. Important aspects of biophysics have been derived from physiology, especially in studies involving the conduction of nerve impulses. One important scientific product of World War II—the development of vastly improved electronics—largely resulted from radar devices that had been used primarily for locating aircraft. Another product, the atomic bomb, was constructed by way of nuclear reactors that could, in peace time, provide an abundant supply of radioactive isotopes, which are now of great value not only in biophysical research but also in biochemistry and medicine. These two disparate advances were important to the work of two Nobel Prize winners, Alan Hodgkin and Andrew Huxley, who showed how the flow of sodium and potassium across the membranes of nerves can be coupled to produce the action potential, a brief electrical event that initiates the action potential, which propagates the nervous signal.

A model of the nerve axon proposed by Hodgkin and Huxley grew from a 19th-century confluence of ideas. Julius Bernstein, an experimental neurophysiologist, used physical chemical theories to develop a membrane theory of nervous conduction; Hodgkin's initial experiments were designed to test specific predictions of the Bernstein hypothesis. Early in 1938 Hodgkin learned of the important results of a newly developed technique that allowed examination of the time course of nervous conduction. After World War II, Hodgkin, joined by Huxley, again took up the research. They presented their explanation of the mechanism of nervous conduction in five scientific papers between October 1951 and March 1952.

Biological membranes. The availability of radioactive isotopes provided the technology necessary for understanding how molecules are transported across biological membranes, which are the very thin boundaries of living cells; the environment maintained by membranes in cells differs from the external environment and permits cellular function. The Danish physiologist August Krogh laid the groundwork in this subject; his pupil, Hans Ussing, developed the conceptual means by which the transport of ions (charged atoms) across membranes can be identified. Ussing's definition of active transport made possible an understanding, at the cellular level, of the way in which ions and water are pumped into and out of living cells in order to regulate the ionic composition and water balance in cells, organs, and organisms. The molecular mechanism by which these processes occur, however, remains to be discovered.

Conceptual explanation of active transport

In addition to the function of transport, membranes also are utilized as templates on which such molecules as enzymes, which must function in a sequential fashion, can be kept in the requisite order. Although great progress has been made in understanding the mechanisms by which specific atoms are assembled into large biological molecules, the principles involved in the assembly of molecules into membranes, which are organized structures of a higher degree of complexity than large molecules, are not yet very well understood. There is reason to believe that the incorporation of a molecule into a membrane endows it with properties that differ from those of a molecule in solution. A primary task of biophysics is to understand the physical character of these cooperative interactions that are essential to life.

Muscle contraction. One influential early proponent of biophysics—the British biophysicist A.V. Hill—developed exquisitely sensitive temperature sensors for measuring heat generated during muscular contraction; he initiated studies relating this heat to the thermodynamic parameters responsible for it. The electron microscope in the years following World War II made possible the description of muscular contraction at a structural level, although the mechanisms involved in the flow of heat during the process are not yet known. Simultaneously, in the 1960s, but independently, various physicists postulated the sliding-filament theory of muscular contraction, according to which muscles contract by the sliding of one filament along another and not by a springlike coiling. Remarkable advances, based on the use of techniques such as X-ray diffraction and electron microscopy, have made it possible to visualize many of the molecules involved in the process. The entire process of muscular contraction, in terms of an identification of the molecules and a description of the chemical reactions in the muscle fibre, has now been almost completely explained.

Sensory communication. The above comprise a few specific examples of the scope of biophysics. One area, difficult to discuss in specific terms, is that of sensory communication. Because stimuli, particularly those of a visible or auditory nature, can easily be specified in exact physical terms, they have excited the interest of physical scientists since before 1850. Modern electronic techniques make it relatively easy to distinguish true signals from noise; in addition, computers make possible the performance of significant experiments concerning the complex relationship between stimulus and action. Quantitative analysis of sensory response is very difficult, however, because it involves a synthesis of the action of many cells. It has been pointed out that

> An adequate theory of sensory function implies an adequate theory of brain function. And an adequate theory of brain function in its turn requires that the nervous system's behavioral repertory be predictably related to the behaviour of the elements that compose it.

THE STUDY OF BIOPHYSICS

The nature of the biophysicist and his work. A.V. Hill has written that

> Biological phenomena, like many others, show aspects and relations susceptible of physical analysis and interpretation. It is by the choice of problems and by the intellectual processes with which they are formulated and attacked, more than by the particular techniques employed, that a subject can be most clearly defined. There are people to whom physical intuitions come naturally, who can state a problem in physical terms, who can recognize physical relations when they turn up, who can express results in physical terms. These intellectual qualities, more than any special facility with physical instruments and methods, are essential to the make-up of a biophysicist. Equally essential, however, are the corresponding qualities, intuitions, and experience of the biologist. A physicist who cannot develop the biological approach, who has no curiosity about vital processes and functions, who is not willing to spend time in learning the habits of living things, who regards biology simply as a branch of physics has no important future in biophysics. (From *Science*, Dec. 21, 1956.)

Most biophysical research has been carried out by physicists with an interest in biology; therefore, there must be a way by which scientists educated in physics and physical chemistry can find their way into biology and become familiar with problems that may be open to a physical interpretation. Although classically oriented biology departments now often offer positions to biophysicists, they are not substitutes for centres in which biophysical research is of central importance.

Tools of the biophysicist

The biophysicist possesses the ability to separate biological problems into segments that are amenable to exact physical interpretation and to formulate hypotheses that can be tested by experiment. The primary tool of the

biophysicist is an attitude of mind. To this might be added the ability to use complex physical theory to study natural objects—for example, that involved in the X-ray diffraction techniques used to determine the structure of large molecules such as proteins. The biophysicist usually recognizes the utility of new physical tools—*e.g.*, nuclear magnetic resonance and electron spin resonance—in the study of specific problems in biology. But he may also, through previous experience in building specialized equipment to solve physical problems, not have to rely on commercially built instruments.

Applied biophysics. The development of instruments for biological purposes is an important aspect of applied biophysics. Biomedical instrumentation is probably most widely used in hospitals. Applied biophysics is important in the field of therapeutic radiology, in which the measurement of dose is critical to treatment, and in diagnostic radiology, particularly with techniques involving isotope localization and whole body scanning to aid in tumour diagnosis. As aids in diagnosis and patient care computers are of increasing importance. Automation of the chemical analyses routinely carried out in hospitals will soon be a reality. The opportunities for the applications of biophysics seem limitless because the lengthy delay between the development of a research instrument and its application means that many scientific instruments based on physical principles already known will be shown to have important potential for medicine.

Biophysical societies and journals. Many of the ways in which scientists come together to deal with matters of common concern are found in biophysics. The first biophysics department in the United States was probably the Johnson Foundation for Medical Physics, which opened in 1929 at the University of Pennsylvania in Philadelphia. Although biophysics had been developed in England before 1952, stimulated in large measure by A.V. Hill, not until that year was a university department in biophysics begun—at University College, London.

The first society of biophysics was the Dutch Stichting voor Biofysica, in 1932. After the Biophysical Society in the United States was formed in 1957, there shortly appeared similar groups in Japan (1960), in Great Britain (1961), and in the Soviet Union and many other countries. The International Organization for Pure and Applied Biophysics was formed in 1961; after that organization became a member of the International Council of Scientific Unions in 1966, its name was changed to the International Union for Pure and Applied Biophysics. Thus, biophysics has taken its place with the older established disciplines of mathematics, physics, and chemistry, and has become an equal partner with biology, physiology, and biochemistry.

BIBLIOGRAPHY. LUIGI GALVANI, *De viribus electricitatis in motu musculari commentarius* (1791; Eng. trans., *Commentary on the Effect of Electricity on Muscular Motion*, 1953), Galvani's own account of the discovery of animal electricity; *On Animal Electricity, Being an Abstract of the Discoveries of Emil Du Bois-Reymond*, ed. by H. BENCE JONES (1852), an English abstract of the work of this great 19th-century German scientist; A.V. HILL, "Why Biophysics?" *Science*, 124: 1233–1237 (1956), probably the best concise statement of the attributes of a biophysicist and the qualities needed to make important scientific contributions in the field; NOBEL FOUNDATION, *Les Prix Nobel en 1962* (1963), contains the lectures delivered by Crick, Wilkins, Watson, Kendrew, and Perutz when they received the Nobel Prize; R. OLBY, "Francis Crick, DNA and the Central Dogma," *Daedalus*, 99:938–987 (1970), a good historical account of the discovery of DNA and the subsequent events; J.D. WATSON, *The Double Helix* (1968), a fascinating first-person account of the discovery of the molecular conformation of DNA; A.L. HODGKIN, "The Ionic Basis of Nervous Conduction," *Science*, 145:1148–1153 (1964), his Nobel prize lecture; A.F. HUXLEY, "Excitation and Conduction in Nerve: Quantitative Analysis," *Science*, 145:1154–1159 (1964), his Nobel prize lecture; H.E. HUXLEY, "The Mechanism of Muscular Contraction," *Scient. Am.*, 213:18–27 (1965), a definitive semipopular exposition of the mechanism of muscular contraction; A.K. SOLOMON, "A Short History of the Foundation of the International Union for Pure and Applied Biophysics," *Q. Rev. Biophys.*, 1:107–124 (1968).

(A.K.S.)

Biosphere

Before the coming of life, the Earth was a bleak place, a rocky globe with shallow seas and a thin band of gases—largely methane, ammonia, hydrogen sulfide, and water vapour. It was a hostile and barren planet. This strictly inorganic state of the Earth is called the geosphere; it consists of the lithosphere (the rock and soil), the hydrosphere (the water), and the atmosphere (the air). Energy from the Sun relentlessly bombarded the surface of the primitive Earth and in time—millions of years—chemical and physical actions produced the first evidence of life, formless, jellylike blobs that could collect energy from the environment and produce more of their own kind.

This generation of life in the thin outer layer of the geosphere established what is called the biosphere, the "zone of life," an energy-diverting skin that uses the matter of the Earth to make living substance. When man appeared, the biosphere became more and more influenced by his presence and was transformed into the anthroposphere by his activities. Within a relatively short span, geologically speaking, man attained a threatening mastery over the Earth and began exploiting it. The age of ecological enlightenment has brought with it a new term, the ecosphere, which implies a responsible stewardship of Earth. Beyond and superimposed on these spheres lies another dimensional sphere, the noosphere, a figurative envelope of conceptual thought, or reflective impulses produced by the human intellect, that can be imagined to weigh upon the entire globe in a mystical way. It is not scientifically measurable, of course, but its presence is strangely felt and its influence all-pervading.

From G. Hardin, *Biology: Its Principles and Implications*, 2nd ed. (Copyright © 1966); W.H. Freeman and Company

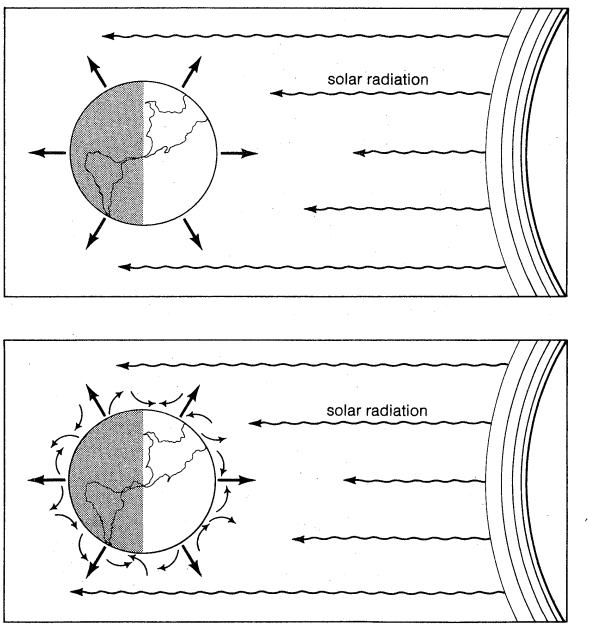


Figure 1: *Effect of the Earth on the dissipation of solar energy into space.*
(Above) Direct dispersal of energy in the geosphere stage, prior to biosphere development. (Below) Transformation and diversion of energy in the biosphere as a result of the action of living things.

GENERAL FEATURES

Preconditions of the biosphere. The Earth is an ideal medium for life. It is at precisely the proper distance from the Sun to receive neither too much nor too little sunlight. It spins on its axis at a rate fast enough to allow the daytime side to warm in sunshine and the nighttime side to cool. Its mass—and therefore its gravity—is such that it holds a wide variety of molecules, including the lighter ones that otherwise would drift off into space. Its magnetic field deflects back to space the Sun's highly energetic radiation, which otherwise would destroy life if free to bombard the Earth's surface directly. The bio-

sphere, then, is the result of many improbable events that occurred during the time the Earth cooled from hot molten rock to its present crusted state. Life arose gradually as each succeeding event occurred (see the article LIFE for a detailed account of this process).

The life-supporting sphere

The biosphere is a system characterized by the continuous cycling of matter and an accompanying flow of solar energy in which certain large molecules and cells are self-reproducing. Water is a major predisposing factor, for all life depends on it. The elements carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur, when combined as proteins, lipids, carbohydrates, and nucleic acids, provide the building blocks, the fuel, and the direction for the creation of life. Energy flow is required to maintain the structure of organisms by the formation and splitting of phosphate bonds. Organisms are cellular in nature, always contain some sort of enclosing membrane structure, and all have nucleic acids that store and transmit genetic information.

Levels of organization. All life on Earth depends ultimately upon green plants, as well as upon water. Plants utilize sunlight in a process called photosynthesis to produce the food upon which animals feed, to provide, as a by-product, oxygen, which most animals require for respiration. At first, the oceans and the lands were teeming with large numbers of a few kinds of simple single-celled organisms, but slowly plants and animals of increasing complexity evolved. Interrelationships developed so that certain plants grew in association with certain other plants, and animals associated with the plants and with one another to form communities of organisms, including those of forests, grasslands, deserts, dunes, bogs, rivers, and lakes. Living (biotic) communities and their nonliving (abiotic) environment are inseparably interrelated and constantly interact upon each other. For convenience, any segment of the landscape that includes the biotic and abiotic components is called an ecosystem. A lake is an ecosystem when considered in totality as water, nutrients, climate, and all of the life contained within it. A given forest, meadow, or river is likewise an ecosystem. One ecosystem grades into another along zones termed ecotones, where a mixture of plant and animal species from the two ecosystems occurs. A forest considered as an ecosystem is not simply a stand of trees but is a complex of soil, air, and water, of climate and minerals, of bacteria, viruses, fungi, grasses, herbs, and trees, of insects, reptiles, amphibians, birds, and mammals.

Anatomy of the biosphere

Stated another way, the abiotic, or nonliving, portion of each ecosystem in the biosphere includes the flow of energy, nutrients, water, and gases and the concentrations of organic and inorganic substances in the environment. The biotic, or living, portion includes three general categories of organisms based on their methods of acquiring energy: the primary producers, largely green plants; the consumers, which include all the animals; and the decomposers, which include the micro-organisms that break down the remains of plants and animals into simpler components for recycling in the biosphere (see ECOSYSTEM). Aquatic ecosystems are those involving marine environments and freshwater environments on the land. Terrestrial ecosystems are those based on major vegetational types, such as forest, grassland, desert, and tundra. Particular kinds of animals are associated with each such plant province (see AQUATIC ECOSYSTEM; TERRESTRIAL ECOSYSTEM).

Ecosystems may be further subdivided into smaller biotic units called communities. Examples of communities are the organisms in a stand of pine trees, on a coral reef, in a cave, a valley, a lake, or a stream. The important consideration in the community is the living component, the organisms; the abiotic factors of the environment are excluded.

A community, in turn, is a collection of species populations. In a stand of pines may be many species of insects, of birds, of mammals, each a separate breeding unit, but each dependent on the others for its continued existence. A species, furthermore, is composed of individuals, single functioning units identifiable as organisms. Beyond this level, the units of the biosphere are the units of the organism: organ systems composed of organs, organs composed of tissues, tissues of cells, cells of molecules, and molecules of atomic elements and energy. The progression in the biosphere, therefore, proceeding upward from atoms and energy, is toward fewer units, larger and more complex in pattern, at each successive level.

From P.B. Weisz, *The Science of Zoology* (copyright 1966); used with permission of McGraw-Hill Book

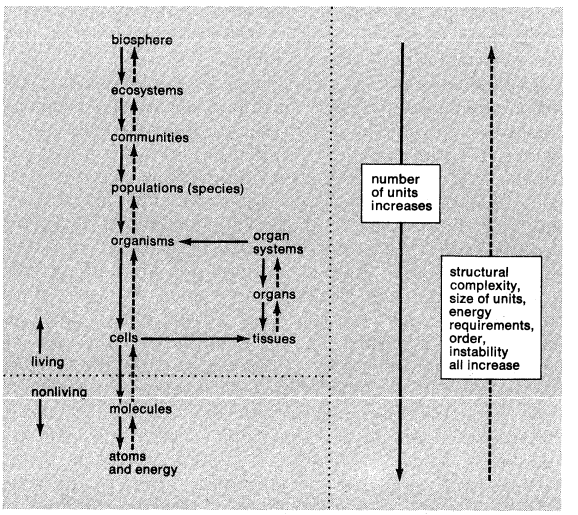


Figure 2: Hierarchy of levels in the biosphere.

ENERGY FLOW IN THE BIOSPHERE

Energy and organization. A unique characteristic of life is that it is an organized system capable of creating more order from less order. This seems contrary to the general trend of the universe, in which there is a tendency to move toward maximum disorder (entropy) as expressed in the second law of thermodynamics. Life results from the steady-state, or balanced, situation in which there is a flow of energy from the Sun to the Earth and then to the cosmic cold of outer space. The first law of thermodynamics, the conservation of energy, states that energy is neither created nor destroyed but remains constant for the universe. It does not indicate how energy is transformed but only that all the energy within the system must be accounted for.

The constancy of energy in the universe

In order for sunlight to organize life on Earth, it must irradiate the surface with electromagnetic frequencies harmonious with the peculiar chemical bonds of organic molecules. Whereas the very short-wave ultraviolet, gamma rays, and X-rays are destructive of organic molecules, and the long-wave infrared radiation is absorbed and dissipated as heat, the near-ultraviolet and visible wavelengths interact well with matter to stimulate the formation of bonds and the arrangement of organic molecules.

The single most important photochemical reaction in the world is photosynthesis, the union of carbon dioxide and water in plants through the interaction of sunlight and chlorophyll molecules. In the photosynthetic process, light energy is absorbed by chlorophyll to convert carbon dioxide and water into carbohydrate and oxygen. This photochemical event is a stepwise process by which electrons are energized within the chlorophyll molecule and raised to higher energy levels with the formation of carbohydrate. In the process, which is complex, high-energy phosphate bonds are formed, and adenosine triphosphate (ATP) results. The end products within green plants are carbohydrates and energy-rich compounds that become food for plant-eating (herbivorous) organisms and a succession of animal-eating organisms. The food web of ecosystems is provisioned in this way.

The importance of phosphorus

The universal fuel within all organisms is ATP, in which high-energy phosphate bonds store energy and release it when the bonds are broken. The complex reactions within a cell that lead to the formation of ATP combine at least one phosphorus atom per adenosine molecule. Without phosphorus no life could exist, since the entire linkage between photosynthesis and cellular activity

would be missing. One process by which the energy stored in ATP is utilized is in the transformation of glucose sugar to sucrose, or ordinary sugar.

Efficiency of utilization of solar energy. The intensity, as well as the composition, of the sunlight irradiating the Earth's surface is important to life. The tremendous amount of solar energy incident upon the Earth's outer atmosphere each day is distributed unevenly over the world, with the greatest amount of solar radiation reaching the desert regions and the least amount reaching the polar regions, where the slanted rays of the sunlight through the atmosphere are long. Temperate regions of the world, where food crops are grown, receive an intermediate amount of solar radiation, but a substantial fraction of this sunlight is received during the winter, when temperatures are too low for maximum plant growth. When intense winter cold grips the Arctic landscape, there is little light and no plant growth; only higher members of the food chain, birds and mammals, move about the land. In the tropics, temperatures are warm and constant throughout the year; the land is irradiated by sunlight more evenly throughout the seasons and plant growth is abundant. Although desert regions receive by far the greatest amount of sunlight, the lack of water limits plant productivity.

The conversion of energy into matter

Only about 25 percent of the sunlight reaching the ground is in wavelengths useful for photosynthesis, and only a fraction of useful light is available to green plants. An understanding of the performance of the biosphere requires an appreciation of the efficiency of the primary production process. The efficiency of energy conversion from incident sunlight into organic matter (net primary productivity) seldom is as large as 3 percent and usually is 1 percent or less. Estimates for a cornfield, for example, show that only slightly more than 1 percent of the total sunlight striking it ends up as part of the corn plants; about 44 percent is used in the evaporation of water from soil and plants, 54 percent is reflected or dissipated as heat, and a fraction of a percent is consumed by the respiration of plants and animals in the field. Not that all these other processes are not important, for indeed they are necessary to the proper functioning of the biosphere, but only a little more than 1 percent of the total incident energy ends up in new plant material. Other estimates of the net productivity of various vegetation types are:

Vegetation type	percent productivity
Tropical rain forest	1.4
Perennial grass–herb field	1.05
Coniferous forest	0.9
Deciduous temperate forest	0.4
Temperate lake	0.3
Desert	0.03

Animals grazing on the primary production convert only about 10 to 15 percent of the energy stored in plant material into animal tissue, and man eating the animals converts only about 1 percent of this energy stored in animal tissue into human tissue. The total conversion process from sunlight to plants to animals to man operates at an overall efficiency of about 0.001 percent.

Energy balance of organisms. Organisms utilize heat energy as well as light. All plants and animals must remain in reasonable energy balance. A plant leaf has a temperature that depends upon the amount of radiation absorbed, the flow of air over its surface, and its rate of transpiration, or gas exchange. The temperature a plant leaf assumes for any set of environmental conditions is of great importance to the plant, since all chemical reactions, including photosynthesis, are influenced by temperatures. A plant is coupled to its environment through the exchange of energy, gases, and nutrients.

A warm-blooded (homeothermic) animal remains in energy balance with approximately a constant body temperature. It moves about in its environment in such a manner that the amount of radiation it absorbs, the amount of energy it exchanges by convectional heating or cooling, and the quantity of energy it consumes by respirational water loss and evaporative cooling are such that its body temperature remains within a narrow range. A cold-blooded (poikilothermic) animal, on the other hand, has much less control over its body temperature. It nevertheless seeks the combination of radiation, air temperature, wind speed, and humidity that will result in an energy flow compatible with its body-temperature limits. Each and every animal in the world, depending upon its coloration, body size, shape, quantity of thermal insulation (fat, fur, or feathers), metabolic and water loss rates, and preferred body-temperature range and limits, has a well-defined climate space within which it must live in order to survive. Each species is restricted to its particular climate space as defined by its physiology and general body characteristics. Many animals will live in a somewhat restricted climate space, but none can live beyond its maximum climate space limits. The animals of the world are distributed accordingly (see DISTRIBUTION OF ORGANISMS).

From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

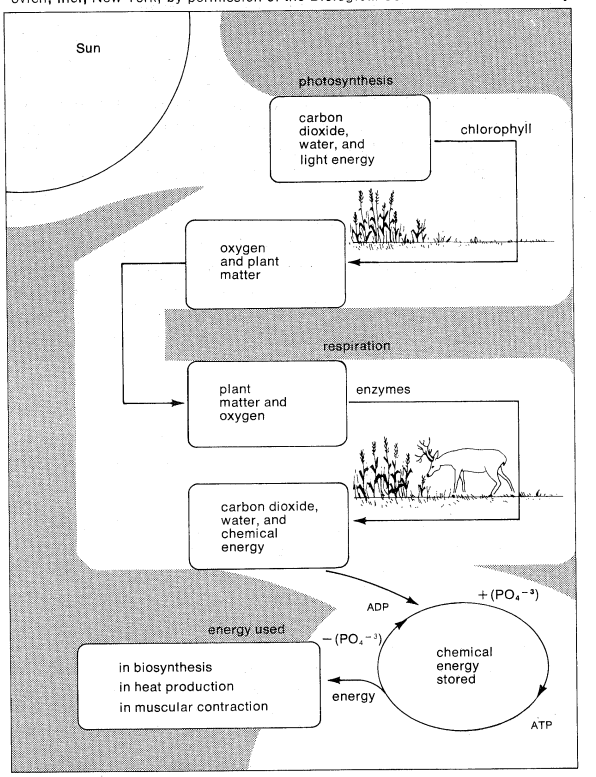


Figure 3: Energy relationships in the biosphere.

CYCLING OF MATTER IN THE BIOSPHERE

The general pattern of chemical cycles in nature. All life depends upon the cycling of matter (nutrients and water), as well as upon the flow of energy. Clearly, no organism can grow, propagate, and continue its kind for millions of generations without replenishing the elements that support it. Like birth and growth, death and decay are rules of the living landscape. Even the purely physical world has its cyclic processes of evaporation and condensation of water and the rise and erosion of rocks. In contrast to the unidirectional flow of energy through the ecosystem, whereby sunlight is absorbed by plants and heat is emitted to space at every conversion of energy from the eaten to the eater, any living and most nonliving entities emerge from the surface of the Earth only to return to their original point of origin in one form or another. Such a movement is called a biogeochemical cycle, in reference to the biological and geological phases of chemical substances, impelled by the Sun's energy.

All organisms contain water and utilize carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur in order to form carbohydrates, proteins, fats, lipids, and other structural materials. Most elements are obtained by plants through compounds such as water, H_2O; carbon dioxide, CO_2; nitrate, NO_3; ammonia, NH_3; sulfate, SO_4; and hydrogen sulfide, H_2S. Oxygen is taken in by animals

through respiration, bound up in organic products, and converted to water, in which form it is eventually returned to the environment. Plants utilize soil and water and, through photosynthesis, return oxygen to the air. Phosphorus, which is so desperately needed by organisms that its abundance or lack often limits populations, does not cycle easily in the ecosystem. Eutrophication is a condition of lakes that results from phosphorus enrichment of the water. Much phosphorus is lost from soils by erosion, and the sediment deposits in the bottom of the ocean keep phosphorus out of circulation for extended periods. Nitrogen, the most abundant of atmospheric constituents, is a relatively inert gas that most organisms cannot use directly but that is essential to life. Fortunately, certain species of bacteria and some blue-green algae are able to utilize gaseous nitrogen that diffuses into the soil from the atmosphere. These soil organisms convert nitrogen (N_2) into ammonia (NH_3). Some plants utilize the ammonia directly, but most depend upon the oxidation of ammonia to nitrite (NO_2^-) and then to nitrate (NO_3^-) by other soil bacteria before it is absorbed by plant roots. This process is known as nitrification.

Mineral elements such as calcium, sulfur, magnesium, potassium, and boron are taken up in solution from soils by the roots of plants and are returned to the soil by decomposers, organisms that promote breakdown and decay of the dead litter of forests or fields. Although rainwater slowly washes out, or leaches, some minerals from an ecosystem, it also brings in with it some nutrients during the precipitation process. As rainwater runs down mountains and hills it gradually picks up nutrients and makes them available to green valleys below. When man cuts the forest and exposes the soils to sun and rain, he accelerates the leaching process, often with disastrous consequences. In order to maintain high agricultural productivity he must then apply increasing quantities of synthetic fertilizers to replace those nutrients washed away.

The mixed blessing of rain

The carbon and oxygen cycles. The total biosphere contains approximately 20,000,000,000,000,000 (2×10^{16}) tons of carbon, mostly in the form of inorganic carbonates in the rocks and oceans, and in organic fossil fuel deposits such as coal, oil, and natural gas. The atmosphere contains 7×10^{11} tons of carbon in the form of carbon dioxide, and the green plants of the world contain 4.5×10^{11} tons as carbohydrates and other organic compounds. The exchange of carbon dioxide with the atmosphere by means of photosynthesis and respiration in plants results in a net annual productivity of the land of about 2.5×10^{10} tons per year and of the oceans of about 2×10^{10} tons per year. During the daytime, photosynthesis often produces a 12 percent drop in atmospheric carbon dioxide in the vicinity of plants, but at night, respiration by soil bacteria, plants, and animals often produces a 25 percent increase in carbon dioxide concentration near the ground.

Man, by burning fossil fuels, releases carbon dioxide into the atmosphere at a rate of 5 to 6×10^9 tons per year, which should increase the atmospheric concentration by 2.3 parts per million (ppm) per year; in fact, the increase is only 0.7 ppm per year. Much of the carbon dioxide released from fossil fuels may have gone into the oceans, and it is possible that a considerable fraction has been bound up in new plant growth. (This is not unlikely, since photosynthesis increases correspondingly, for most plants, with an increase of carbon dioxide concentration in the range of 300 to 400 ppm; only when the concentration approaches 1,000 ppm does the photosynthetic reaction rate approach saturation.) It would seem that man is inadvertently fertilizing the land and ocean by releasing to the atmosphere the carbon dioxide bound up in fossil fuels.

New plants from old

Carbon dioxide is extremely soluble in water. It is not surprising, therefore, that the oceans play a significant role in the global distribution of carbon dioxide and that rainfall sometimes contains about 0.3 cubic centimetre of CO_2 per litre of water. Carbon dioxide combines with water to form carbonic acid (H_2CO_3), which dissociates into hydrogen ions (H^+) and bicarbonate ions (HCO_3^-); the bicarbonate ions, in turn, dissociate into hydrogen and carbonate ions (CO_3^{--}). These reactions can be expressed:

$$CO_2 + H_2O \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons 2H^+ + CO_3^{--}.$$

These reactions are readily reversible, the direction depending upon the concentration of the components. The amount of carbon present as bicarbonate or carbonate depends upon the alkalinity or acidity (pH) of the water. If the water is alkaline, more carbon is present as carbonate than if the water is acid. High amounts of dissolved carbon dioxide in water produce high plant productivity (*e.g.*, algal blooms), but often this results in higher respiration, depleted oxygen, and subsequent fish kills. Acidic waters usually have low productivity.

Another gas of enormous significance to life is oxygen, which is cycled between the lithosphere, the atmosphere, and the biosphere. Plants are primarily responsible for the presence of atmospheric oxygen through the photosynthetic process. Oxygen in metabolism and in the production of energy-rich phosphorus bonds provides the power for all higher forms of life. Although oxygen is utilized within cell constituents such as mitochondria for the release of energy and the synthesis of ATP, other cellular bodies called peroxisomes appear to protect the cell from too much oxygen, which would result in destruction of the cell. Hence it seems that life has had to evolve within an environment in which some organisms utilize oxygen, some must be protected against oxygen, and some generate oxygen, all at the same time.

The nitrogen cycle. The atmosphere contains nearly 80 percent nitrogen. Ironically, most green plants are unable to use free nitrogen and must have it converted to soluble compounds of nitrogen, such as ammonia (NH_3), nitrite (NO_2), or nitrate (NO_3), which can then be taken up by their roots and the nitrogen converted into amino acids and plant proteins. The Earth's primitive atmosphere apparently contained ammonia, so the necessity for conversion of nitrogen into soluble products (nitrogen fixation) did not arise until more recently. Nitrogen fixation, or nitrification, is performed by a few species of micro-organisms. The reverse process—by which soluble compounds of nitrogen are reduced to molecular nitrogen (N_2)—is called denitrification and is accomplished by other micro-organisms. Micro-organisms that decompose the remains of dead plants and animals reduce amino acids containing nitrogen to ammonium ions and other products. This process is known as ammonification. The rate at which nitrogen fixation removes nitrogen from the atmosphere is almost balanced by the rate at which denitrification returns nitrogen to the atmosphere. (The large-scale and widespread use of fertilizers by man may be upsetting this balance, however.)

The indispensability of micro-organisms in the movement of nitrogen

Nitrogen is a versatile element in living processes because of its ability to form many different kinds of compounds and to release energy when moving from one compound to another. Nitrogen is cycled in nature from reduced inorganic compounds to oxidized compounds by atmospheric oxygen, with the release of energy; when oxygen is unavailable, oxidized nitrogen compounds can in turn oxidize organic compounds. Two main types of bacteria and algae participate in nitrogen fixation. One lives in a mutualistic partnership (symbiosis) with higher plants; the other is a free-living form that derives energy directly from sunlight and indirectly from plant materials. *Rhizobium* species, the most abundant of the root-nodule symbiotic bacteria, are found on the roots of legumes, alders, and buckthorns. Legumes, which include peas, soybeans, and alfalfa are often used in agriculture to increase the productivity of the land. Symbiotic nitrogen fixers seem to have a critical need for certain trace elements such as molybdenum or cobalt. When these are absent, the bacteria do not function effectively, and the rate of nitrogen fixation is reduced.

Symbiotic micro-organisms occur only in terrestrial ecosystems as far as is known. Among the nonsymbiotic nitrogen fixers are aerobic soil bacteria such as *Azotobacter*, which supply fixed nitrogen in grasslands and in marine and other ecosystems where symbiotic micro-organisms are absent. Blue-green algae are an important source of fixed nitrogen in rice paddies and in aquatic

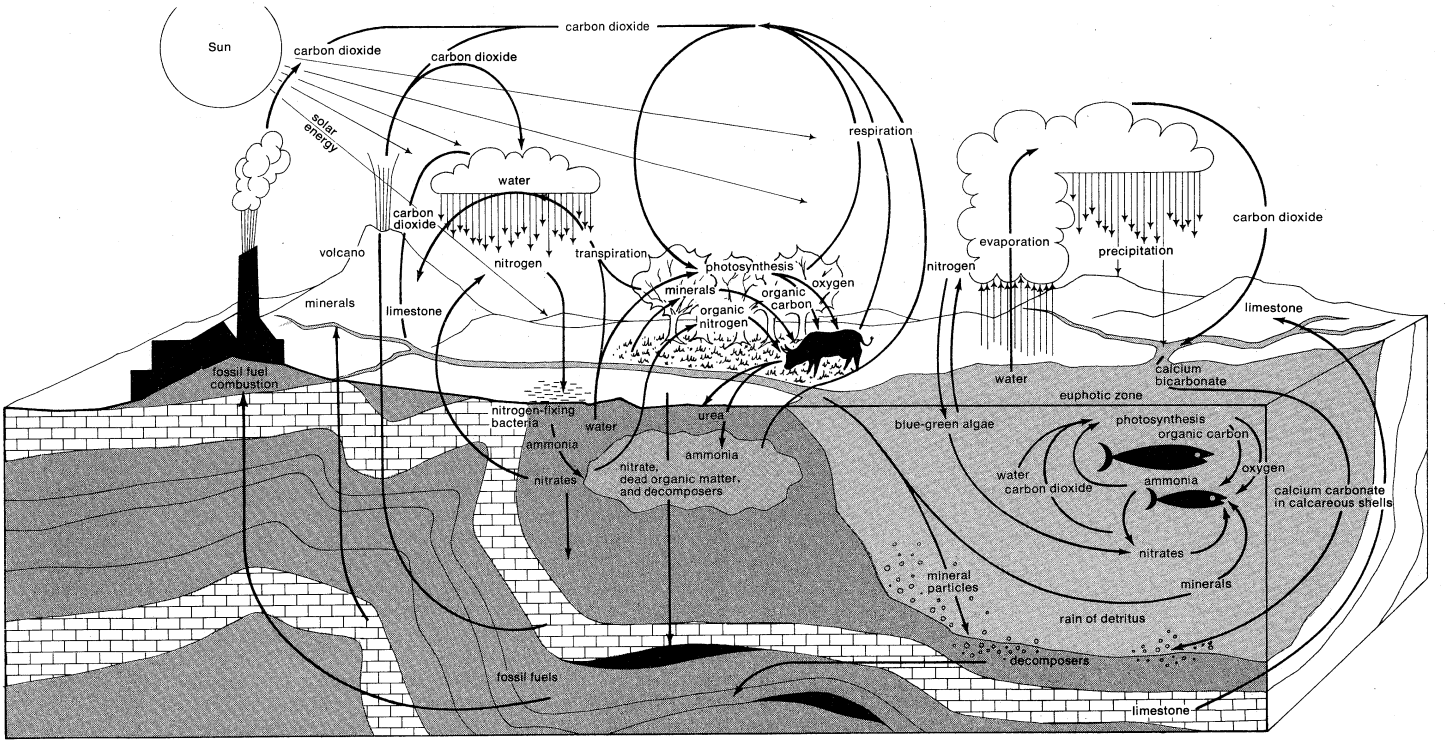


Figure 4: Major cycles of the biosphere.
From "The Biosphere" by G. Evelyn Hutchinson. Copyright © (1970) by Scientific American, Inc.; all rights reserved

systems. One free-living nitrogen-fixing bacterial genus lives only without oxygen is *Clostridium*.

A group of micro-organisms known as autotrophs function without an organic source of energy by oxidizing various forms of soluble nitrogen compounds. *Nitrosomonas* species oxidize ammonium to nitrite with the release of water and energy. *Nitrobacter* species, in turn, oxidize nitrites to nitrates with the release of energy. Nitrates are highly soluble in water and are easily taken up by roots and assimilated by plants.

Pseudomonas species and some fungi use nitrate as an oxygen source; they draw on the energy available in glucose and phosphate to convert the nitrate to molecular nitrogen and then to nitrous oxide and nitric oxide. Some denitrifiers reverse the process only partway by reducing nitrate to nitrite or ammonia. Denitrification proceeds best under anaerobic conditions, since whenever oxygen is present it is more efficient for organisms to use it than to use the oxygen bound up in nitrate ions. This suggests that denitrification would occur best in the soil and in deeper portions of the waters of the world. From the enormous amount of nitrogen fixed annually, however, about 92,000,000 tons, it is obvious there must be enormous anaerobic reserves in the world in order to return this amount of nitrogen to the atmosphere each year. It is entirely likely that man's use of fertilizers is causing more nitrogen fixation to occur than the biosphere can return through denitrification, in which case there would be a gradual buildup of nitrates, nitrites, and ammonia.

The nitrogen cycle is obviously extremely complex and very important in the balance of the biosphere. Many things about it are not yet clearly understood.

Mobile sulfur in the biosphere

The sulfur cycle. For materials to cycle easily throughout the biosphere they must be not only water-soluble but volatile as well. In addition to carbon and nitrogen, sulfur is a highly mobile constituent of the biosphere. Furthermore, all proteins incorporate sulfur in their molecular structure to form bonds that give them well-defined three-dimensional shapes. Just as carbon and nitrogen are reduced during the formation of proteins, so also is sulfur. In distinction to carbon, however, which requires green plants and sunlight for reduction, nitrogen and sulfur reductions are most often accomplished anaerobically by micro-organisms. Most of this anaerobic activity occurs in oxygen-deficient soils, bogs, and swamps. The sulfur cycle is not nearly so well understood as are the carbon and nitrogen cycles.

When an organism dies and decays, much of its incorporated sulfur is returned to a mineralized state by bacteria and fungi, but some is reduced under anaerobic conditions directly to sulfides, such as the evil-smelling hydrogen sulfide (H_2S), some of which escapes to the atmosphere. Man's burning of fossil fuels sends massive quantities of sulfur dioxide into the atmosphere as a pollutant. When the sulfur dioxide combines with water in rainfall it forms dilute sulfuric acid (H_2SO_4). Inorganic sulfate (SO_4), formed by the decomposition of organic matter, is readily soluble in water and serves as a further source of sulfur to plants and animals. Sulfate is reduced under anaerobic conditions to elemental sulfur or to sulfides by bacteria. Large quantities of hydrogen sulfide occur in the anaerobic (deeper) portions of aquatic ecosystems. The anaerobic bacteria utilize the sulfate in metabolic oxidation much as bacteria denitrify nitrate and nitrite.

There are sulfur-utilizing bacteria that play the same role with sulfur compounds as nitrifying bacteria do with nitrogen compounds. These include the green and purple photosynthetic bacteria; the green bacteria oxidize sulfide to sulfur and the purple bacteria generate sulfate.

Sulfur precipitates in lake waters as iron sulfide in the presence of iron under anaerobic conditions. Some sulfide is insolubly bound with iron in the bottom muds of lakes, bogs, swamps, etc., where other elemental metals, such as copper, cobalt, cadmium, and zinc, also become bound as precipitates.

The vital role of water

The water cycle. Water is a ubiquitous and unique substance necessary for life on Earth. Water in liquid and vapour states in the atmosphere regulates and ameliorates the climates of the Earth. Strictly speaking, water conforms to a gaseous cycle, but because of its great importance in the biosphere it is here discussed separately. Life began in the oceans, evolved in the waters of the world, and spread upon the land; yet life has always remained dependent upon the availability of water. The relative distribution and availability of water on the land determines the vegetative character of the landscape. Water erodes and sculpts the rocky surface of the Earth, transports nutrients and sediments, and forms the lakes, swamps, rivers, and seas. Sunlight evaporates water from land and sea into the atmosphere, where it is transported

with the global circulation of air and precipitated onto the surface as rain or snow.

The topography of continents and islands affects very strongly the precipitation pattern of the Earth. Windward mountain slopes are wet, as warm moist air rising into cooler air expands and condenses its moisture as rain or snow. The leeward sides of mountains are dry, since the moisture has already been wrung from the air on the other side of the divide, and the air moving down the leeward face compresses and warms, thereby retaining whatever moisture it still has. Semi-arid regions with sparse vegetation often occur to the leeside of mountain ranges. Cold polar ice caps have relatively low precipitation since very cold air can contain little moisture. Deserts generally occur where the air is stable. The trade winds move toward the Equator from cooler latitudes, picking up moisture and heat as they go; hence the coast lines of southern California, Mexico, and Chile are relatively dry, and the equatorial regions are wet.

Water has amazing properties, particularly when compared with most other forms of matter known in nature, that make it chemically and physically suitable for life. It is a liquid at ordinary temperatures, contracts on cooling down to 4° C (39° F), then expands on further cooling to the freezing point at 0° C (32° F). Solid water, or ice, is thus less dense than liquid water and floats on water. The fact that water expands on freezing makes it a powerful agent for the breakdown and fragmentation of rock into soil particles. Water warms up less rapidly and in turn cools more slowly than other substances. Lakes and oceans, therefore, have a different temperature from the adjoining landmasses, with a seasonal lag. Water tends to hold dissolved material in solution and also has the greatest surface tension of any known liquid. Moist air is less dense than dry air and rises above it, contributing to the weather dynamics of the atmosphere.

The global distribution of water

Taken over the world as a whole, the horizontal distribution of moisture must always add up to zero—the amount of water falling as precipitation is equal to the amount of water vapour taken up by evaporation. For any single region of the world, however, there may be a water surplus or a water deficit, differences that are theoretically balanced by ocean currents or river runoff. The Northern Hemisphere, for example, has an excess precipitation over the Southern Hemisphere, and the difference is redistributed by means of ocean currents. The drier the continent, the smaller the fraction of the annual precipitation that runs off. Australia and Africa, as a whole, are relatively dry continents in which large areas are covered by deserts and more than 75 percent of the annual precipitation is lost through evaporation.

In order for evaporation to occur, two conditions must exist. There must be energy available, which is derived largely from sunlight, and there must be a water vapour gradient from a moist surface to air that is less moist. In other words, heat and dryness aid evaporation.

Interrupting this cycle of simple precipitation and evaporation are the organisms of the world, which divert water for their own use. Water relationships are most dramatic in land situations. Soils hold considerable amounts of water. On the average, worldwide, this may be a reserve of about ten centimetres (four inches) in depth. In some parts of the world, such as North Carolina, where soils are thin, the total amount of water held in the soil may not exceed five centimetres (two inches), but in the deep volcanic soils of east Africa as much as 50 centimetres (20 inches) is held in the soil.

Plants take up water largely through their roots and evaporate water through the leaves in a process known as transpiration. A green plant may transpire three to five millimetres (about ⅛ to ⅕ inch) of water a day, depending upon the amount of energy available. Maximum rates in sunny, hot, irrigated regions may reach eight millimetres (⅓ inch) a day.

An important ratio for plants, known as the transpiration ratio, is the amount of water used to the weight of accumulation of dry matter in the plants. For corn (maize) it is 317 grams of water per gram of dry weight, for cotton 568, brome grass 880, deciduous trees 825, and evergreen trees 140. Taken over a whole growing season, although highly variable among crops, one can estimate that a production of 20 fresh-weight tons of crop will require 2,000 tons of water from the soil. Of the 20 tons of crop only about three tons comprise water molecules utilized in photosynthesis, the remainder being evaporated as transpiration.

In addition to their physiological need for water, large numbers of terrestrial animals use bodies of water as sanctuaries. The eggs of many insects are laid in water, larvae live in water, and the adults emerge from ponds or rivers to spend their adult lives on the land. Amphibians occupy a narrow zone between land and water, in which they can move from the harsher, more variable climate of land into the milder, more comfortable conditions of water. All land animals give off moisture when they exhale, and some animals pant or sweat to evaporate moisture and thus cool themselves.

The sedimentary cycles of essential minerals. While many elements give rise to gaseous compounds that are significant parts of biogeochemical cycles, some elements in the biosphere cycle primarily through water transport and sedimentation in bodies of water. For any soluble but nonvolatile compound a natural cycle is only possible when life is involved, since otherwise the compounds wash from the land to the rivers and oceans, where they remain in the sediments. Green plants pick up the compounds in the nutrient-rich soil water and convert them into plant cells where they may be passed to animals in the food chain; they eventually return to the soil through death and decay of plants and animals. Some compounds soluble in water are carried into the atmosphere by water evaporation and returned to the surface in rain.

The transport of essential minerals

The elements calcium, potassium, silicon, and magnesium each have important biological roles in the biosphere. Magnesium, for instance, is an essential element in the chlorophyll molecule, and calcium and silicon help form the hard parts of shells, bones, and teeth of animals. Iron, manganese, and sodium are present in organisms in minute, or trace, amounts, but nevertheless are vitally important. As mentioned earlier, phosphorus is probably the most important element among those that form nonvolatile compounds. A deficiency of phosphorus is most often responsible for poor crop production. Free phosphorus as such is not found in the atmosphere. If it were not for green plants absorbing its salts from the soil and transporting it to the leaves, phosphorus would not rise above the surface of the ground. So it is also for trace elements such as iron and manganese. Some trace elements such as vanadium, cobalt, nickel, and molybdenum are found primarily in aquatic plants, since they accumulate in bottom sediments.

Phosphorus is soluble in acidic waters; under special environmental conditions it is bound up as calcium phosphate or iron (ferric) phosphate. Phosphorus is extremely scarce in the geosphere, concentration normally being a few parts per thousand million. But great supplies of phosphorus are found in bird guano, the excrement of fish-eating gulls, cormorants, pelicans, and penguins, on the islands and in the ocean sediments off the coast of Peru. Artificial, or chemical, fertilizers are made from phosphate rocks and marine phosphates. Great quantities of these phosphates are used in detergents and wash into lakes and streams. Such phosphate enrichment creates rapid and excessive growth of algae, especially the blue-green algae. Increased decay and respiration result in oxygen depletion of the water and the suffocation of more sensitive species of fish, usually the game fish. This enrichment process, called eutrophication, usually results in a simpler animal and plant community, a shortening of the food web, and a less stable ecosystem.

The problem of phosphate enrichment

MAN'S PLACE IN THE BIOSPHERE

Man as an exploiter of nature. For perhaps 3,000,000 years or more, man lived in reasonable balance with the organisms about him. Parasites, disease, and the difficult search for food kept man's numbers low and he was on equal footing with other animals within the natural system. Although he utilized the plants and animals that

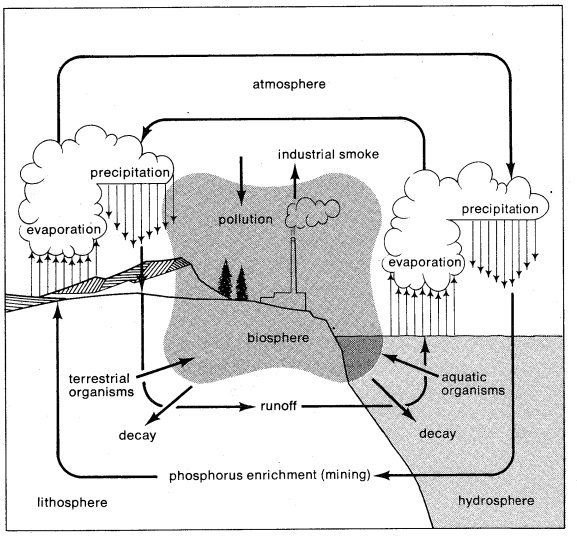


Figure 5: The intensive cycling (eutrophication) of phosphorus, nitrogen, and sulfur in the biosphere.
Adapted from E.S. Deevey, Jr., "Mineral Cycles," copyright ©1970 by Scientific American, Inc.; all rights reserved

surrounded him, the extent of his depredations was limited and reversible since his numbers were few. More than 10,000 years ago, however, man learned to select and cultivate plants, a progressive step that helped free him from the labour of bare subsistence and allowed him to engage in creative endeavours and to congregate in ever larger groups for mutual benefits. His heightened capacity for reasoning, his memory, and his ingenuity led him to improve still further his well-being. It was then that the consequences of man's activities began to extend beyond the borders of his limited fields and towns. He may have extinguished such Pleistocene animals as the woolly mammoth. His use of fire on the prairies appears to have maintained grasslands where otherwise trees might have grown. The smoke from these fires and from extensive slash-and-burn agriculture filled the skies with haze long before photochemical smog ever stung his eyes. By his ingenuity and inventiveness he learned to release energy by burning coal and oil, to manufacture machines that would greatly amplify the labours of his hands, to work metals and to forge new alloys, while at the same time selecting and cultivating better crops and improving his domesticated stock of animals. These various skills and practices eventually encouraged the world-wide proliferation of man.

The exploitation of resources

Man discovered the benefits of mining the great ore deposits, pumping the reservoirs of oil, digging the rich coal beds from carboniferous strata, plowing and planting the prairies, cutting the forests, and generally helping himself to the abundant and apparently limitless resources of the world. In the process he became sedentary rather than nomadic, he improved his health, extended his life-span, reduced his working time, added leisure hours, lived more comfortably, increased the abundance of his food supply, and increased in numbers exponentially (10 to 100 to 1,000 . . .). He could hunt, fish, harvest, and exploit without concern for the consequences to the environment. Then suddenly, very suddenly, within the most recent decades of man's time on Earth, he began to realize that resources were limited and hastily instigated some attempts to conserve the environment. Still, his behaviour generally belied any conviction that he should live in a prudent society. Man's population still doubles in number every 35 years; his demand for domestic products doubles every 14 years; and his consumption of energy doubles, in the United States, every 8.5 years and worldwide at a slower, yet rapidly increasing, rate. Man manufactures exotic products that are generally nondegradable (*i.e.*, they do not decay) but are carried by the sedimentation cycle to accumulate in the bottoms of lakes, rivers, and oceans or are piled in refuse heaps or are strewn about the landscape. Considering that these products consume resources that are nonrenewable, man's failure to recycle these materials may be seen as an act of irresponsibility toward the needs of future generations. The careless and greedy consumption of resources by any group of people cannot be justified in the context of the ultimate well-being of mankind.

When the first men landed on the Moon in July 1969 and reported during the voyage their views of the planet Earth, seen as a lonely, isolated, utterly self-dependent satellite of the Sun suspended in space, the ordinary citizen began to realize how limited and finite his space and resources really were. The biosphere was seen indeed as a very thin shell, protected from Sun and space by a veil of air and supported by a fragile and disturbed rocky crust. Coupled with this new appreciation was the realization that the world's population was expanding at an alarming rate, a rate that, if continued for another century, almost certainly would mean mass starvation, pandemic outbreaks of disease, and general degradation of society. It no longer was self-evident that man could manage such a complex society, keep it in good health, and maintain a high standard of living indefinitely.

Man as a manager of nature. If man expects to have a future on earth, it is clear that he must use his resources in the most prudent manner possible. Conservation does not mean hoarding; it means the wise management of resources to provide a continuing supply for a long time into the future. It suggests continual renewal of a resource and recycling, reusing, and recovering the products produced. Conservation of a natural area may mean its maintenance in a natural state for purposes of enjoyment or study in order to understand and appreciate the complexities of ecological laws. Man, of course, must continue to reserve forests for timber and other lands for agricultural production. But he must also set aside some regions for recreational purposes. The best use of each resource must be carefully determined. Once land is abused, it is very costly to reclaim it and put it to more suitable use. It is nearly impossible, for example, to return urban or industrial land to agricultural land use.

Man is not only a part of his ecosystem but is the most powerful force within it. He is simultaneously the most precious resource within the biosphere and the most dangerous. In many parts of the world the human population already exceeds the capacity of the land to feed it. By modern standards two-thirds of the world's population is ill housed, ill fed, and ill clothed. Through clever management of resources, including the management of his own numbers, man must improve the standard of living for all peoples or risk the slow degradation of mankind.

The growing need for new sources of energy

Man must further devise new energy sources to replace the coal, gas, and oil reserves that will have been exhausted, perhaps within 200 years. Nuclear energy is one of man's great hopes for the future; failure to achieve its potential in the form of a breeder reactor as a positive energy source would be one of the great tragedies in human history (see CONSERVATION OF NATURAL RESOURCES).

Man's ability to grow and survive in a complex world depends upon his ability to develop more productive, more adaptable, and more resistant strains of domesticated plants and animals through scientific principles of breeding (see ANIMAL BREEDING; PLANT BREEDING).

Diets in various parts of the world are very different from one another. Eastern peoples eat primarily plant foods, whereas Western peoples consume a considerable amount of meat. Grain consumption in India is about 350 pounds per person per year, of which 320 pounds per person per year are eaten directly. In the U.S., 1,700 pounds per person per year of grain is consumed, but only 200 pounds per person per year is eaten directly. Most of the grain is fed to cattle and indirectly is consumed by man as meat. Animals are such an important part of the Western diet that a great deal of selection and breeding has resulted in highly productive stock. Food is not the only reason man has selected and bred animals, for cattle, horses, and dogs are used for draft purposes, sheep for wool, goats for fleece, and cats and dogs for companionship.

Man's phenomenal success with breeding crop plants for

production around the world has become known as the "green revolution." Along with developing new hybrids it was necessary to produce a revolution in techniques of fertilizing, harvesting, storing, distributing, and marketing. Although man can come close to feeding the present world population, the chances of continuing to do so in the future are diminishing. If the world population ever approaches 12,000,000,000 (in 1971 it was about 3,706,000,000), it is likely that mass starvation will occur in many underdeveloped countries. Man will need to make the fullest use of potential food resources, both marine and terrestrial, if he is to support a larger global population than now exists.

Man has successfully selected specific strains of domesticated plants and animals and, in principle at least, can do the same with respect to the human race. Racial groups of man differ in various ways, and yet within each racial group there is a vast amount of diversity in, among other things, facial features, body build, height, shape, physical coordination, and mental traits. Racial groups, of course, overlap a great deal with respect to genetically determined traits that affect such things as blood substances, serum proteins, hemoglobin, and enzymes.

Average body sizes of man vary enormously, from the Pygmy to the tall Watusi. Facial features of Caucasians differ from Orientals. Some of these differences are simply the result of isolation of peoples, while other features have resulted from selective pressures of an environmental nature. Dark-pigmented skin is common in the tropics and serves to filter out excess ultraviolet radiation, some of which is required for the production of vitamin D by the skin. (Too much ultraviolet radiation, however, can result in excessive vitamin D production and calcium deposits that can cause crippling.) In temperate climates, on the other hand, the slanted rays of the Sun irradiate the surface with much less intensity, and inhabitants of such regions benefit from transparent skin, which allows sufficient amounts of vitamin D to be generated. It is possible that peoples whose members have long limbs and slender bodies have an advantage in the tropics over short stocky peoples because they can dispel heat quicker. By contrast the Eskimo has a thick, heavy-set body that seems better adapted to conserve body heat in the Arctic climate.

The mass production of men

Man could select and breed for particular characteristics by working with each consecutive generation of persons, but the process would be slow. Much more efficient and more revolutionary ways of selecting for specific human types now loom as possibilities. One method, nuclear implantation, has been successful in remaking amphibians. Unfertilized eggs are removed from the oviduct of a female, the nuclei of these removed and replaced by the nuclei of body cells of chosen animals. The eggs are then replaced in the oviduct of the females from which they came, the embryos develop, and normal birth occurs. The result is the replication many times over of identical animals, each having the traits selected for; the entire procedure is accomplished within one generation. Whether man can do the same with his own species is a moot point. To most people the idea is repulsive, yet the fact remains that man is today on the threshold of guiding his own evolutionary destiny and may perhaps someday be forced to take such measures.

The need for recognition of personal responsibility

Epilogue: the prospect. The point has been made abundantly regarding the untoward effects of man's activities on the biosphere. Past centuries of public and private greed may now cost man centuries of deprivation. Hope diminishes with each decade. There is a persistent personal unconsciousness regarding the numerous individual actions that affect nature; these seemingly insignificant events, when multiplied by man's numbers, constitute a tremendous assault on the biosphere. As a civilization develops, its technology improves and the needs of its citizens become displaced by wants, simple unalloyed desires for the trappings of technology.

The burden must be shared by both private and public sectors of society. Industries and institutions, which historically are heavy despoilers of the environment, must be held accountable for their actions; they must realize their responsibility to the continuance of the biosphere. But man as an individual must learn to live with fewer unnecessary items, so as to drain less heavily on the Earth's limited resources.

BIBLIOGRAPHY. E.P. ODUM, *Fundamentals of Ecology*, 3rd ed. (1971), a general ecology text, at an intermediate and advanced level; E.J. KORMONDY, *Concepts of Ecology* (1969), an introductory ecology text, limited in coverage but very good; "The Biosphere," *Scient. Am.*, vol. 223, no. 3 (1970), an entire issue devoted to the ecology of the biosphere, with excellent general articles; NATIONAL ACADEMY OF SCIENCES, NATIONAL RESEARCH COUNCIL, *Resources and Man* (1969), an excellent summary report on natural resources, including coal, oil, nuclear energy, and man's demand for power; PHILIP HANDLER (ed.), *Biology and the Future of Man* (1970), an analysis of modern biology with a perspective for the future of research in all fields of biology and, to a more limited extent, in medicine; N.M. JESSOP, *Biosphere: A Study of Life* (1970), a college text with emphasis on the integration of living things and their environment; W.L. THOMAS, JR. (ed.), *Man's Role in Changing the Face of the Earth* (1956), a record of a symposium, consisting of several definitive essays of remarkable insight into problems of the biosphere.

(D.M.G.)

Biotic Interactions

In a biological community no species lives in isolation from other species. Plants grow together in spatial patterns determined in part by competition and are fed upon by characteristic herbivorous animals, which in turn serve as food for carnivorous animals. Both plants and animals are attacked by parasites. Animals interact with each other competitively and sometimes beneficially. Some of these interactions are temporary, casual, and of minor importance, whereas others are permanent, vital, and of major significance.

Biotic interactions may occur between members of the same species (intraspecific interactions) or between two or more species (interspecific interactions). They may involve nutritional benefits, space in which to live, shelter or protection, transport, or reproductive capability. Many interactions are highly specialized and complicated, involving adaptive changes in structure, function, behaviour, and ecology. Some create negative effects upon the interactants, reducing survival or reproductive success; other interactions produce positive effects, increasing survival or reproductive success. In some cases, members of an interaction are apparently unaffected.

Biotic interactions are significant not only because they influence individual species but also because they constitute the principal stabilizing, connective linkages among the various species contained in a biological community. In a general sense, species in a biological community persist in relative harmony because of biotic interactions, despite individual gains and losses. As a consequence of this web of interactions, the community as a whole persists, with all species ultimately contributing to its continuation.

INTERACTIONS WITHIN A SPECIES

The individuals in a population often interact with other members of the same population. Such intraspecific interactions can be either negative, as in competition for limited resources, or positive, as in group protection, parental care, and social integration. Because these matters are of such great concern in biology, they are generally treated under behavioral studies; for details see the article SOCIAL BEHAVIOUR, ANIMAL.

Negative interactions. Intraspecific competition provides the ultimate force limiting the abundance of a species in a community. Natural enemies may be present or absent, but others of the same species are always present, providing the competition that insures against overexploitation of available resources and the irreversible destruction of a community due to overpopulation.

Cannibalism, fighting, and territorial defense are some of the overt methods by which competition manifests itself. Many animals, both herbivores and carnivores, exhibit cannibalism when overcrowded, and some are cannibalistic whenever they contact another of their own

species. The codling moth caterpillar in the apple is of the latter sort; hence, there is never more than one worm per apple. Mice and rats, when crowded, undergo psychological disturbances leading to excessive fighting and litter destruction. Territoriality, the dividing up of resources, mostly space, is a common means of defense against overcrowding. It differs from simple aggression in that, instead of involving individuals against individuals, it involves groups of individuals, breeding pairs, families, or herds. Migration away from overcrowded conditions is another common intraspecific response. It may be individualistic, with each member of a population going its own way, or it may involve mass flight, such as occurs with locusts, butterflies, and birds (see MIGRATION, ANIMAL).

Migration

Positive interactions. On the positive side, numerous interactions among individuals of the same species lead to clustering, aggregation, swarming, herding, and, ultimately, development of social groups. Tent caterpillars live together in large colonies, foraging together in masses and resting at night in communally spun silk tents. Isolated caterpillars are unable to survive; only the aggregation thrives. Parental care provides the basis for many aggregations of like individuals. Parental protection of the egg, as occurs in birds, some insects, and some fish, is the simplest form of care. Feeding and protection of the young is a more advanced level of care. Living together in social groups, in which one reproductive adult, the queen, is served by her own progeny, is the ultimate in parental care. In such cases the offspring remain with the parents, providing protection for the colony, care of younger members, nest building and cleaning services, and food gathering and storage. Division of labour to attend to these services is common, particularly in insect societies.

Some species live in habitats where—because of small size and weak flight ability, limited and scattered resources, and relatively low population densities—the chances of the mature sexes meeting for reproductive purposes are low. One method for overcoming this hazard is the phenomenon of autoparasitization, in which one sex lives as a virtual parasite on the other, an unusual form of intraspecific interaction. Certain insects exhibit this mode of life: the males of scale-insect parasites develop as hyperparasites on their own females. Some deep-sea angler fish also employ this habit, with the very small adult male living in permanent attachment to the body of the larger female. In such situations the association of the sexes is facilitated, and reproduction in hazardous environments is insured.

THE RANGE OF INTERSPECIES ASSOCIATIONS

Positive, negative, and neutral effects

One system of analyzing two-species interactions assigns positive (+), negative (−), or neutral (0) value to the interactants on the basis of how they affect each other. If the survival or reproduction of a species is enhanced, it is a gain and given a plus value. If survival or reproduction is reduced, it is a loss and given a minus value. If survival or reproduction is unaltered and the species unaffected, it rates a zero value. From such a rating scheme, a two-species (A, B), three-effect (+,0,−) table can be constructed in which the bases for association (symbiosis) of all biotic interactions can be defined (see table).

Mutualism (+,+) is an association of two different species that results in mutual benefit or gain. Obligative mutualism requires that both members interact together, or neither survives. Facultative mutualism, or protocooperation (+,+), refers to a less-rigid association in which one or both members can survive in the absence of the other. In commensalism (0,+) one member benefits while the other remains unaffected. Cases in which the small or weak species is unaffected while the larger or stronger benefits, especially in nutritional relations, is sometimes called allotrophy (+,0).

In parasitism (−,+) the smaller organism benefits at the expense of the larger. Many parasites are microorganisms that cause disease, but not usually the death, of the host. A special case of parasitism, which involves insects that do kill their hosts, is called parasitoidism.

species A (large, strong) \ species B (small, weak)	+	0	−
+	+, + interdependence (obligative mutualism) protocooperation (facultative mutualism)	+, 0 commensalism	+, − herbivory predation
0	0, + commensalism	0, 0 neutralism	0, − amensalism (competition)
−	−, + parasitism parasitoidism	−, 0 amensalism (antibiosis)	−, − mutual antagonism

Predation (+,−) refers to an association in which a large or strong animal consumes a small or weak animal. Herbivory (+,−), the consumption of plants by animals, is generally comparable to predation. Amensalism (0,−) occurs when one organism loses while the other organism is unaffected. Unilateral competition, to which the term competitive exclusion refers, and antibiosis, in which one species remains unaffected while a competitor is deprived, are examples of amensalism. In antagonism (−,−) both species lose or are harmed. Neutralism (0,0) is an association in which neither species is affected by the presence of the other, the two species sharing a localized resource or habitat.

The following sections will provide selected examples of the various forms of biotic interactions, based on the definitions above. The emphasis rests on the pattern of the interactions rather than the symbiotic associations themselves.

CONSUMPTION: ORGANISMS EATING ORGANISMS

Herbivory: animals eating plants. Herbivory, the consumption of plants by animals, constitutes one of the most important classes of biotic interactions. It serves as the main connection between all plant and animal life, the basic food association (trophic link) through which the composition, dynamics, and variation of community structure takes place. All higher organisms, primary and secondary carnivores and scavengers—essentially all the remaining links in the community food chains and food webs—are more or less dependent on this basic link.

The basis of community food webs

Herbivory is encountered at all levels of animal life, from the smallest to the largest animal, and in all habitats. No plant species is exempt, whether in the Arctic tundra, the tropical forest, or the desert and whether on the highest mountain, along the seashore, or in midocean. It is a reciprocal interaction, or coaction, wherein the number or quantity of plants is altered by the number and consumption rates of the herbivores and in which, conversely, the number of herbivores is limited by the quantity of plants. This coaction, therefore, may approach a state of balance in which the amount of plant material is just sufficient to sustain the number of herbivores associated with it, and the herbivores collectively remove only the surplus plant material produced. The manner in which specific herbivores respond to the results of this interaction is dependent on their food habits. Species that require one kind of plant material are very sensitive to changes in abundance of their preferred food, whereas less-specialized species simply shift their attentions to other plants when the quantity of a given one fluctuates.

The degree of herbivory is rather wide. A plant may be totally consumed, as it is when a protozoan or a fish consumes unicellular planktonic algae. Vital parts of a plant may be eaten, leading to the eventual death of the individual, as occurs when bark beetles girdle a pine tree or when birds eat the tender tips of germinating plants. When only nonvital parts of a plant are eaten—including leaves, twigs, and roots—the plant may be stunted or not affected at all. In the latter case, which is quite common, the herbivore is truly living on "surplus" plant life.

Herbivory is one factor responsible for limiting the

amount of plant life in a community and, therefore, influencing the composition of communities. It serves as an important means of recycling plant tissues back into the nutrient minerals, carbon dioxide, and water from which they were derived and enters into the determination of the numbers and kinds of animals associated with the community.

Many biologists contend that herbivorous organisms rarely influence the abundance of plant life notably, since plants manage to survive, or persist, without noticeable injury in communities, and starvation does not seem to play an important role in the survival of herbivores. Agricultural experience, however, suggests that substantial interactions between plants and herbivores do occur, since plant pests can drastically reduce the growth or productivity of crop plants. Successful attempts to use specific herbivorous insects to control certain weed species (*e.g.*, cactus in Australia or the poisonous Klamath weed in California) and the remarkable recovery of plant populations when an adverse factor is removed (*e.g.*, a pest) or a stimulating factor is added (*e.g.*, nutrients) suggest that herbivore–plant interactions are of critical importance to community composition. Plant–animal interactions are also illustrated in the overgrazing that occurs when deer, protected against predators, increase to such numbers that a plant community is threatened by excessive browsing. Grazing by stock animals, sheep, and cattle is an important factor in the composition of range and pasture lands; if excessive, grazing can actually destroy such communities. A rapid increase in plant growth brought about by excessive nutrient deposition (eutrophication) in bodies of water through runoff of soil fertilizers and discharge of municipal and industrial wastes results in "blooms" of aquatic algae and the disruption of the normal trophic balance between algae and algal feeders.

Herbivores to control weeds

The concentration of a few plant species into large areas, as is the general practice in agriculture, provides a potential for excessive herbivory. Such large tracts of single crops (monocultures) concentrate and augment the numbers of herbivores (when insects, they are usually called phytophages) to the point that massive population outbreaks commonly occur. By contrast, natural communities, which are diverse in species, are much less prone to support pest outbreaks: an expression of the ecological principle of stability in diversity in community life.

Predation: animals eating animals. Just as most plant species are fed on by herbivores, so are most animal species subject to predation by other animals. By feeding upon and reducing the numbers of both herbivores and other carnivores, predators constitute the higher links in the nutritional pathways characteristic of communities. Predators substantially affect the numbers of herbivores and thus influence their impact on plants. When the numbers of predators fluctuate, their effects upon prey and plants vary accordingly.

Quantitative effects of predators on prey

The quantitative effects of a predator population upon a prey population are determined by the number of prey consumed per predator and the total number of predators. In areas where prey are abundant, they become easier to find, more are consumed, and predator numbers increase through augmented reproduction and concentration. As predators increase, they eventually cause a reduction in prey numbers. Eventually, because of the biotic interaction between the two species, the number of predators also declines. This constitutes the predator–prey oscillation characteristic of many animal populations. An example is the cyclical pattern of abundance shown by the lynx and the snowshoe hare in the northern coniferous forests. These species fluctuate with a periodicity of about ten years. Biologists agree that predator density follows and is very likely determined by prey density. The opposite conclusion, that predator numbers determine prey numbers, has been less generally accepted. More commonly accepted are the predator-induced oscillations of insect prey. The biological reason for this is that insect predators commonly attack and destroy immature prey individuals, so that the effects of predation are more important in determining the population status of the prey. Vertebrate predators, on the contrary, seem to concentrate on the weak, sick, or old, the effects of which are less important in the population dynamics of the prey.

Because of the obvious importance of and interest in the mechanics of predation, predator–prey interactions have been the subject of much scientific research. Laboratory and theoretical studies have been made of isolated predator–prey "systems" and of predator–prey "models." The degree to which such systems and models simulate natural predator–prey population fluctuations has been used as a basis for judging the "soundness and realism" of the scientific methods and for "explaining" the process of predation. Simple mathematical expressions have been devised that represent predator and prey numbers and indicate that predators vary in numbers in a smooth, oscillatory manner purely as a result of prey abundance changes, and vice versa. This latter result, that prey oscillate due to predation, is the important new "evidence" that prey–predator interactions are, indeed, reciprocally coupled.

The importance of predation in determining prey abundance and community stability can be illustrated by several examples. When man (a predator) overexploits his animal resources (in effect, prey), they decline in numbers sufficiently to affect subsequent utilization (consumption). Such is the result when fish, whales, sea otters, buffalo, and other animals are excessively hunted. Whales and sea otters have declined drastically; buffalo became exterminated as a wild animal, and grizzly bears, mountain lions, and elks similarly are threatened with extinction.

Predators to control pests

An illustration approaching the point from the opposite side is the employment by man of predators to control pests, a technique called biological pest control. The effective suppression of the cottony-cushion scale pest of citrus in California in the late 1800s by use of an imported lady beetle, *Rodolia cardinalis*, is a classic example. Since that time more than 100 successful cases of biological control, leading to permanent suppression of pests, have been accomplished.

The unintentional unleashing of new pests through the use of pesticides, which cause the destruction of natural-enemy predators, illustrates the delicate balance between predator and prey that exists in most communities.

Occasionally, outbreaks of predators occur, with resultant drastic impact on prey populations. An example of such a case is in the Great Barrier Reef, Australia, and near Guam, in the western Pacific, where the crown-of-thorns starfish (*Acanthaster planci*), a predatory species, has increased in numbers that have seriously reduced the abundance of many corals. Sometimes the predator attacks a valuable resource of man. The whelk, a marine snail, attacks oysters. An American whelk, *Urosalpinx cinerea*, has invaded oyster beds off southern England, causing severe reductions in harvest since the late 1950s. And man has always been uneasy about both bird and mammal predators—such as hawks, eagles, coyotes, wolves, and foxes—that attack his domesticated animals.

PARASITIC INTERACTIONS

Parasitism. In practically every community, plant and animal populations are burdened with parasites, which include viruses, fungi, protozoans, nematodes, many marine and freshwater coelenterates, flukes and tapeworms, leeches, some insects, certain birds, and higher plants such as dodder, mistletoe, and broomrape. A parasite usually does not kill its host (if it does, the parasite itself may die before reaching a new host), but it generally reduces the growth rate, survival ability, and reproductive capacity of the host. From an ecological standpoint, therefore, parasitism does constitute a considerable burden on a host population.

Ectoparasites and endoparasites

Parasites that live on the body surface of the host are called ectoparasites. They do not commonly cause disease in their hosts but rather suck blood or create superficial damage to the skin; examples include leeches, fleas,

lice, and ticks. Other parasites live inside the host's body —either in cells or in spaces lined by cells (*e.g.*, intestine, blood vessels, mouth, etc.). Parasites that live within host cells—such as many bacteria and viruses—are called intracellular endoparasites; those that inhabit spaces within the host's body are called intercellular endoparasites.

Many disease-causing organisms, or pathogens, are endoparasites carried from host to host by some other organism; malaria, for example, is caused by a protozoan endoparasite transmitted by mosquitoes. In such situations, the biotic interaction involves three necessarily coexisting species: the pathogen, the carrier (or vector), and the host. A number of plant diseases transmitted by insects are dependent on carrier density, pathogen abundance and virulence (disease-producing power), and host density. Host susceptibility is also a significant factor in influencing the outcome of host–parasite interactions. In Europe the Dutch elm disease fungus (*Ceratocystis ulmi*) is a minor pest of elm trees. When this pathogen established itself in North America, however, it severely infected native American elms, which were more highly susceptible than their European counterparts. A similar event occurred when the chestnut blight, a fungus native to chestnuts in China and Japan, invaded North America and devastated stands of native chestnuts.

Not all invading parasites are micro-organisms, however. The sea lamprey, an ectoparasite on fish, attaches itself to its host by means of a sucker-like mouth with which it sucks blood and soft tissues. The sea lamprey is native to the North Atlantic, living in the ocean but breeding in coastal rivers and streams. It was able to move into Lake Ontario but no farther, being blocked by Niagara Falls. After construction of the Welland Canal, however, which bypassed the falls, the lamprey invaded the remaining Great Lakes and began decimating the populations of commercially valuable fishes.

Intermediate hosts

An intermediate host—interposed between the target host and parasite—adds ecological complications to host–parasite interactions: the parasite, host, and intermediate host must all be present in the same community in suitable numbers at the same time. The liver fluke that attacks sheep and cattle, for example, requires an aquatic snail as an intermediate host; there is no direct transmission of the fluke from sheep to sheep. Sheep, snail, fluke, pasture, and aquatic habitat must all coexist in the same community. Many other common parasites similarly require an intermediate host; *e.g.*, malaria, sleeping sickness, wheat stem rust. One reason that such complicated three-component interactions are so common is that with at least one of the organisms the reproductive capacity is very high; this increases the likelihood of continuation of the pathogen. Furthermore, in the case of parasites carried by insects, the motility and numerical abundance of the insect carrier facilitates parasite dissemination.

Parasitoidism. Parasitoids—insects that parasitize other insects—differ from true parasites in that, like predators, they always kill their hosts. Most are free-living in the adult stage. In a typical life history, the female parasitoid lays an egg in or on a host; the ensuing larva feeds upon host fluids and tissues, eventually killing the host; and the fully developed larva pupates and later emerges as an adult. Most species of parasitoids have a distinct preference as to the stage of host attacked, whether egg, larva, pupa, or adult.

Parasitoids are among the most important agents in limiting the numbers of other insects. For this reason they, along with true insect predators, are frequently sought and used by man as control agents for insect pests. Almost all insect species are attacked by one parasitoid or another. Some parasitoids are gregarious, with several living in one host simultaneously; others are solitary, with only one developing in a host. Certain wasp parasitoids exhibit polyembryony, wherein one egg deposited in a host may give rise to dozens or thousands of parasitoid embryos, and the host is literally overwhelmed by the developing parasitic larvae. In some parasitoid flies the female lays her eggs on leaves, and the host becomes infected by eating the egg-bearing leaves. Other parasitoid flies lay their eggs on the backs of hosts, and the larvae hatch out and bore into the host body.

The parasitoid's life cycle is always well attuned to that of its host. If the host is a leaf feeder and found only in the foliage of trees, that is where the female parasitoid searches for them. If the host has but one generation a year, so usually does the parasitoid. If the host passes the winter hibernating in the soil, the parasitoid generally does likewise, emerging from the soil in the spring soon after the host does.

Hyperparasitism

A parasitoid may itself sometimes be parasitized. In such cases its parasites are called hyperparasites. The hyperparasitic habit, commonly found in insects, may extend to several levels; for example, an aphid can be attacked by a solitary, endoparasitic aphidiid wasp, which, in turn, may be parasitized while still within the host's body by an endoparasitic cynipid wasp. Moreover, the cynipid can itself be attacked by an ectoparasitic pteromalid wasp. In theory even additional levels of parasitism are possible.

The free-living habit of the adult parasitoid enables it to spread rapidly from host to host throughout a community. Because the host almost always dies as a result of a parasitoid attack, few host-immunity responses have evolved. Only when parasitoids attack unusual hosts—those not in their normal host range—can a host survive and sometimes elicit an immune response to further attacks. The true parasite and the parasitoid can sometimes interact in their association with a host insect. Some parasitoid wasps coincidentally distribute pathogens from host to host when they deposit their eggs.

Brood parasitism. Brood parasitism, the laying of eggs in another animal's nest to be reared therein by the host, occurs in certain birds, including the cuckoo and the cowbird, and in certain insects, including cuckoo wasps. The female cuckoo bird lays an egg in the host's nest, where the hatched cuckoo chick either kills or pushes out the host's brood, so that it is reared alone until it can fly. The cuckoo wasp accomplishes much the same result, except that its host is one of the social wasps or bees. It enters a colony and lays an egg in one of the compartments, or cells. After hatching, the parasitic larva destroys the normal cell occupant and is taken care of by the colony workers.

The cuckoo habit is a balanced one, not being so efficient as to suppress unduly the host population but effective enough to assure perpetuation of the habit. In most cases, the reproductive habits of the host species enable progeny to be produced at times when the brood parasite is not active.

AMENSALISM AND ANTAGONISM

Competition. Competition occurs when two species utilize a common, limited resource (such as food, space, or moisture), and, in so doing, one species interferes with, injures, or deprives the other. When the competitors are of the same species, intraspecific competition occurs (dealt with earlier). Populations of two species cannot persist together for very long in the same community when both compete for and are limited by a common resource. This principle, known as Gause's hypothesis, further suggests that the species having the greater competitive advantage—manifested in greater searching ability, more rapid exploitation of the resource, greater numerical growth rate per unit of resource attained, greater aggressive or fighting ability—in time gains more and more of the limited resource. Competitive interactions between species are resolved in either of two ways: (1) the exclusion or displacement of one species by the other; or (2) the coexistence of the two competing species as a result of minor differences in habit or because their numbers are limited by factors other than the resource under consideration. When a species invades the habitat or community of another species and both occupy the same niche (the totality of its resource needs, its habits, and its mode of life), either the incumbent will persist and the invader will disappear, illustrating competitive exclusion, or the invader will become established and the

Gause's hypothesis

incumbent will disappear, resulting in competitive displacement.

Actual competition is difficult to see in nature, since a community usually contains "superior" competitors only, the lesser adapted species having already been ejected from the area. When a species invades a new habitat, however, either through its own dispersive powers (rare) or through man's activities (common), competition can be observed in operation. Hence, when settlers established themselves in western North America and brought with them grains and forage grasses for their livestock, they also brought, unwittingly, many weed seeds from Europe. Many of those weeds and foreign forage plants became established, forcing out or eliminating many native plants. The Klamath weed (*Hypericum perforatum*) is a remarkable example of the way in which weedy competitors can succeed in a new habitat. Gaining a foothold in California a few years before 1900, Klamath weed established itself as the dominant perennial on low-elevation rangelands. Several decades later, millions of acres of land were infested. Only when the weed itself was suppressed by biological control agents in the late 1940s were the displaced native species able to return to their old habitats. In England the shrub *Rhododendron ponticum*, an invader from the Continent, has become a pest in wooded areas, displacing holly and preventing regeneration of oaks and other desirable trees.

Insect invaders

Many insect invaders have displaced incumbent species through competition. The Argentine ant (*Iridomyrmex humilis*) and the imported fire ant (*Solenopsis saevissima richteri*), both invaders of North America from South America, have displaced many native ant species. When the fruit fly *Dacus dorsalis*, the maggots of which feed on tropical fruits, invaded Hawaii in the mid-1940s, it apparently displaced an earlier invader, the Mediterranean fruit fly *Ceratitis capitata*, from all the low-elevation habitats that the latter had occupied for so many years. The Mediterranean fruit fly was forced up into higher elevations on volcanic slopes, where coffee remained its major host, leaving the lower slopes and shores to the newer invader.

In aquatic habitats, aggressive fish such as carp and perch commonly displace trout and other game fish. When the long-legged crayfish (*Potamobius leptodactylus*), a native of eastern Europe, was introduced into some central European lakes, it totally displaced the native broad-legged crayfish (*P. astacus*).

Laboratory experiments also show the nature of competition. When two species of the single-celled protozoan *Paramecium* compete for food and space in small glass tubes, one displaces the other, even though both thrive when grown alone. Similar results are obtained when two flour beetles, *Tribolium confusum* and *T. castaneum*, are cultured together in jars of medium. Either species, living alone, limits its own population numbers (intraspecific competition) through such self-regulatory actions as cannibalism (adults and larvae feed on eggs and pupae), interference with female egg laying, and defilement of the common food supply. When populations of the two species are grown together, one species is eventually eliminated; the other persists. The interesting difference in these experiments is that, possibly because of slight variations in genetic composition of the organisms used, the same species does not always survive in every experiment. Rather, there is an element of probability as to the outcome. The probability of winning the competition can be varied, sometimes substantially, by altering the physical conditions for growth, such as temperature, moisture content, or humidity.

When two species of closely related water fleas, *Daphnia pulicaria* and *D. magna*, are cultured together with algae as food, both species increase in numbers for about three weeks; then the *D. magna* population ceases to grow and instead gradually declines to extinction at the end of the sixth week. The competition for the common food supply, not important while the populations were small, soon becomes severe enough to affect strongly the less adapted of the two competitors. On a different food supply, yeast cells, *D. magna* actually increases at a greater rate at first, but after about 24 days it reaches a peak and declines to extinction, as in the previous case.

Coexistence of potentially competitive species

It often happens that two species with similar niches occur in adjacent geographic areas and partially overlap. Coexistence in the area of overlap is then possible as a consequence of the continuous migration of the "excluded" species from its area of exclusivity. Displacement takes place, but the dispossessed are constantly being replaced by new immigrants into the "overlap" area. Eventually, if the area of overlap is broad enough, niche separation may occur through the process of evolution. Niche separation, or displacement, is the shifting of one or the other competitor species from the common, limited resource to another resource not competitively limited. This can be a different food, different spatial habitat, or different nesting site. An example of niche displacement is that of the finches on the Galápagos Islands. Some of the 12 or so finch species that have evolved from a mainland insect-eating finch have become vegetarian, whereas others have remained insectivores; some occupy ground-level habitats, others inhabit trees, and one lives on cactus; some eat seeds, others fruit. These are all ways by which the original stock extended into unutilized niches.

Occasionally, niche displacement is reflected in a structural or behavioral change in the species involved, a phenomenon called character displacement. The change in character, or feature, occurs in the region of overlap but not in the regions of exclusive occupancy. A good example is provided by the same Galápagos finches. Two species living on separate islands have approximately the same bill length; when they occur together on other islands, they exhibit altered bill dimensions, which reflect changes in food resources brought about by competitive interactions. A similar displacement of characters is seen in two related ant species—*Lasius flavus* and *L. nearcticus*—that coexist in the eastern United States and differ in at least eight features. These differences are not apparent in *L. flavus* in areas where it occurs alone in the western United States.

Antibiosis. Antibiosis occurs when one species interferes with or injures another through the secretion of a chemical substance. This commonly occurs among the bacteria and fungi and to a lesser extent among higher plants and some animals. The observance of such an effect in certain *Penicillium* molds is well-known. A bactericidal substance secreted into the growth medium prevented the growth of competing micro-organisms and led to the discovery of the antibiotic penicillin. A more complicated example concerns the grass *Aristida oligantha*, which invades old-field communities. It secretes phenolic acids, which inhibit the development of nitrogen-fixing bacteria and the blue-green algae of the soil. This, in turn, suppresses the production of available nitrogen in the soil, thus slowing the invasion of other competing, nitrate-requiring plants into the grass community.

Plant resistance

Antibiosis is the basis of the resistance exhibited by certain varieties of crop plants to insect attack. Corn varieties resistant to the European corn borer, for example, contain benzoxazolinones, chemicals that adversely affect larval growth and survival.

The discovery that many plants and some animals contain or secrete chemicals injurious to competitors or natural enemies has led to development of the study of allelopathy; the chemicals are called allelochemics. This phenomenon—the suppression of some higher plants by chemicals released by another higher plant—has been extended to include chemical defenses of plants against herbivores, phytophagous insects against predators, and the resistance of hosts to parasitoids.

Some of the more interesting examples of plant-to-plant antibiosis based on allelochemics include the chaparral plants, whose toxic phenolic secretions are washed by rains into the soil, where they inhibit the germination and growth of herb seeds close enough to provide competition. The milkweed-feeding caterpillar of the monarch butterfly acquires toxic chemicals from its host plant, thereby rendering the adult butterfly distasteful to bird predators.

Antibiosis can take forms other than chemical defense. They include immune responses, whereby hosts defend themselves against invading parasites. Immune responses of a host to an invading organism can involve engulfment (phagocytosis), chemical inactivation (precipitation), dissolution (lysis), or encystment.

Mutual antagonism. Mutual antagonisms occur when a biotic interaction between two species results in harm or death to both. Most mutual antagonisms occur as a result of competition for a limited resource. In some cases two pathogenic species together invade a host and thereby bring about its death, but in the process they destroy themselves; either pathogen alone would normally not destroy the host. An insect parasitoid and a predator occasionally engage in competitive interactions that result in death to both competitors, a phenomenon called synnecrosis.

When more than one insect parasitoid parasitizes a given host, competition for the host results. For example, when the female wasp parasitoids *Praon exsoletum* and *Trioxys complanatus* both attack the same aphid host, usually only one parasitoid survives, but occasionally the competition is so disruptive that the host dies as well, and with it all competitors. A variation of this habit is superparasitism, or the depositing by one species of more eggs into a host than can survive. The usual, adaptive outcome of such a situation is the death of all but one, an example of intraspecific competition. When hosts are in relatively short supply, however, superparasitism to an excessive degree commonly takes place. The result often is the death of both the parasitic larvae and the host itself.

Superparasitism

It has been said that the true parasite rarely exploits its host to the point of death, for the death of the host means the death of the parasite. The "prudent" parasite has evolved to avoid such drastic interaction, and the host likewise has evolved a resistance (immunity) to a parasitic species of long-standing association. When the pathogen, however, invades a new host species or subspecies or a previously unparasitized, susceptible host population, the resulting association may be one of extreme virulence, with rapid onset of death of the host.

Another example of mutual antagonism that is clearly harmful to both interacting species is the case in which cattle consume *Halogeton* plants, whose leaves contain a poison that frequently kills the cattle.

COMMENSALISM

In many communities there are species that benefit through interaction with other species, while the latter remain indifferent or unaffected. Such unilateral benefits include attainment of nutrients, space or support, shelter or protection, and transport. The commensal population is, by virtue of the association, increased in numbers through improved survival. If the commensal is obligatorily associated with a host, it cannot live in a community that does not contain the host in question.

Nutritional commensalism. Nutritional commensals, or allotrophs, are well illustrated by the many organisms found in the alimentary tracts of higher animals. Such endosymbiotes use host waste products for food and acquire suitable water supplies and a suitable microclimate in which to live. No harm is done to the hosts. Endosymbiotic bacteria and yeasts are found in most large mammals, particularly cattle, horses, sheep, deer, and buffalo.

Endosymbiotes

Numerous bacterial and yeast colonies grow on the bark, twigs, and leaves of trees and large shrubs, particularly in tropical habitats. These nonparasitic organisms live on the nutritious exudates and the decomposing bits and flakes of dead bark that coat the surfaces of such plants. In the soil many fungi and bacteria live on nutrients derived from the root exudates of higher plants.

Aquatic organisms with poor dispersal powers or poor food-gathering capabilities often attach themselves to species on whom they can depend to provide those qualities. The small crab *Lissocarcinus* lives commensally on the surface of sea cucumbers, being protected by colour camouflage from natural enemies and gaining its food simply by diverting bits of the plankton that is pulled into the sea cucumber's mouth by feeding currents. Clown fish (*Amphiprion*), which habitually swim among the tentacles of sea anemones, are unaffected by the stinging cells (nematocysts) of the host, but they receive protection from small predators and also partake of surplus bits of any food captured by the anemone. The well-known remora and the pilot fish also obtain their food from the activities of such hosts as a shark, marlin, or swordfish. The remora attaches itself to its host by a sucker organ, whereas the pilot fish swims in close association with the host. Both commensals feed on the leftovers of their host's meals.

Numerous birds are food commensals of other animals. Cattle egrets in Africa follow herds of elephant, buffalo, or antelope, feeding on insects and grubs turned up by the grazing activities of such animals. The cowbird (*Molothrus*) behaves similarly with cattle in the Americas. In northern Europe, ptarmigan travel with caribou in order to get the insects uncovered from the semifrozen sod. Gulls, lapwing plovers, and herons follow the farmer's plow to attain soil organisms turned up to the surface. Gulls and albatrosses follow ships at sea, feeding on garbage tossed overboard or on small fish stirred up by the passing vessel. Vultures follow carnivores such as lions, leopards, jackals, and hyenas and scavenge upon what is left of a killed carcass.

Scavenging as a form of commensalism

Insects such as biting lice, fleas, and louse flies, commonly categorized as ectoparasites, are more correctly designated ectosymbiotic commensals; they feed on feathers, sloughed-off flakes of skin, or waxy epidermal exudates. Only in the few cases of blood sucking are they regarded as true parasites.

Some insects and spiders closely resemble ants and associate with army ants in their foraging columns. These mimics feed on the booty flushed up by the foraging ant column. Certain ant mimics live within the nests of the host and as such are termed symphiles. Such symphiles include some of the rove beetles, claviger beetles, and certain mirid bugs.

Physical commensalism. The need for living space is a common basis for commensalism. In overcrowded communities, offshore marine habitats, or tropical rain forests, many nonmotile (sessile) micro-organisms are unable to find adequate sites on which to become established; hence, many of them attach themselves to the surfaces of other plants or animals. Most sessile or slow-moving marine animals—sponges, corals, bivalves, snails, turtles, and whales—carry ectocommensals such as algae, hydroids, and barnacles about with them. Woody plants physically support numerous orchids, ferns, bromeliads, grasses, and mosses of the tropics in tropical habitats. Trees in the temperate zone provide support for mosses and lichens.

Insects and mites live in bird's nests or rodent burrows, feeding on dead organic litter. Nest inquilines—insects living in the nests of bees, wasps, termites, or ants—subsist on excess food stores, nest materials, or dead hosts. Some inquilines apparently find the internal physical habitat of these social insect nests favourable for their own well-being.

Batesian mimicry, the morphological and colour-pattern resemblance of certain animals to other animals, is a form of protective commensalism. Edible butterflies, through natural selection by predatory birds or toads, come to resemble, or mimic, distasteful butterflies, which are avoided by these same predators. Drone flies gain protection from predatory toads by mimicking bees that sting and hence are avoided. In such situations the mimicking species benefits by avoidance of predation, whereas the model is little affected, neither being harmed nor benefitted by the mimicry (see also MIMICRY).

Protective commensalism

Examples of protective commensalism include the bird species that build their nests near those of aggressive species, such as predatory birds or certain venomous or predacious insects. In Europe, house sparrows and starlings often locate their nests at the sides of an eagle's nest. In North America, house sparrows and grackles often nest in close contact to nests of ospreys. In both

cases the "protector" species are interested in other prey: the eagles usually attack large rodents or larger birds; the osprey preys on fishes. The cordon bleu weaver, of Africa, nests directly above colonies of the predatory, fiercely stinging papernest wasps, and the Asiatic woodpecker *Micropternus* often builds nests inside the larger nest complexes of pugnacious ants of the genus *Crematogaster*. These weavers and woodpeckers rarely feed on their insect protectors, and the latter appear never to attack the birds in their nests.

MUTUALISM

Facultative mutualism: protocooperation. Many plant and animal species interact facultatively in ways that are general and indirect but beneficial to both species. These relations, called protocooperation, can be considered the first evolutionary step toward mutualism.

The interactions that occur between soil bacteria and fungi and between them and higher plants growing in the soil are protocooperative. No species is dependent on such an association, but collectively all microflora and higher plants, with associated soil fauna, participate in determining soil composition, structure, and fertility. In the zone of plant roots, soil bacteria and fungi interact with each other, some producing nutrients (metabolites) required by others and all obtaining nutrients ultimately from root exudates and decaying organic matter. Plants benefit from the actions of this microflora by acquiring needed mineral nutrients and carbon dioxide.

Plants also interact protocooperatively with grazing herbivores. Although this relationship constitutes herbivory, grazing herbivores such as deer or cattle maintain a characteristic plant association in their habitual grazing habitats. Removal of the grazing herds soon exposes the range or pasture to invasion by aggressive pioneer plants, including woody shrubs, brambles, and trees. If the grazers are returned, the carrying capacity of the range, much reduced at first, is gradually restored as the grasses and browse plants return. The reason for this interaction is that grasses and browse plants grow from their bases rather than from their tips, so grazers rarely injure the growing crowns of such plants. Shrubs and seedling trees, on the other hand, grow from the tips of their stems and hence are usually destroyed by grazing.

Maintenance of pasturage by grazing animals

The natural control of herbivore populations by natural enemies provides an indirect, protocooperative benefit to plants. These natural enemies (predators, parasites, and pathogens), whose direct influence on herbivorous and other animals has already been described, are some of the major factors that keep herbivore numbers in check. At a community level, this interrelatedness of trophic levels, plant–herbivore–carnivore, constitutes a large-scale protocooperative interaction.

Some of the relationships exhibited between ants and aphids are examples of protocooperation. Benefits are provided to both ant and aphid, but the relationship is often quite loose and facultative (a few such ant–aphid associations are obligatory and are treated in the next section). Generally, the ant member forages for food on trees and shrubs infested with such honeydew-secreting species as aphids, mealybugs, and some scales, collecting the sugary material and transporting it to its nest as food for developing young. In some cases the ant actually stimulates the aphid to secrete honeydew directly into its mouth. Some ant species even protect the honeydew producers from natural enemies, the consequences of which are noticeable: ant-attended trees usually bear much heavier infestations of aphids.

Plants whose flowers are pollinated by insects and birds benefit protocooperatively, particularly when the pollinator is a general one, and the plant is attended by many different pollinator species. Many plants, particularly those with colourful, showy flowers bearing nectar glands, are the beneficiaries of cross-pollination accomplished for them by insects. The insect, of course, benefits from the supply of food in the form of pollen and nectar. Honeybee colonies are used commercially to ensure pollination of many agricultural crops.

An interesting and apparently widespread form of protocooperation, called cleaning symbiosis, appears most noticeably in birds and fish. The Egyptian plover picks insect pests from the backs of buffalo, antelope, giraffes, and rhinoceroses and even leeches from the open mouths of crocodiles. The cattle egret in America performs the same function. In the offshore marine habitat, certain fishes function habitually as cleaners of other fishes, nibbling away at ectoparasites, wounded tissues, and dead flesh. Even predatory fish search out such cleaning symbiotes and remain passive while they are worked over. Such fish cleaners are often concentrated in fixed sites, called cleaning stations, where other fish come to be cleaned.

Cleaning symbiosis

Müllerian mimicry, the close similarity in appearance of two or more unrelated species in which each species is more or less distasteful to predators, is a form of protocooperation. Such mimicry presumably benefits all participating species (see MIMICRY).

Obligative mutualism: interdependency. In many communities the most stable, persistent, and interdependent associations between species are those based on obligative mutualism. Since the association is mandatory for the survival of each participant, the coexistence of both in a given habitat is always required. Because of this, ways have evolved to assure the perpetuation of the association from one generation to the next. Some of these means are intricate, involving adaptations in structure, behaviour, or life history so as to guarantee the coexistence of the partners. On the other hand, some mutualistic relationships are maintained simply through the high probability of mutual encounters as a result of the population density or great reproductive powers of one or both mutualists, or symbiotes.

How nature assures continuation of interdependencies

A number of bacteria–protozoan associations occur in aquatic environments. In certain cases, endosymbiotic bacteria, which exist in the cytoplasm of protozoan flagellates, are able to digest cellulose in quantities to provide for themselves and their hosts. Associations of fungi and single-celled algae include the well-known lichens, in which the fungal member penetrates algal cells with feeding tubes or haustoria, thus effecting an intimate physical interconnection. The mutual benefits derived from this association are nutrient exchange, maintenance of water and mineral balance, and resistance to drying or to extreme temperatures. Whether the association is truly mutualistic is difficult to resolve on formal grounds; some authorities contend that the fungi parasitize the algae. Ecologically, however, there is no question that the lichen is far better able to cope with its environment and to invade many more habitats than can either partner living alone.

Single-celled algae also enter into symbioses other than lichen formation, particularly in marine habitats where photosynthesis is restricted. Algae are symbiotic with, among other groups, protozoans, sponges, coelenterates, rotifers, flatworms, mollusks, echinoderms, and tunicates. The algal cell provides oxygen and manufactured food to its partner and, in turn, receives physical support, water, minerals, and a proper environment. Some algae are acquired by ingestion (the more primitive mechanism), whereas others are inherited; *i.e.*, transmitted during cell division of the host (the more advanced mechanism).

The association of algae with a motile coral polyp is accompanied by remarkable behavioral alterations in the host, which reinforce the benefits derived. Free-living larvae (planulae) of certain corals that contain yellow-green algae move toward light, whereas planulae lacking such algae are unresponsive to light. Corals endowed with algae grow more rapidly, take on different shapes, and are much denser than corals without algal members. In Caribbean mangrove swamps, certain algae-containing anemones distribute themselves in zones of intermediate light intensity; bleached anemones become indifferent to light or shade.

The important association of nitrifying bacteria with the roots of certain mostly leguminous plants is well-known. Nitrogen fixation—the extraction of nitrogen from the environment—occurs only after the bacteria

The unusual case of nodule bacteria

have invaded the plant roots and stimulated the host to form nodules that encapsulate the bacteria (a response of the host to infection). This association is closely related to parasitism. Nevertheless, the mutual benefits to the interactants are substantial: an environment and nutrients for the bacteria and improved fertility of the soil (by addition of nitrogen compounds) for the plant host.

Many insect species (perhaps most) contain microbial endosymbiotes, including bacteria, rickettsiae, fungi, yeasts, and protozoans. These organisms provide required nutrients for the host—often vitamins, occasionally digestive enzymes, and sometimes simple foodstuffs such as glucose sugar. The host provides the symbiote with a protected microhabitat containing food, water, minerals, and the proper chemical environment. The endosymbiotes are rarely found in the free-living state, and their hosts are unable to survive in their absence. They may be either extracellular—residing in the mouth, gut, rectum, blood spaces, or excretory tubules—or intracellular—living in the cytoplasm of various cells of the host. Because of the intimacy of such symbioses and the dependent nature of the associations, intricate mechanisms have evolved for the transmission of the endosymbiote from one generation of the host to the next. In some triatomid and lygaeid bugs, intestinal symbiotes are deposited as fecal material with the egg, and newly hatched larvae regain the symbiotes by consuming this material. In some pentatomid bugs the symbiotes are smeared on the oviposited egg shell, reinfecting the new hatchling as it emerges from the shell. In the olive fly (*Dacus oleae*) the egg is smeared with bacteria as it passes down the ovipositor, the bacteria penetrating the egg and becoming enclosed in the developing embryo. Finally, there are cases where the symbiotes penetrate the female ovary to invade the egg before it is even deposited; this occurs in some roaches, weevils, and ants.

One of the most notable mutualistic associations is that between certain wood-eating insects and cellulose-digesting protozoans. The wood roach (*Cryptocercus punctulatus*) acquires the intestinal cellulose-digesting symbiotes at an early age, retaining them for life. The wood-eating termite loses its symbiotes at each molt (shedding of skin) during its immature stages. The termites, however, reinfect themselves: newly molted nymphs ingest anal secretions from nonmolting nymphs and thereby obtain the intestinal organisms.

Insects that cultivate fungi

Numerous insects have become adapted to the use of fungi as specific food resources. These include the wood-boring insects that introduce and maintain fungal growths in their nest galleries and termites and ants that cultivate fungal "gardens" in their subterranean colonies. Four groups of wood-boring insects possess this habit: ambrosia beetles, bark beetles, timber borers, and sawflies (wood wasps). The adult ambrosia and bark beetles introduce the fungus into the brood galleries they make when they invade a tree. The adult timber borers and sawflies lay their eggs on or beneath the surface of the bark of a tree, introducing the fungus along with the eggs. The fungi then grow in the galleries formed in the wood and provide food for the developing insect larvae.

Myrmicine ants and certain tropical termites also are fungal "gardeners." The fungus-raising termites apparently utilize the fungi to control the localized climate of the nest rather than to obtain food. In termite nests where fungal gardens thrive, both the relative humidity and temperature are maintained at relatively constant, high, favourable levels as a result of the metabolic activities of the fungi. Fungus-culturing ants, however, clearly acquire nutrients from their gardens. Leaf-cutting ants, including the parasol ant, clip off and carry to their nests pieces of fresh leaves on which the fungi grow. The gardens are established usually in enlarged underground cavities, well ventilated, and tended by a class of workers who remove undesirable fungi and bacteria. All members of the colony feed on strands of the cultured fungi.

Certain insects are obligatory mutualists of the plants they pollinate. The California desert yucca can be pollinated only by the pronuba moth; the moth, in turn, is wholly dependent on the yucca flower ovary as a place to deposit its egg and develop its young. The common Smyrna fig can fruit only after pollination by the wasp *Blastophaga psenes*. This is a much more complicated process, since three different kinds of figs are involved in proper sequence: one for maintaining an overwintering population of wasps, one for producing adult wasps during the growing season, and one (the edible fig) that requires pollination but is otherwise unavailable for wasp reproduction.

Insects also enter into protective relationships with plants. The ant *Pseudomyrmex ferruginea*, which lives on and derives its food from the bull-horn acacia, protects the acacia from intruding vines, competing plants, and herbivorous insects. The ant depends upon the acacia, and the acacia cannot survive without the ant. A concomitant development of this mutualism is that this particular acacia has lost the chemical defenses against insect defoliators that other acacias possess.

NEUTRALISTIC INTERACTIONS

The final category of interspecies association commonly encountered in communities is neutralism, the persistent appearance of two or more species together with neither benefit nor harm accruing to any. It is a commonplace that seems self-evident and without significance; thus, it is often ignored in discussions of symbiosis. Indeed, in some cases—perhaps most—this interaction is trivial or accidental, yet in many communities such interspecific associations are characteristic. Most often, neutralistic associations occur when a common resource for several species is highly localized within the community, as when numerous species aggregate about water holes or streams. Insectivorous birds and predatory insects and rodents are persistent followers along foraging columns of army or driver ants, not to prey on the ants but rather upon the organisms stirred up by the mass advance. Neutralistic associations are characteristic in the distribution of trees in forests. The common situation is mixed stands of different species rather than pure stands of single species.

Simple aggregations

POPULATION EFFECTS OF INTERACTION

At the population level the interaction between different species can have two different effects, particularly when the interaction influences the survival or reproductive success of either species. One effect is on the numbers of individuals in the interacting populations. The other is on the qualities, or properties, of the individuals in these populations. The first effect—the quantitative one—is the basis for the balance of nature, the state of rareness or commonness of different species. The second effect—the qualitative one—is an aspect of evolution. Biotic interactions, when influencing the survival of a species, tend to produce changes in such aspects as structure, function, behaviour, and colour of the affected species. Since inherited changes in one species can serve as the basis for change in the other, biotic interactions between two species often lead to mutually induced changes in both, or coevolution.

Ecological aspects. The effect of one population on the numbers of another can be either positive or negative. A positive effect occurs when the survival or reproductive success of individuals of one species, or both, is enhanced and leads to an increase in the numbers of one or the other interacting population. Hence, an insect that consumes pollen in seeking a meal will generally bring about the pollination of the flowers of the host plant concerned; the result is an increase in the reproductive success of the host plant. Algae in such hosts as coral polyps stimulate the growth rate of coral colonies, and the enhanced growth of the coral benefits the algae.

Negative effects on a population

A negative effect occurs when the survival or reproductive success of one or both interacting populations is reduced, usually leading to a reduction in numbers of the affected populations. Negative interactions are those resulting from herbivory, predation, or parasitism, in which the food, prey, or host species suffers and may therefore diminish in abundance, or from competition or

mutual antagonism, in which both interacting species may suffer, be reduced in numbers, or even be eliminated from the habitat altogether.

In cases in which the biotic interaction is obligatory for one or both participating species, both species must appear together in the same community for the interaction to persist and for one or both participants to survive. A host species that is dependent upon an endosymbiote to provide necessary dietary components will usually not survive in the absence of the symbiote, and vice versa. A predator in consuming prey thereby reduces the numbers of the prey population to a point at which the prey becomes scarce and more difficult for the predator to find. The predator population then comes into balance with the diminished prey population. Any later increase in the prey results in an increase in the predator, and so on.

Evolutionary aspects. Since all species gradually change in response to selective factors, biotic interactions can serve as the means by which one of the interacting species brings about evolutionary change in the other. Just as predators can alter the numbers of a prey, and vice versa, evolutionary change in one interacting population can result in the evolutionary response of the other. For example, if a prey, through natural selection, becomes swifter or more agile in escape, a predator may develop, also through natural selection, capabilities for better pursuit. Plants, under the selective pressure of herbivory, often evolve a capability to produce toxic poisons that discourage herbivores. Subsequently, however, the herbivore, especially if it is dependent solely on the particular plant group involved, tends to develop tolerance for or resistance to the poisons. These examples of coevolution illustrate the dynamics of the process, the constant influence of each participant upon the other.

Convergent evolution

Since two interacting species tend to influence each other's evolution, it is not surprising that evolutionary convergence occurs in dissimilar organisms involved in similar biotic interactions. Convergence is the acquisition, through natural selection, of similar structure, function, or behaviour by unlike and totally unrelated species. One interesting example of convergence is a consequence of insect herbivory. A variety of unrelated insect species attack members of the Cruciferae (Brassicaceae), the plant family that includes mustard, cabbage, brussels sprout, radish. All crucifers contain chemical irritants or poisons that affect many animals and many micro-organisms. Such toxins very likely were evolved as chemical defenses in response to herbivore pressure. The insect species that now attack crucifers, however, have not only become capable of detoxifying the poisons but in some cases are even able to utilize them, either as attractants or as feeding stimulants.

Through coevolution, interspecific interactions among mutualists have led to mutual adaptations of unusual complexity. The mutualistic associations between the yucca plant and the pronuba moth, the Smyrna fig and the fig wasp, and the bull-horn acacia and the acacia ant have already been described. An even more striking example concerns the mimicry developed in the orchid, *Ophrys insectifera*, commonly called the fly orchid or bee orchid. The flowers of this unusual orchid—pollinated chiefly by a species of bumblebee—simulate very closely the shape and colour of a female bumblebee, so much so that male bumblebees frequently try to mate with it and in that attempt pollinate the flower.

Polymorphism is the occurrence in the same species of two or more different forms, distinguishable by colour, pattern, function, body dimensions, shape of the appendages, or behaviour. Often, polymorphism in one species occurs as a consequence of biotic interactions with another species. Character displacement in overlapping competitor species may present a situation in which the taxonomist often wonders whether he is not dealing with three or four different species rather than just two. Similarly, the discovery of noninterbreeding species, identical in every morphological detail but coming from different regions, also poses problems for the taxonomist. Such so-called sibling species often can be recognized only through differences in the nature of their biotic interactions. An example is the two parasitoid species *Trioxys complanatus* and *T. pallidus*. They are virtually identical in structure, yet each attacks and is specific to altogether different aphid hosts.

The taxonomic utility of biotic interactions

Biogeographic aspects. The geographic distribution of interdependent species also presents some interesting problems, particularly when the interrelationship is obligatory. When a herbivore is limited to a particular food plant (a not infrequent occurrence), the herbivore obviously cannot extend geographically farther than the range of the host plant. Parasites and predators are similarly limited by the distributions of their hosts or prey. On the other hand, neither the host plant, in the first case, nor the host or prey species, in the second, is so limited. On the contrary, some very interesting cases of population explosions or pest outbreaks have occurred when a species has, by invasion into a new region, escaped the natural restraints imposed by its normal predators or parasites.

Where the association of two species together is mandatory for the survival of each, as occurs in obligate mutualism, for example, the biogeographic problems are of a different sort. As has been shown, many obligate mutualists have, through coevolution, developed mechanisms for the transmission of the association from one generation to the next. Such transmission has reached the peak of development in the case of those insects that pass on their internal symbiotes via the ovarian egg. The symbiote never reaches the outside; it is passed on as faithfully as are the genes of the parents to the offspring.

A most interesting example of the value of knowledge of biotic interaction and geographic distribution is the case of the introduction of the common fig into California in the mid-19th century. Seedling figs were imported, planted, and nourished to maturity but failed to set fruit, even though they flowered profusely. Only when it was realized that the small fig wasp (*Blastophaga*) was required to insure fruiting in the Mediterranean countries from which the imported fig stock came was the problem solved. Importation and colonization of the fig wasp restored the obligatory mutualism between the fig and the fig wasp, and the plants were able to set fruit.

BIBLIOGRAPHY. W.C. ALLEE *et al.*, *Principles of Animal Ecology* (1949), a comprehensive reference work, with detailed discussion of both positive and negative biotic interactions among animals; P. BUCHNER, *Tier und Pflanze in intrazellularer Symbiose* (1921; Eng. trans., *Endosymbiosis of Animals with Plant Microorganisms*, rev. ed., 1965), a detailed, authoritative treatment of symbiosis between insects and micro-organisms; T.C. CHENG, *Symbiosis: Organisms Living Together* (1970), a semipopular book, encompassing types and origins of symbiotic associations, including human parasites and their life cycles; P. DEBACH, *Biological Control of Insect Pests and Weeds* (1964), a technical treatise with detailed accounts of the biotic interactions on which biological control is based; and "The Competitive Displacement and Coexistence Principles," *A. Rev. Ent.*, 11:183–212 (1966); C.S. ELTON, *The Ecology of Invasions by Animals and Plants* (1958), a popular and well-written account by a noted ecologist; P.L. ERRINGTON, "Predation and Vertebrate Populations," *Q. Rev. Biol.*, 21:144–177, 221–245 (1946); G.F. GAUSE, *The Struggle for Existence* (1934), a classic work that suggested many of the explanations now proved regarding biotic interactions; N.G. HAIRSTON, F.E. SMITH, and L.B. SLOBODKIN, "Community Structure, Population Control, and Competition," *Am. Nat.*, 94:421–425 (1960); G. HARDIN, "The Competitive Exclusion Principle," *Science*, 131:1292–1297 (1960); S.M. HENRY, *Symbiosis*, 2 vol. (1966–67), a survey of the symbiotic associations of micro-organisms, plants, and marine organisms (vol. 1), and a consideration of the symbiotic associations of invertebrates, birds, ruminants, and other organisms (vol. 2); C.B. HUFFAKER, "Experimental Studies on Predation: Dispersion Factors and Predator-Prey Oscillations," *Hilgardia*, 27:343–383 (1958); L.B. KEITH, *Wildlife's Ten-Year Cycle*, (1963), generalizations regarding the oscillations of prey and predators in nature; L. MARGULIS, "Symbiosis and Evolution," *Scient. Am.*, 225:48–57 (1971); R.H. WHITTAKER and P.P. FEENY, "Allelochemics: Chemical Interactions Between Species," *Science*, 171:757–770 (1971).

(P.S.M.)

Bird

Birds are numerous, conspicuous, and widespread. Their world, like man's, is primarily one of sight and sound, and they resemble man in possessing colour vision, a rather poor sense of smell, and a limited auditory range. Like man, birds care for their young and most are diurnal. Thus their ways are easily appreciated and studied by man. Their ability to fly stimulated the minds of poets and inventors for centuries before man's first successful flights.

Man has long used birds and their eggs for food, falcons for hunting, pigeons for carrying messages, the guano of seabirds for fertilizing crops, and the feathers of many species for ornaments. Some birds are destructive to man's crops and small livestock, and others act as carriers for various diseases that affect man. Bird-watching is a hobby of increasingly large numbers of people, and amateur naturalists have made many important contributions to the science of ornithology.

GENERAL FEATURES

The smallest living bird is generally acknowledged to be the bee hummingbird of Cuba, which is 6.3 centimetres (2½ inches) long and weighs less than three grams (about one-tenth of an ounce). The largest living bird is the ostrich, which may stand 2.5 metres (eight feet) tall and weigh 135 kilograms (300 pounds). Some extinct birds were even larger: the largest of the moas of New Zealand and the elephant birds of Madagascar may have reached ten feet in height. Among flying birds, the wandering albatross has the greatest wingspan, up to 3.5 metres (11½ feet), and the trumpeter swan perhaps the greatest weight, 17 kilograms (38 pounds). A Pleistocene condor-like bird, *Teratornis incredibilis*, had an estimated wingspan of about five metres (16½ feet) and was by far the largest known flying bird.

The ability to fly has permitted an almost unlimited radiation of birds, so that they are now found virtually everywhere on earth, from occasional stragglers over the polar ice caps to complex communities in tropical forests. In general the number of species found breeding in a given area is directly proportional to the size of the area and the diversity of habitats available. The total number of species is also related to such factors as the position of the area with respect to migration routes and wintering grounds of species that nest outside the area. In the United States, Texas and California have both the largest number of species recorded (545 and 461, respectively, including both resident and migrant species) and breeding (300 and 286). Seven hundred and seventy-five species, 650 of them breeding, have been recorded from North America north of Mexico. The figures for Europe exclusive of the U.S.S.R. are 577 and 420, and the figures for the U.S.S.R. are 704 and 622. Costa Rica, with an area of only about 20,000 square miles and a known avifauna of at least 758 species, probably has the most diversified for its size of any country.

IMPORTANCE TO MAN

Wild birds and their eggs have been at least incidental sources of food for man since his origin and still are in most societies. The eggs of some colonial seabirds, such as gulls, terns, and murres, or guillemots, and the large young of some shearwaters (muttonbirds) are even now harvested in large quantities. As man changed from hunting to agrarian cultures, several species of birds became domesticated. Of these, chickens, ducks, geese, and pigeons, descended from the red jungle fowl (*Gallus gallus*), the mallard duck (*Anas platyrhynchos*), the greylag goose (*Anser anser*), and the rock dove (*Columba livia*), respectively, were taken in early and have been selectively bred into many varieties. After the discovery of the New World, the turkey (*Meleagris gallopavo*), which had already been domesticated by natives, and the Muscovy duck (*Cairina moschata*) were brought to Europe and have since produced several varieties. Guinea fowl (*Numida meleagris*) were procured from Africa and kept not only for food but because they are noisy when alarmed, thus warning other fowl, as well as their owners, of the approach of intruders. In addition to being a food source, pigeons have long been bred and trained for carrying messages, and the ability of frigate birds to "home" to their nesting colonies has enabled natives of the South Seas to send messages by these birds.

Domestication

With the development of modern culture, hunting evolved from a foraging activity to a sport, in which the food value of the game became secondary. Large sums are now spent annually on hunting waterfowl, quail, grouse, pheasants, doves, and other game birds. Sets of rules and conventions have been set up for hunting, and in one elaborate form of hunting, falconry, there is not only a large body of specialized information on keeping and training falcons but also a complex terminology, much of it centuries old.

Feathers have been used for decoration since early times. Their use in the headdresses of American Indians and New Guinea natives, as well as of more civilized peoples, are well known. Feather robes were made by Polynesians and Eskimos, and down quilts, mattresses, and pillows are part of the European culture. Large feathers have often been used in fans, thereby providing an example of an object put to opposite uses—for cooling as well as for conserving heat. Whereas most feathers used in decorating are now saved as by-products of poultry raising or hunting, until early in this century, egrets, grebes, and other birds were widely shot for their plumes alone. Ostrich farms have been established to produce plumes. The early use of large quills for writing led to the development of pens, and feathers have long been used on arrows and fishing lures.

Many birds are kept as pets. Small finches and parrots are especially popular and easy to keep. Of these, the canary (*Serinus canaria*) and the budgerigar of Australia (*Melopsittacus undulatus*, sometimes called parakeet) are widely kept and have been bred for a variety of colour types. On large parks and estates, ornamental species like peafowl (*Pavo*) and various exotic waterfowl and pheasants are often kept. Zoological parks in many cities import birds from many lands and are a source of recreation for millions of people each year.

Birds as pests

With the rise of agriculture, man's relationship with birds became more complex. In regions where grain and fruit are grown, depredations by birds may be a serious problem. In North America various species of blackbirds (*Icteridae*) are serious pests in grainfields; while in Africa a grain-eating finch, the red-billed quelea (*Quelea quelea*), occurs, like locusts, in plague proportions so numerous that alighting flocks may break the branches of trees. The use of city buildings for roosts by large flocks of starlings and blackbirds is also a problem, as is the nesting of albatrosses on airplane runways on Pacific islands. As a result of these problems, conferences on the control of avian pests are held with increasing frequency.

Although birds are subject to a great range of diseases and parasites, few of these are known to be capable of infecting man. Notable exceptions are ornithosis (or psittacosis), caused by one or more viruses that are transmitted directly to man from pigeons, parrots, and a variety of other birds, a serious and sometimes fatal disease resembling virus pneumonia. Encephalitis, an inflammation of the brain, is also serious and is transmitted from birds to man and to his domestic animals by biting arthropods, including mosquitos. Wild birds may also act as reservoirs for diseases that adversely affect domesticated birds. Much work has been done recently on the ecology of viruses, with more and more of them being found in birds.

The study of birds has contributed much to both the theoretical and practical aspects of biology. Darwin's studies of the Galápagos finches and other birds during the voyage of the "Beagle" were important in his formulation of the idea of the origin of species through natural selection. Study collections of birds in research museums still provide the bases for important studies of geographic variation, speciation, and zoogeography, because birds

are one of the best known of animal groups. Early work on the domestic fowl added to the development of both genetics and embryology. The study of animal behaviour (ethology) has been based to a large extent on studies of birds by Konrad Lorenz, Nikolaas Tinbergen, and their successors. Birds also have been the primary group in the study of migration and orientation and the effect of hormones on behaviour and physiology.

Birds feature prominently in mythology and the literature of many countries. Some of their attributes, real or imagined, have led to their symbolic use in art as in language. The esthetic and recreational pleasures of bird-watching are increasingly being recognized.

Man's impact on bird populations has become increasingly strong. Since 1680, approximately 80 species of birds have become extinct and an even larger number are seriously endangered. While pollution and pesticides are important factors in the decline of certain large species, such as the peregrine falcon, osprey, and brown pelican, the destruction of natural areas and introduction of exotic animals and diseases have probably been the most devastating. Concerted efforts are required to insure the survival of rare species and to learn as much as possible about them.

NATURAL HISTORY

Behaviour. Birds depend to a great extent on innate behaviour, responding automatically to specific visual or auditory stimuli. Even much of their feeding and reproductive behaviour is stereotyped. Feather care is vital to keep the wings and tail in condition for flying and the rest of the feathers in place where they can act as insulation. Consequently preening, oiling, shaking, and stretching movements are well developed and regularly used. Some movements, like the simultaneous stretching of one wing, one leg, and half the tail (all on the same side) are widespread if not universal among birds. Stretching both wings upward, either folded or spread, is another common movement, as is a shaking of the whole body beginning at the posterior end. Other movements have evolved in connection with bathing, either in water or in dust. Such comfort movements have frequently become ritualized as components of displays.

Many birds maintain a minimum distance between themselves and their neighbours, as can be seen in the spacing of a flock of swallows perched on a wire. In the breeding season most species maintain territories, defended areas ranging from the immediate vicinity of the nest to extensive areas in which a pair not only nests but also forages. The frequency of actual fighting is in birds greatly reduced by ritualized threat and appeasement displays. Birds range from solitary (*e.g.*, many birds of prey) to highly gregarious, like the guanay cormorants of the Peru Current off the west coast of South America, which nest in enormous colonies of hundreds of thousands and feed in large flocks with boobies and pelicans.

Vocalizations

Auditory signals, like visual ones, are almost universal among birds. The most familiar vocalization of birds is that usually referred to as "song." It is a conspicuous sound (not necessarily musical) that is used, especially early in the breeding season, to attract a mate, to warn off another bird of the same sex, or both. As such it is usually associated with establishing and maintaining territories. Individual variation in songs of many species is well known, and it is believed that some birds can recognize their mates and neighbours by this variation. Many other types of vocalizations are also known. Pairs or flocks may be kept together by series of soft location notes. Alarm notes alert other individuals to the presence of danger; in fact, the American robin (and probably many other species) uses one note when it sees a hawk overhead and another when it sees a predator on the ground. Begging calls are important in stimulating parents to feed their young. Other calls are associated with aggressive situations, courtship, and mating. Nonvocal sounds are not uncommon. Some snipe and hummingbirds have narrow tail feathers that produce loud sounds when the birds are in flight, as do the narrowed outer primaries of the American woodcock. The elaborate courtship displays of grouse include vocalizations as well as stamping of the feet and noises made with the wings. Bill clapping is a common part of courtship in storks, and bill snapping is a common threat of owls.

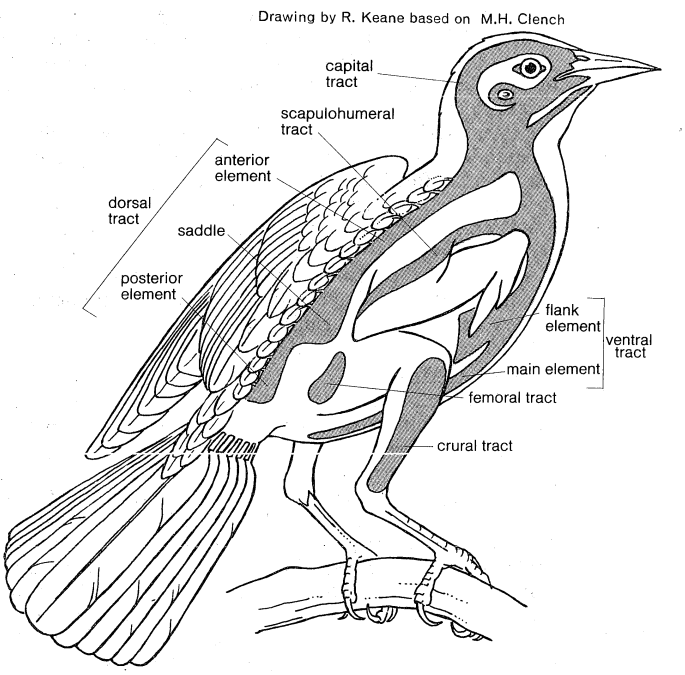


Figure 1: Basic body feather tracts on a generalized songbird. Shaded areas show the right half of each tract.

The nest

Most birds build nests in which the eggs are laid. Nests vary widely: they may be a scrape in the sand, a deep burrow, a hole in a tree or rock, an open cup, a globular or retort-shaped mass with a side entrance tube, or an elaborately woven hanging structure. The materials with which nests are made also vary widely. Some nests are lined with small stones, others are built of dirt or mud with or without plant material. Sticks, leaves, algae, rootlets, and other plant fibres are used alone or in combination. Some birds seek out animal materials such as feathers, horsehair, or snakeskin. The nest materials may be held together by weaving, sewing, or felting the materials themselves or with mud or spider webs. Swifts use saliva to glue nest materials together and to attach the nest to the supporting structure. In at least one species of swift, the entire nest is made of saliva and is the prized ingredient of birds' nest soup in the Orient. All birds incubate their eggs, except megapodes (mound builders), which depend on the heat generated by decaying vegetation or other external sources, and brood parasites, which lay their eggs in the nests of other species. Murres and the king and emperor penguins build no nest but incubate with the egg resting on top of the feet. In most birds a brood patch is developed. This bare area on the abdomen is edematous (fluid filled) and highly vascularized and is in direct contact with the eggs during incubation. Its development during the breeding season is under hormonal control. When the parent is off the nest, adjacent feathers are directed over the brood patch, and it is usually not apparent. A few birds (*e.g.*, boobies) keep their webbed feet over the eggs during incubation.

Incubation takes from 11 to 80 days, depending at least in part on the size of the bird and the degree of development at hatching. Most songbirds and members of some other groups are hatched nearly naked and helpless (altricial) and are brooded until well able to regulate their body temperature. They are fed by the parents until after they are capable of flight. The young of numerous other birds, such as chickens, ducks, and shorebirds, are hatched with a heavy coat of down and are capable of foraging for themselves almost immediately (precocial). Still others, such as the petrels and the auks, are downy

when hatched but remain in the nest and are fed by their parents.

The length of time parents care for young birds varies widely. Young megapodes can fly shortly after hatching and are entirely independent of their parents; young royal albatrosses may spend up to 243 days at the nest and in the area immediately around it before they can fly. The length of time needed to attain independence is related to size and condition at hatching. Ground-nesting birds tend to take less and hole-nesting birds more time than the average.

The number of eggs in a set varies from 1 to about 20. Some species invariably lay the same number per clutch (determinate laying), whereas in the majority the number is variable (indeterminate laying). In species of the latter category, clutch size tends to be smaller in tropical regions than in cold ones. There is also a tendency for birds in warm regions to make more nesting attempts in a given season. In the Arctic, where the season is very short, the cycle of breeding and the molt that follows it are telescoped into a minimum of time.

Feeding habits. The earliest birds were probably insectivorous, as are many modern ones, but the latter have evolved many specializations for catching insects: swifts, swallows, and nightjars have wide gapes for catching insects on the wing; some woodpeckers can reach wood-boring grubs while others can catch ants by probing anthills with their long, sticky tongues; thrashers dig in the ground with their bills; tree creepers and woodhewers probe bark crevices; and warblers glean insects from many kinds of vegetation. Raptorial birds have evolved talons and hooked bills for feeding on larger animals, and vultures have bare heads and tearing bills for feeding on carrion. Herons have spearlike bills and trigger mechanisms in the neck for catching fish, while kingfishers, terns, and boobies plunge into the water after similar prey. Long-billed waders probe for worms and other invertebrates. Of the many kinds of birds that feed on plant material, most use seeds, fruit, or nectar, which are high in food value; leaves and buds are eaten by fewer species. While some kinds of birds feed entirely on a single kind of food, others may take a wide range of foods, and many have seasonal changes.

FORM AND FUNCTION

Body proportions. Birds arose as warm-blooded, arboreal, flying animals with forelimbs adapted for flight and hindlimbs for perching. This basic plan has become so modified, through the course of evolution, that in some forms it is difficult to recognize. The maximum size attainable by flying birds is limited by the fact that wing area varies as the square of linear proportions, and weight or volume as the cube. On the other hand, the minimum size is probably governed by another aspect of the surface-volume ratio: the relative increase, with decreasing size, in surface through which heat can be lost. The largest flying birds have highly pneumatic skeletons (part of the bone is replaced by air spaces) and other adaptations for reducing weight; the small size of some hummingbirds may be facilitated by the decrease in heat loss resulting from their becoming torpid at night.

When birds lose the power of flight, the limit on their maximum size is lifted, as can be seen in the ostrich and other ratite birds. Some birds (auks, diving petrels, and certain ducks) use the wings for propulsion underwater as well as in the air. When birds that "fly" underwater lose the ability to fly in air, the wings become highly modified as paddles, as in the penguins.

Wing shape

The types of flight found in birds vary considerably. At least two major types of modifications for gliding or soaring are found. The albatrosses and some other seabirds have long, narrow wings and take advantage of winds over the oceans, whereas some vultures and hawks have broad wings with slotted tips and make more use of updrafts and winds deflected by hills. Short, broad wings are characteristic of chicken-like birds, which fly up with a rush of rapid wing beats. Birds like ducks, pigeons, and falcons, which fly rapidly with continuous wing beats, tend to have moderately long, pointed wings, while swifts and hummingbirds, with their narrow, curved wings fly rapidly and manoeuvre easily. The shape of a bird's tail also appears to be related to flight. Forms with deeply forked tails, such as frigate birds, terns, and some swallows, manoeuvre easily, whereas the opposite extreme, long, graduated tails, are often found in rapid, direct fliers like some parrots and doves. Woodpeckers and some other climbing birds have strong tail feathers with stout shafts, which they use as props while on the trunks of trees.

The bipedal gait, dictated by modification of the forelimbs for flight, necessitates manipulating food by the bill and feet and poses problems in balance. The relative lengths of the segments of the legs must be such that as the bird shifts from a standing to a sitting position, its centre of gravity remains over the feet. As some birds moved out of the trees and became terrestrial or aquatic, their legs were accordingly modified. The toes became shorter and the opposable first toe was lost in rapidly running forms like rheas and ostriches, and the toes became very long in birds that walk on aquatic vegetation or soft ground. In very large, slow-moving birds like moas, the leg bones became very heavy. Wading birds developed long legs, and climbing birds developed short legs with strongly curved, sharp claws. In swimming and diving birds, webs developed between the toes or lobes on the sides of the toes.

Feathers and molt. Feathers are unique to birds and characteristic of them. Like the scales of reptiles, and those on the feet of birds, feathers are made of keratin, a fibrous protein also found in hair. Feathers vary considerably in structure and function (Figure 2). Contour feathers form most of the surface of the bird, streamlining it for flight and often waterproofing it. The basal portion may be downy and thus act as insulation. The major contour feathers of the wing (remiges) and tail (rectrices) and their coverts function in flight. Contour feathers grow in tracts (pterylae) separated by bare areas (apteria) and develop from follicles in the skin.

The typical contour feather consists of a tapered central shaft, the rachis, with paired branches (barbs) on each side. An unbranched basal section of the rachis is called the calamus, part of which lies beneath the skin. The barbs, in turn, have branches, the barbules. The barbules on the distal side of each barb have hooks (hamuli) that engage the barbules of the next barb. The barbs at the base of the vane are often plumaceous; *i.e.*, lacking in hamuli and remaining free of each other. In many birds each contour feather on the body (but rarely on the wings) is provided with a complex branch, the aftershaft, or afterfeather, that arises at the base of the vane. The aftershaft has the appearance of a second, smaller feather, growing from the base of the first. Down feathers have loose-webbed barbs, all rising from the tip of a very short shaft. Their function is insulation, and they may be found in both pterylae and apteria in adult birds. They also constitute the first feather coat of most young birds. Filoplumes are hairlike feathers with a few soft barbs near the tip. They are associated with contour feathers and may be sensory or decorative in function. Bristlelike, vaneless feathers occur around the mouth, eyes, and nostrils of birds. They are especially conspicuous around the gape (corners of the mouth) of birds that catch insects in the air. Some bristles function as eyelashes on ground-dwelling birds, and the bristles over the nostrils may serve as filters.

The contour feathers are shed and replaced (molted) at least once a year, usually just after the breeding season. In addition, many birds have at least a partial molt before the breeding season. A typical series of molts and plumages would be juvenile plumage, postjuvenile (also called first prebasic) molt, first winter (or first basic) plumage, first prenuptial (or prealternate) molt, first nuptial (or alternate) plumage, first postnuptial (first annual, or second prebasic) molt, second winter (or basic) plumage, etc. Molt of the remiges and rectrices usually occurs as part of the annual molt and can be serial, from

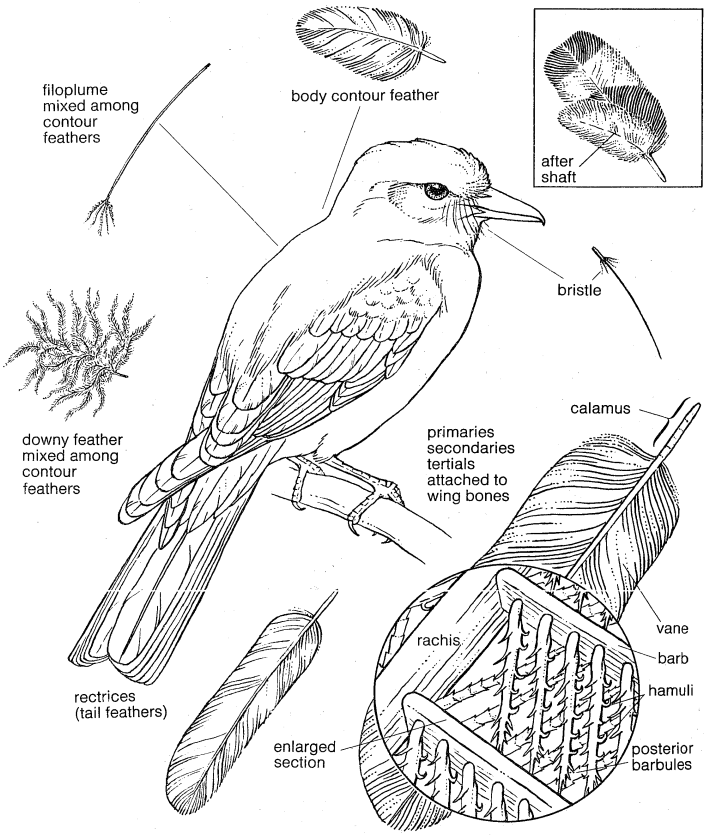


Figure 2: Feather types and their distribution on a typical perching bird.
Drawing by R. Keane

the innermost feather out (centrifugal), from the outermost in (centripetal), or simultaneous. Normally it is symmetrical between the right and left sides.

Coloration

Colour in birds is caused by pigments or structure. Buffs, red browns, dark browns, and blacks are caused by melanins, pigments synthesized by the bird and laid down in granules. Yellows, oranges, and reds come from carotenoid or lipochrome pigments; these originate at least in part from the food and are diffused in the skin and feathers. Porphyrin feather pigments occur in birds but less frequently than melanins and carotenoids. Blue colours in feathers are structural, based on a thin, porous layer of keratin overlying melanin pigment. Most greens result from the addition of yellow pigment to the structural blue colour. Iridescent colours result from thinly laminated structure of the barbules and are enhanced by underlying melanin deposits.

Birds' feet are covered with scales like those of reptiles. The scales are occasionally shed, but the timing of this molt is not known. The toes are tipped with claws, and vestigial claws are not infrequently found on the tips of the first two digits of the wing.

The bill is covered with a sheet of keratin, the rhamphotheca, which in petrels and a few other birds is divided into plates. In birds that probe for food (kiwis, woodcock, etc.), many sensory pores are found near the tip of the bill. Both melanins and carotenoids are found in the rhamphotheca and in the scales of the feet.

The skin of a bird is almost without glands. The important exception is the oil (uropygial) gland, which lies on the rump at the base of the tail. The secretion of this gland contains approximately one-half lipids (fats) and is probably important in dressing and waterproofing the plumage. In a few birds, the secretion has a strong, offensive odour. Some birds, in which the oil gland is small or absent, have a specialized type of feather (powder down) that grows continuously and breaks down into a fine powder, believed to be used in dressing the plumage.

Skeleton. The avian skeleton (Figure 3) is notable for its strength and lightness, achieved by fusion of elements and by pneumatization (*i.e.*, containing air spaces). The skull represents an advance over that of reptiles in the relatively larger cranium with fusion of elements, made possible by birds having a fixed adult size. Birds differ from mammals in being able to move the upper mandible, relative to the cranium. When the mouth is opened, both lower and upper jaws move: the former by a simple, hingelike articulation with the quadrate bone at the base of the jaw, the latter through flexibility provided by a hinge between the frontal and nasal bones. As the lower jaw moves downward, the quadrate rocks forward on its articulation with the cranium, transferring this motion through the bones of the palate and the bony bar below the eye to the maxilla, the main bone of the upper jaw.

The number of vertebrae varies from 39 to 63, with remarkable variation (11 to 25) within the cervical (neck) series. The principal type of vertebral articulation is heterocoelous (saddle shaped). The three to ten (usually five to eight) thoracic (chest) vertebrae each normally bear a pair of complete ribs consisting of a dorsal vertebral rib articulating with the vertebra and with the ventral sternal rib, which in turn articulates with the sternum (breastbone). Each vertebral rib bears a flat, backward-pointing spur, the uncinate process, characteristic of birds. The sternum, ribs, and their articulations form the structural basis for a bellows action, by which air is moved through the lungs. Posterior to the thoracic vertebrae is a series of ten to 23 fused vertebrae, the synsacrum, to which the pelvic girdle is fused. Posterior to the synsacrum is a series of free caudal (tail) vertebrae and finally the pygostyle, which consists of several fused caudal vertebrae and supports the tail feathers. The sternum consists of a plate lying ventral to the thoracic cavity and a median keel extending ventrally from it. The plate and keel form the major area of attachment for the flight muscles. The bones of the pectoral girdle consist of the furcula (wishbone) and the paired coracoids and scapulas (shoulder blades). The sword-shaped scapula articulates with the coracoid and humerus (the bone of the upper "arm") and lies just dorsal to the rib basket. The cor-

Drawing by R. Keane based on L. Darling and L. Darling, *Bird* (1962); Houghton Mifflin Company

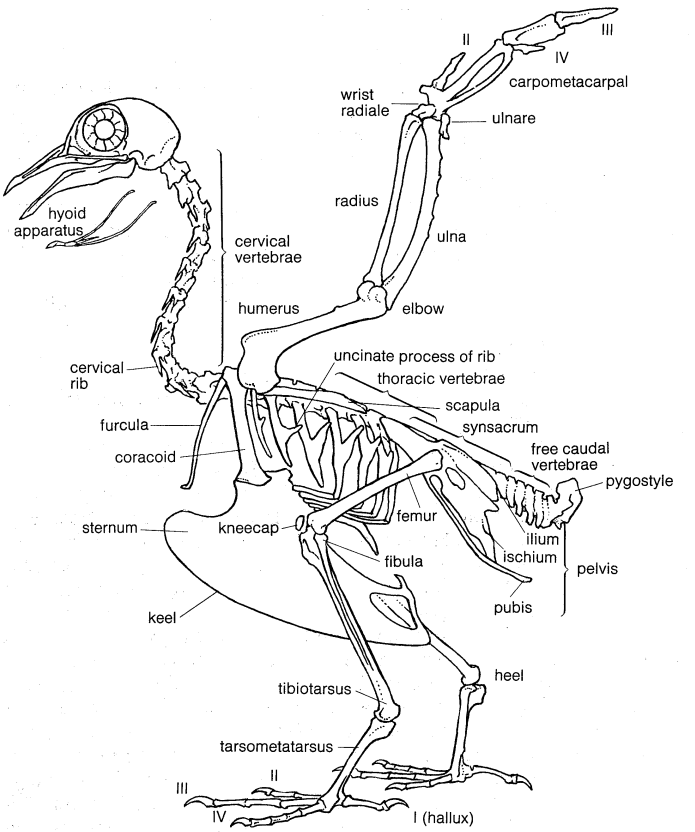


Figure 3: Pigeon skeleton, with the near wing raised and the far wing omitted.

acoid articulates with the anterior (forward) edge of the sternum and with the scapula, humerus, and furcula. The furcula connects the shoulder joints with the anterior edge of the keel of the sternum. It consists of paired clavicles (collarbones) and, probably, the median, unpaired interclavicle.

Wing skeleton

The bones of the forelimb are modified for flight with feathers. Major modifications include restricting the motion of the elbow and wrist joints to one plane, reduction of the number of digits, loss of functional claws, fusion of certain bones of the "hand" (the metacarpals and most of the carpals) into a carpometacarpus, and modification of the elements, especially those toward the tip of the limb (distal), for the attachment of feathers. The wing bones are hollow, and the cavity in the humerus, at least, is connected with the air-sac system. As a general rule, large flying birds have proportionally greater pneumaticity in the skeleton than small ones. The highly pneumatic bones of large flying birds are reinforced with bony struts at points of stress. The humerus, radius, and ulna are well developed. The secondary flight feathers are attached to the ulna, which thus directly transmits force from the flight muscles to these feathers and is therefore relatively heavier than the radius. Two small wrist bones are present: the radiale, or scapholunar, and the ulnare, or cuneiform. The former lies between the distal end of the radius and the proximal part (the part toward the body) of the carpometacarpus. When the elbow joint is flexed (bent), the radius slides forward on the ulna and pushes the radiale against the carpometacarpus, which in turn flexes the wrist. Thus the two joints operate simultaneously. The U-shaped ulnare articulates with the ulna and the carpometacarpus. Anatomists differ on which bones of the reptilian "hand" are represented in the bird's wing. Embryological evidence suggests that the digits are II, III, and IV, but it is possible that they are actually I, II, and III. The carpometacarpus consists of fused carpals (bones of the wrist) and metacarpals (bones of the palm), metacarpals II and III (or III and IV) contributing the greater part of the bone. The phalanges (bones of the "fingers") are reduced to one each on the outer and inner digits and two on the middle one. The primary flight feathers are attached to the carpometacarpus and digits, the number attached to each being characteristic of the various major groups of birds.

Drawing by R. Keane based on A. Bellairs and C.R. Jenkin, "The Skeleton of Birds," in J. Marshall (ed.), *Biology and Comparative Physiology of Birds*, vol. 1; Academic Press, Inc.

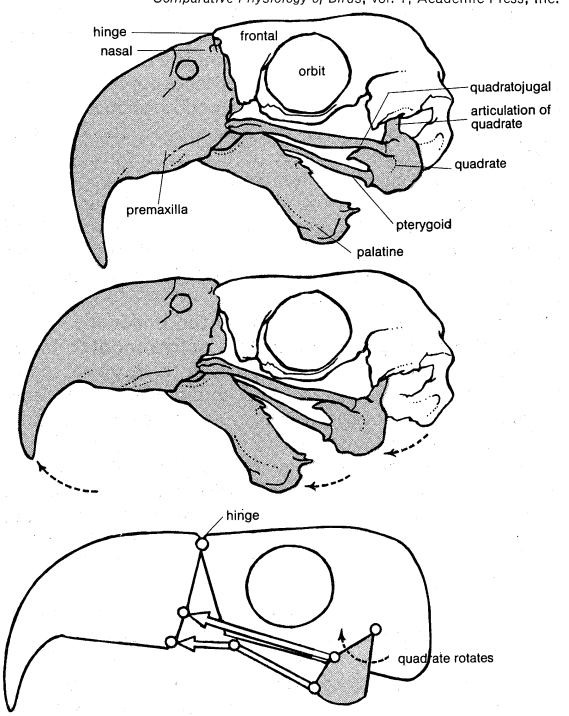


Figure 4: *Kinesis of the cranium of a macaw.* (Top) With upper mandible lowered. (Centre) With upper mandible raised. (Bottom) Showing forces acting on mandible.

Pelvis and leg

The pelvic girdle consists of three paired elements, the ilia, ischia, and pubes, which are fused into a single piece with the synsacrum. The ilium is the most dorsal element and the only one extending forward of the acetabulum (the socket of the leg). The ilium is fused with the synsacrum and the ischium, the latter fused with the pubis. All three serve as attachments for leg muscles and contribute to the acetabulum, which forms the articulation for the femur. The leg skeleton consists of the femur (thighbone), tibiotarsus (main bone of the lower leg), fibula, tarsometatarsus (fused bones of the ankle and middle foot), and phalanges (toes). The fibula is largest at its proximal (upper) end, where it forms part of the knee joint and tapers to a point distally, never forming part of the ankle joint. The latter joint is simplified, there being but two bones involved: the tibiotarsus, consisting of the tibia (the "shinbone" in man) fused with the three proximal tarsals (upper ankle bones), and the tarsometatarsus, resulting from the fusion of metatarsals I through IV and the distal row of tarsals. Metatarsals II through IV contribute most to the tarsometatarsus. The basic number of phalanges (sections) on the toes is two, three, four, and five, respectively; *i.e.*, one more than the number of the toe. Most birds have four toes, the fifth being always absent, but there are many variations in the number of digits, or phalanges, representing reductions of the basic arrangement.

Drawing by R. Keane based on L. Darling and L. Darling, *Bird* (1962); Houghton Mifflin Company

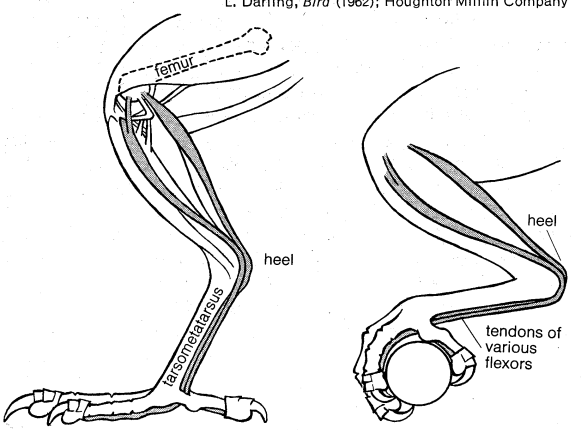


Figure 5: Perching mechanism of a pigeon with the leg extended and flexed.

The basic avian foot is adapted for perching. The first, or hind, toe (hallux) opposes the other three, and the tendons for the muscles that bend the toes pass behind the ankle joint in such a way that when the ankle is bent the toes are also. The weight of a crouched bird thus keeps the toes clasped around the perch.

Muscles

Internal organs. The cardiac (heart) muscles and smooth muscles of the viscera of birds resemble those of reptiles and mammals. The smooth muscles in the skin include a series of minute feather muscles, usually a pair running from a feather follicle to each of the four surrounding follicles. Some of these muscles act to raise the feathers, others to depress them. The striated (striped) muscles that move the limbs are concentrated on the girdles and the proximal parts of the limbs. Two pairs of large muscles move the wings in flight: the pectoralis, which lowers the wing, and the supracoracoideus, which raises it. The latter lies in the angle between the keel and the plate of the sternum and along the coracoid. It achieves a pulley-like action by means of a tendon that passes through the canal at the junction of the coracoid, furcula, and scapula and attaches to the dorsal side of the head of the humerus. The pectoralis lies over the supracoracoideus and attaches directly to the head of the humerus. In most birds the supracoracoideus is much smaller than the pectoralis, weighing as little as one-

twentieth as much; in the few groups that use a powered upstroke of the wings (penguins, auks, swifts, hummingbirds, and a few others), the supracoracoideus is relatively large. Avian striated muscles contain a respiratory pigment, myoglobin. There are relatively few myoglobin-containing cells in "white meat," whereas "dark meat" contains far more. The former type of muscle is used in short, rapid bursts of activity, whereas the latter is characteristic of muscles used continuously for long periods and especially in muscles used during diving.

The circulatory system of birds is advanced over that of reptiles in several ways: (1) there is a complete separation between the pulmonary (lung) and systemic (body) circulations, as in the mammals; (2) the left systemic arch (aortic artery) is lost, blood passing from the heart to the dorsal aorta via the right arch; (3) the postcaval vein is directly connected with the renal portal that connects the kidneys with the liver; and (4) the portal circulation through the kidneys is greatly reduced. Birds' hearts are large—0.2 to over 2.4 percent of body weight, as opposed to 0.24 to 0.79 percent in most mammals.

The avian lung differs from the type found in other land vertebrates, in containing fine tubes (capillaries) through which air passes and through the walls of which gas exchange takes place. Several pairs of nonvascular air sacs are connected with the lungs and extend into the pneumatic parts of the skeleton. The sound-producing organ in birds is the syrinx, located where the trachea (windpipe) divides into the bronchial tubes. The sounds are made by the flow of air setting up vibrations in membranes formed from part of the trachea, bronchi, or both. Muscles between the sternum and trachea or along the trachea and bronchi vary tension on the membranes.

The avian digestive system shows adaptations for a high metabolic rate and flight. Enlargements (collectively called the crop) of parts of the esophagus permit the temporary storage of food. The stomach is typically divided into a glandular proventriculus and a muscular gizzard, the latter lying near the centre of gravity of the bird and compensating for the lack of teeth and the generally weak jaw musculature. Otherwise, the digestive system does not vary markedly from the general vertebrate type.

Like reptiles, birds possess a cloaca, a chamber that receives digestive and metabolic wastes and reproductive products. A dorsal outpocketing of the cloaca, the bursa of Fabricius, controls antibody-mediated immunity in young birds. The bursa regresses with age, and thus its presence or absence may be used to determine age.

The testes of the male bird are internal, like those of reptiles. Intromittent organs are found in only a few groups (waterfowl, cracids, tinamous, ratites). The distal part of the vas deferens (the seminal sac) becomes enlarged and convoluted in the breeding season and takes on both secretory and storage functions. In passerine birds, at least, this enlargement and the adjacent part of the cloaca form a cloacal protuberance, a swelling visible on the outside of the bird. Usually only the left ovary and oviduct are functional. The albumen, membranes, and shell are laid down in the oviduct as the egg moves down it. The gonads and accessory sexual organs of both sexes enlarge and regress seasonally. In the breeding season, the testes of finches may increase over 300-fold in volume over their winter size.

Temperature control

Birds are homeothermic (warm-blooded) and maintain a body temperature of approximately 41° C (106° F). It may be 1–1.7° C less during sleep and up to 2° C higher at times of great activity. Feathers, including down, provide effective insulation. In addition, layers of subcutaneous fat add further insulation in penguins and some other water birds. Reduction of heat loss from the feet in cold weather is accomplished by reducing blood flow to the feet and by a heat-exchange network in the blood vessels of the upper leg, so that the temperature of blood flowing into the unfeathered part of the leg is very low.

Drawing by R. Keane based on L. Darling and L. Darling, *Bird* (1962); Houghton Mifflin Company

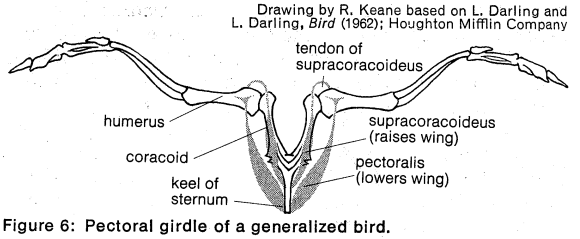


Figure 6: Pectoral girdle of a generalized bird.

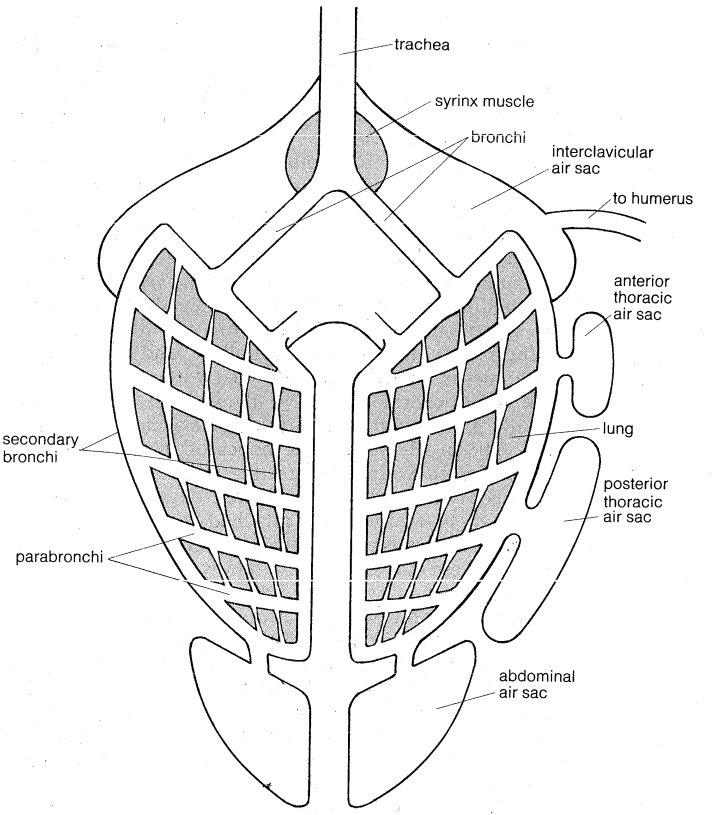


Figure 7: Avian lung and air sac system in a generalized bird.
Drawing by R. Keane based on A.L. Thomson (ed.), *A New Dictionary of Birds;* British Ornithologists' Union and J. Marshall (ed.), *Biology and Comparative Physiology of Birds* (1961); Academic Press

Birds differ from mammals in lacking sweat glands, hence heat loss is accomplished by rapid panting, which reaches 300 respirations per minute in domestic hens. Some heat dissipation can be accomplished by regulation of blood flow to the feet. In hot climates, overheating is often prevented or reduced by behavioural means, concentrating activities in the cooler parts of the day and seeking shade during the hot periods. Temporary hypothermia (lowered body temperature) and torpor are known for several species of nightjars, swifts, and hummingbirds. Torpor at night is believed to be widespread among hummingbirds. The heart rate of birds varies widely—from 60 to 70 in the ostrich to over 1,000 per minute in some hummingbirds.

The kidneys lie in depressions on the underside of the pelvis. The malpighian bodies (the active tubules of the kidney) are very small in comparison to those of mammals, ranging from 90 to 400 per cubic millimetre. Over 60 percent of the nitrogen is excreted as uric acid or its salts. There is some resorption of water from the urine in the cloaca, with uric acid remaining. There is no urinary bladder, the urine being voided with the feces. In marine birds, salt is excreted in a solution from glands lying above the eyes through ducts leading to the nasal cavity.

EVOLUTION AND PALEONTOLOGY

Archaeopteryx

The earliest known fossil bird is *Archaeopteryx lithographica* discovered in Upper Jurassic deposits in Bavaria. This bird was about the size of a magpie. It resembled some reptiles and differed from Recent birds in many ways: (1) the jaws contained teeth set in sockets; (2) the articulations between the vertebrae were amphicoelous (concave at both ends); (3) there were only six sacral vertebrae; (4) the long tail was made up of a series

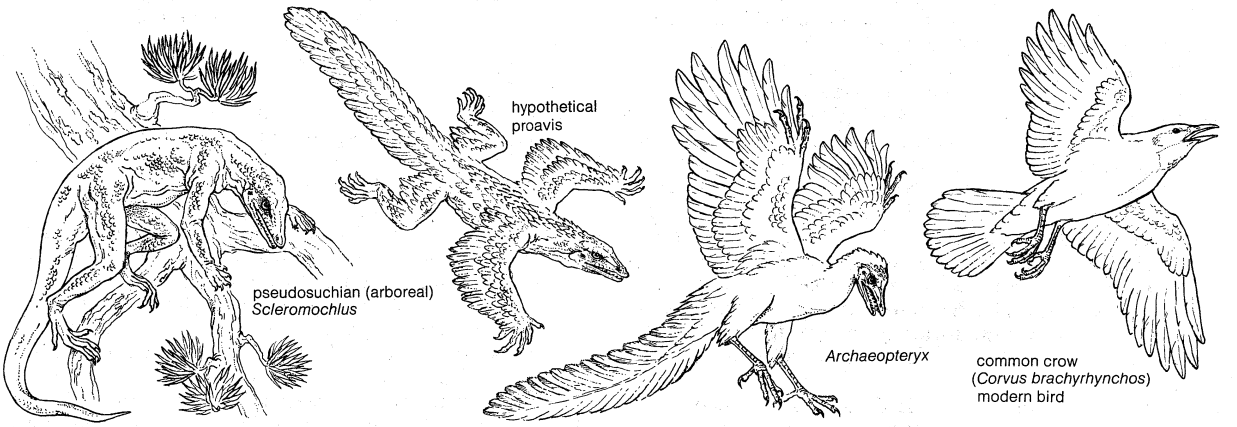


Figure 8: Four stages in the evolution of modern birds.
Drawing by R. Keane

of free vertebrae each bearing a pair of rectrices; (5) the slender ribs lacked articulations with the sternum and uncinate processes (flat upward projections); (6) ventral ribs (gastralia) were present behind (posterior to) the sternum; (7) the sternum was short and not keeled; (8) the bones were not pneumatic; (9) the third metacarpal bone in the wing was fused to the carpals, but the first two metacarpals were free, resulting in three movable digits of the "hand," all with functional claws; (10) the fibula was as long as the tibia; (11) the metatarsal bones were free; (12) the cerebral hemispheres were elongated and slender, and the cerebellum lay behind the midbrain, not overlapping it from behind or crowding it downward. Avian characters of *Archaeopteryx* included the possession of feathers, the elongated, backward-directed pubis, the furcula, and the opposable hallux. In the structure of the beak, eye, and jaw articulation, in the fusion of the third metacarpal with the carpals, and in the fusion of each of the distal tarsals with the corresponding metatarsal, *Archaeopteryx* was intermediate between reptiles and modern birds.

The absence of a keel on the short sternum indicates that *Archaeopteryx* did not fly but glided. The opposable hallux, indicative of the perching type of foot, and the clawed digits of the hand point to an arboreal existence. From the arrangement of feathers on the wing and the number and arrangement of bones in the limbs, it appears that *Archaeopteryx* was near the main line of avian evolution. From the fact that the skull was diapsid (*i.e.*, had two "windows") and from certain features of the limb bones, it appears that *Archaeopteryx* was descended from reptiles of the Triassic order Thecodontia.

The origins of flight

By the Triassic Period (225,000,000 years ago) a group of small bipedal reptiles, the pseudosuchians, were well established. Their skulls had much in common with that of *Archaeopteryx*, although they had heavier jaws and smaller eyes. It is likely that one group of pseudosuchians became arboreal. The advantages of such a life would be safety from large terrestrial predators and an abundance of insect food. Once these reptiles were in the trees, selective pressures would favour mutations leading to many avian features. The swaying of branches would favour the evolution of the grasping foot. Use of the forelimbs in climbing from branch to branch would favour enlargement of the claws and elongation of the forelimb, which was short in the bipedal ancestors. Greater visual acuity and more effective coordination are of special advantage in arboreal animals. Natural selection thus favoured larger eyes, narrower snouts (permitting better forward vision), and greater development of the cerebral lobes and cerebellum of the brain. The smaller jaws may also indicate the advantages of lightness, balance, and a specialization for feeding on insects as opposed to the apparently more general carnivorous diet of the terrestrial ancestors. Perhaps most important was the development of homeothermy (internal temperature control). A warm-blooded insect eater has an enormous advantage in being able to capture insects when they are cold and slow to react. It is also advantageous in the wind-moved environment of the treetops. In addition to increased food intake and advanced respiratory and circulatory systems, however, homeothermy requires effective insulation. It is likely that feathers evolved to fill this requirement, although many authorities believe the origin of feathers was directly connected with flight. Just how feathers evolved from reptilian scales is unknown, but it is known that the two are similar in chemical composition and that some pseudosuchians had scales bearing an imprint of a feather-like pattern on their surface. Elongated feathers on the forelimb and tail may have evolved for balancing and for gliding to produce the *Archaeopteryx* stage.

In the evolution of modern birds from an *Archaeopteryx*-like form, the development of active flight must have occurred early. This meant an increase in size of the muscles moving the wing and the development of a keel on the sternum as an added area of attachment for these muscles. As the tail took on more of a steering function and less of a supportive one, it became shorter and more readily moved as a unit. Feathers became increasingly specialized for different functions, and at the same time, the trends in the development of the eyes, brain, and respiratory and circulatory systems associated with the evolution of the homeothermic, arboreal, gliding types continued. By the time birds became strong fliers, they were ready to radiate out into many new environments; and by the Cretaceous Period (136,000,000 to 65,000,000 years ago), they had begun to do so. This radiation has produced the large array of adaptive types known today.

The lightness and pneumaticity of bird bones makes them poor candidates for fossilization. As might be expected, heavyboned diving birds and large flightless birds are disproportionally represented in the record.

Cretaceous birds

One of the best known groups of fossil birds consists of *Hesperornis* and its relatives. These birds were highly specialized foot-propelled divers of the Upper Cretaceous. The known species of *Hesperornis* were up to six feet long and had completely lost the power of flight. The sternum lacked a keel; the humerus was small and weak; and the other, more distal, elements of the wing were missing. The pelvis and hindlimb had a strong but superficial resemblance to those of modern loons and grebes—the pelvis was narrow; the femur short and stout with a hingelike articulation with the pelvis; the tibiotarsus long, with a long cnemial crest (a projection at its upper end); and the tarsometatarsus laterally compressed. Two major features (and several less obvious ones) indicate, however, that the resemblance was the result of convergent evolution: the ischia and pubes were free for most of their length, and the cnemial process was made up entirely of the patella; in the loon, this process is derived from the tibiotarsus. *Hesperornis* was remarkable for three features: it had teeth set in grooves, not sockets, in

the maxilla and mandible; the phalanges of the stout fourth toe had a unique rotary ball-and-flange type of articulation; and the free tail vertebrae had broad lateral projections and limited vertical motion, indicating that the tail was somewhat beaver-like in its action. *Baptornis*, a contemporary relative of *Hesperornis*, was smaller and less strongly modified. While flightless, it had less reduced wings than *Hesperornis*, and it lacked the peculiar modifications of the fourth toe and caudal vertebrae.

Living on the same seas as *Hesperornis* and *Baptornis* was a group of flying birds known as *Ichthyornis* and *Apatornis*. Although not related to gulls, these birds resembled them superficially and may well have been their ecological counterparts. It was long believed that *Ichthyornis* had teeth, like *Hesperornis;* but it is now thought that the toothed jaws formerly thought to belong to *Ichthyornis* were really those of a small mosasaur, a marine reptile.

Large flightless predators

After the extinction of the dinosaurs and before large carnivorous mammals evolved, two groups of large flightless birds evolved to fill a similar niche. From the upper Paleocene to the middle Eocene, *Diatryma* and its relatives were major predators in the Northern Hemisphere. The largest species stood over two metres (seven feet) tall and had stout hooked beaks. They are of uncertain relationships but may have been distantly related to cranes and rails. The second group, that of *Phororhacos* and related genera, had a long history (from the lower Oligocene to the middle Pliocene) in South America, which was without large carnivores until relatively late. Fragmentary Pleistocene material from Florida has also been assigned to this group. The *Phororhacos* line evidently evolved from cariama-like stock and radiated into numerous genera and species, the largest of them (*Onactornis*) standing 2½ metres (eight feet) tall and having a skull 80 centimetres (31 inches) long and 40 centimetres (16 inches) high.

Large grazing or browsing birds appear to have evolved several times. On continents where there are large predators, these birds have always been rapid runners (ostriches, rheas, emus), but on islands lacking such predators, they were slow-moving, heavy-bodied birds. Two such groups were the elephant birds of Madagascar and the moas of New Zealand, the largest in each group approaching ten feet in height. Fragmentary fossil material from Eocene and Oligocene deposits in Egypt indicates that similarly adapted birds occurred there before the advent of large carnivores.

Except for the *Hesperornis* line, teeth appear to have been lost very early in the history of birds, but fish-eating birds have evolved several toothlike structures for grasping their prey. Perhaps the most remarkable adaptation was that of *Osteodontornis* and its relatives, large, flying marine birds that flourished from the lower Eocene to the Miocene. In these birds, there were bony projections of the upper and lower jaws, which were covered by the ramphotheca, forming sharp, toothlike structures.

The fact that fewer bird fossils are found in earlier deposits is well illustrated by expressing the number of known species in a given geological period in terms of the duration of the period. About 35 species of birds are known from Cretaceous deposits, which were laid down over an estimated 71,000,000 years, giving a figure of 0.5 species per 1,000,000 years. The corresponding figure for the Paleocene is 1.2 (12 species in 10,000,000 years) and that for the Eocene, 5.4 (87 species in 16,000,000 years). Up-to-date figures for the later periods are not available, but estimates based on several recent sources are 7.9 per 1,000,000 years in the Oligocene, 12.3 in the Miocene, and 21.8 in the Pliocene. The Pleistocene, which lasted approximately 2,500,000 years has yielded nearly a thousand species of fossil birds. From this it is evident that very little is known about the early avifaunas. It is known, however, that, as might be expected, the birds in the earlier periods differ more from Recent species than do those of the later periods. Of the 12 families of birds recorded from Cretaceous deposits, only two are still extant, whereas the majority of species recorded from the Pleistocene were structurally little, if at all, different from living forms. Thus the absence of a group in the fossil record, especially in the earlier periods, is rarely significant.

The major diversification of birds probably took place in the Cretaceous, which lasted longer than the sum of all subsequent periods, and it must have started early in that period because fragmentary material of foot-propelled divers (*Enaliornis*) and of an early relative of the flamingos (*Gallornis*) are known from Lower Cretaceous deposits of Europe. Upper Cretaceous deposits have yielded, besides *Hesperornis* and *Ichthyornis* and their relatives, diving birds similar to *Enaliornis* (*Lonchodytes*), other early flamingo-like birds, and species in the same suborders as gannets, ibises, rails, and shorebirds.

Paleocene deposits have yielded the earliest known loons, cormorants, New World vultures, and gulls, while the large, flightless predatory birds culminating in *Diatryma* first made their appearance in this period. From the far richer Eocene deposits have come the earliest known rheas, penguins, albatrosses, tropic birds, anhingas, true flamingos, herons, storks, secretary birds, hawks, curassows, cranes, bustards, avocets, auks, sandgrouse, cuckoos, owls, swifts, trogons, rollers, hornbills, and songbirds. Almost certainly all living orders and most living families were in existence by the end of that period. One of the most interesting finds from this period was *Neocathartes*, a long-legged bird allied to the New World vultures. There are several anatomical similarities between this group of vultures and the storks, and the existence of this fossil lends support to the idea that the storks and New World vultures are more closely related to each other than each family is to the birds with which it is usually grouped.

Important Oligocene fossils include the earliest phororhacoids, one of the few groups of fossil birds that is known from enough material from over a long enough time span to show evolutionary trends, in this case, both in size and in bill form.

Fossils of Miocene birds are numerous. Several early groups of pelecaniform, cranelike, and flamingo-like birds are known last from this period, and the first of the Mancallidae, superficially penguin-like auks, appeared. Otherwise, the avifauna was essentially modern.

By Pliocene times, most modern genera were probably in existence. The Mancallidae continued on the California coast at least until the middle of the period.

The appearance and extinction of large birds as well as mammals was a feature of the Pleistocene. Perhaps most notable were the teratorns, "super condors," which were found in North America. These included *Teratornis incredibilis*, the largest known flying bird.

CLASSIFICATION

Distinguishing taxonomic features. In classifying birds, most systematists rely primarily on structural characters. Plumage characters include the number of remiges and rectrices; the presence or absence of down on the feather tracts, on apteria, and on the oil gland; and the presence or absence of an aftershaft. Characteristics of the bill and feet are useful, as is the arrangement of bones in the palate and around the nostrils. The presence or absence of certain thigh muscles and the arrangement of the carotid arteries, the syrinx, and the deep flexor tendons of the toes are employed, as is the condition of the young when hatched. Recently, advances in the study of protein structure and of chromosomes have added new evidence of taxonomic relationships.

Annotated classification. This classification is based primarily on that of the American ornithologist Alexander Wetmore but includes the ideas of a number of other authorities. It is unlikely that most avian systematists would agree on all aspects of one arrangement, but the one presented below will satisfy many. The dagger (†) indicates extinct groups, known only from fossils.

CLASS AVES

Vertebrate (backboned) animals primarily adapted for flight with feathers. Warm-blooded, 4-chambered heart, left syste-

mic arch lost. Lower jaw articulates with cranium via the quadrate; teeth absent in living forms. Reproduction by hard-shelled eggs, nearly always incubated by one or both parents. About 8,600 living species.

†Subclass Archaeornithes

†Order Archaeopterygiformes (*Archaeopteryx*). Upper Jurassic; Europe. Teeth set in sockets; long, unfused caudal vertebrae, each bearing a pair of rectrices; keelless sternum; functional claws on digits of hand. Gliding birds, about 50 cm long.

Subclass Neornithes

†Superorder Odontognathae

†Order Hesperornithiformes (*Hesperornis, Baptornis*). Upper Cretaceous; North and South America. Teeth set in groove in jaws. Flightless, foot-propelled diving birds, 1 to 2 m long.

Superorder Neognathae

Order Tinamiformes (tinamous). Upper Pliocene to present; Central and South America. Superficially quaillike or grouselike ground-dwelling birds with flat, elongated, and rather weak bills and very small tails. Size 15–50 cm. See TINAMIFORMES.

Order Rheiformes (rheas). Lower Eocene to present; South America. Ostrich-like cursorial birds with very small tails and no aftershaft on the feathers. Sexes alike. Length 90–130 cm.

Order Struthioniformes (ostrich). Lower Pliocene to present (the Eocene *Eleutherornis* may belong here); southwestern Asia and Africa (fossils from southern Europe and southeastern Asia). 2-toed (3rd and 4th) running birds. Males black and white, females brown. Aftershafts, down, and filoplumes absent. Largest living bird; length to 180 cm, height 260 cm, weight 136 kg, egg 1.6 kg.

Order Casuariiformes (emus, cassowaries). Pleistocene to present; Australia, New Guinea, adjacent islands. Very large, cursorial (running) birds. Sexes alike, brown (emus) or blackish with brightly coloured wattles and skin on head (cassowaries). Aftershaft very large. Length 130–190 cm. See CASUARIIFORMES.

†Order Aepyornithiformes (elephant birds). Pleistocene; Madagascar (upper Eocene and lower Oligocene fossils from Egypt have been placed here). Very large and graviportal (heavy bodied); height to 3 m; egg weight estimated at 10 kg.

Order Dinornithiformes (moas, kiwis). Upper Miocene or lower Pliocene to present; New Zealand. Very large (to 3 m tall) and graviportal birds (moas) or smaller (length 30–80 cm); almost wingless, nocturnal, probing birds (kiwis).

Order Podicipediformes (grebes). Lower Miocene to present; worldwide. Foot-propelled diving birds with lobed toes, minute tails, and silky plumage. Length 21–66 cm. See PODICIPEDIFORMES.

Order Procellariiformes (albatrosses, shearwaters, petrels). Middle Eocene to present; all oceans, most numerous in Southern Hemisphere. Web-footed marine birds with tubular nostrils; rhamphotheca divided into plates; musky smell. Most have narrow wings and stiff, gliding flight. Length 14–135 cm. See PROCELLARIIFORMES.

Order Sphenisciformes (penguins). Upper Eocene to present; oceans of Southern Hemisphere. Wings flipper-like, for propulsion underwater; webbed feet short and stout; stance upright. Feathers short and dense, molted in patches. Length 40–120 cm (fossil forms to 180 cm). See SPHENISCIFORMES.

Order Pelecaniformes (tropic birds, pelicans, boobies, cormorants, frigate birds). Paleocene to present; worldwide. Water birds with all 4 toes webbed; bill hooked or straight and sharply pointed. Length 50–180 cm. See PELECANIFORMES.

Order Anseriformes (screamers, waterfowl). Middle Eocene to present. Web-footed birds with broad bills containing fine plates or lamellae (waterfowl); or large-footed marsh birds with chicken-like bills (screamers). Length 29–160 cm. See ANSERIFORMES.

Order Phoenicopteriformes (flamingos). Cretaceous to present; discontinuously distributed in warm regions except Australasia. More varied and widely distributed as fossils. Web-footed birds with long legs, long necks, bent bills with lamellae, and much pink or red in the plumage. Share characters with both Anseriformes and Ciconiiformes, but evidently closer to the latter, with which they are sometimes grouped. Length 91–122 cm (some fossil forms smaller). See CICONIIFORMES.

Order Ciconiiformes (herons, storks, ibises, spoonbills). Upper Cretaceous to present; worldwide except in extreme north. Long-legged wading birds with long bills; feet not webbed. Although usually grouped together, herons and storks may prove to belong to different orders. Length 28–152 cm. See CICONIIFORMES.

Order Falconiformes (diurnal birds of prey). Upper Paleocene to present; worldwide. Diurnal raptors with hooked beaks, long talons, and short (hawks, falcons) or very long (secretary bird) legs or carrion-eating birds with weaker claws and tearing bills (vultures, condors). Length 15–150 cm (some fossil forms larger). See FALCONIFORMES.

Order Galliformes (grouse, pheasants, quail, turkeys). Middle Eocene to present; nearly worldwide, except southern South America. Terrestrial or arboreal chicken-like birds; strong, scratching feet; short, rounded wings; feathers with long aftershafts. Length 13–198 cm. See GALLIFORMES.

Order Gruiformes (cranes, rails, coots, cariamas, bustards). Upper Cretaceous to present; worldwide. Diverse group, ranging from small quail-like hemipodes to large long-legged cranes, marsh-inhabiting rails, swimming coots and fin-foots, and cursorial bustards. The Tertiary phororhacoids belong here, as may the very large *Diatryma* and its relatives. Length 11–152 cm (fossils to 200 cm tall). See GRUIFORMES.

†Order Ichthyornithiformes (*Ichthyornis, Apatornis*). Upper Cretaceous; North America. Superficially gull- or ternlike water birds of uncertain affinities. Length approximately 21 to 26 cm (estimated from reconstruction of fossils).

Order Charadriiformes (plovers, sandpipers, gulls, terns, auks). Upper Cretaceous to present; worldwide. 3 basic body plans: Suborder Charadrii—waders (shorebirds), usually feeding on small animals in mud or water, bill variable but often long and used for probing; Lari—web-footed, dense-plumaged water birds feeding by plunging into water for fish, robbing other birds, or scavenging; Alcae—dense-plumaged, web-footed, marine, wing-propelled divers, feeding on fish or invertebrates. Length 13–76 cm. See CHARADRIIFORMES.

Order Gaviiformes (loons or divers). Upper Paleocene to present; Holarctic. Foot-propelled diving birds with webbed feet and pointed bills. Cnemial crest an extension of the tibia. Length 66–95 cm.

Order Columbiformes (sandgrouse, pigeons, doves, dodos). Upper Eocene or lower Oligocene to present; worldwide except in extreme north. Fast-flying birds with pointed wings and weak bills; feed on seeds and fruit. Length 15–84 cm. See COLUMBIFORMES.

Order Psittaciformes (parrots, lories, cockatoos). Upper Oligocene to present; throughout tropics, with some temperate-zone species. Often brightly coloured. Strong-flying, seed-, fruit-, or nectar-eating birds with very stout, hooked bills and zygodactyl feet (*i.e.*, outer toe reversed). Length 9.5–99 cm. See PSITTACIFORMES.

Order Cuculiformes (touracos, cuckoos, roadrunners). Upper Eocene or lower Oligocene to present; worldwide except in extreme north. Long-tailed birds with zygodactyl or semizygodactyl feet. Feed on both fruits and small animals. Most arboreal, a few terrestrial. Some brood parasites. Length 16–70 cm. See CUCULIFORMES.

Order Strigiformes (owls). Eocene to present; worldwide. Nocturnal raptorial birds with hooked beaks, strong talons, and soft plumage. Length 13–69 cm. See STRIGIFORMES.

Order Caprimulgiformes (nightjars, frogmouths, oilbird). Pliocene to present; worldwide except in extreme north. Concealingly coloured, soft-plumaged, nocturnal birds with weak feet and very large mouths. Most feed on insects caught in flight. Length 19–53 cm. See CAPRIMULGIFORMES.

Order Apodiformes (swifts, hummingbirds). Upper Eocene or lower Oligocene to present; worldwide except in extreme north; hummingbirds limited to New World. Rapid-flying birds that feed on the wing on insects and nectar. "Hand" and primaries constitute a relatively great proportion of the wing; feet weak. Length 6.3 to 23 cm. See APODIFORMES.

Order Coliiformes (colies or mousebirds). Unknown as fossils. Africa south of the Sahara. Soft-plumaged birds with long, pointed tails and all 4 toes directed forward. Food largely vegetable, some insects. Length 29 to 36 cm.

Order Trogoniformes (trogons). Pantropical, except Australasia; upper Eocene or lower Oligocene to present. Extremely soft-plumaged arboreal birds; underparts yellow to red, head and neck often iridescent, tail long, black and white. Feet weak; 1st and 2nd toes directed backward. Food insects and small fruit. Length 23 to 34 cm.

Order Coraciiformes (kingfishers and allies). Eocene to present; worldwide except in extreme north. A heterogeneous

group of hole-nesting birds. Many with long, pointed bills and blue or green in plumage. All have 2nd and 3rd or 3rd and 4th toes joined at base. Food largely animal, except hornbills, which eat much fruit. Length 9 to 160 cm. See CORACIIFORMES.

Order Piciformes (woodpeckers, barbets, honeyguides, toucans). Upper Oligocene (possibly upper Eocene) to present. Zygodactyl (rarely 3-toed) hole-nesting birds. Food insects and fruit. Woodpeckers are modified for climbing. Honeyguides are brood parasites. Length 9 to 61 cm. See PICIFORMES.

Order Passeriformes (perching birds). Upper Eocene to present; worldwide. The large complex assemblage of perching birds, containing more than half of the known species of birds. Bill, plumage, and habits highly varied. Length 7.5 to 102 cm. See PASSERIFORMES.

Critical appraisal. It has frequently been stated that birds are one of the best known of animal groups. This is true, in the sense that most of the living species and subspecies in the world have probably been described; but because of inadequacies in the fossil record and repeated cases of convergent evolution within the group, our knowledge of the phylogenetic relationships between orders, suborders, and families of birds is inferior to that of mammals and reptiles.

The taxonomic positions of several bird groups remain open to question. The hoatzin, included above in the Galliformes, is often given its own order, Opisthocomiformes. The touracos, here included in the Cuculiformes, are considered by many authors to warrant separation as Musophagiformes. *Diatryma* and several related genera of extinct flightless predators are often placed in a distinct order, Diatrymiformes, near Gruiformes. The flamingos, which constitute the order Phoenicopteriformes above, are placed in the Ciconiiformes in many classifications. The sandgrouse, family Pteroclidae, are believed by some to be more closely related to the shorebirds (order Charadriiformes) than to the pigeons (order Columbiformes), with which they are usually grouped.

One area particularly in need of study is the relationships among the various groups of ratites (ostriches, rheas, emus, moas, etc). Formerly, some authorities argued that these birds and the penguins arose independently from cursorial reptiles, but it is now generally agreed that all of them passed through a flying stage in the course of their evolution. The ratite groups differ greatly in morphology and yet show remarkable similarities in palate and bill characters. The principal unanswered questions are how many different flightless lines evolved from flying ancestors and from how many different groups were the flying ancestors evolved. On zoogeographic grounds, it is likely that the isolated kiwi-moa, elephant bird, and emu-cassowary lines arose independently from each other and from ratites on the other continents. But the ostriches and rheas could be descended from a common flightless ancestor because of the known former land connections from Asia to North and South America.

Before organic evolution was understood and accepted, animals were grouped on the basis of general similarity. It is now known that many such groupings were unnatural from a phylogenetic standpoint but were instead the result of convergent evolution from different parental stocks. Examples are *Hesperornis*, loons, and grebes, and diving petrels and auks. It is likely that many more examples are not recognized or generally accepted. At least the following groups should be studied with this in mind: herons and storks, *Diatryma* and the phororhacoids, New World vultures and other falconiforms, sandgrouse and pigeons, touracos and cuckoos, and swifts and hummingbirds. These examples are all from the ordinal or subordinal level; examples at lower levels would be far more numerous.

BIBLIOGRAPHY. A.J. MARSHALL (ed.), *Biology and Comparative Physiology of Birds*, 2 vol. (1960–61), and D.S. FARNER and J.R. KING (eds.), *Avian Biology*, vol. 1–2 (1971–72; vol. 3–4, in prep.), two of the most thorough works on avian biology to date in English; A. LANDSBOROUGH THOMSON (ed.), *A New Dictionary of Birds* (1964), definitions and authoritative accounts of a wide variety of topics related to birds; R.T. PETERSON *et al.*, *The Birds* (1963), a well-illustrated, popular account of birds and their biology; E. STRESEMANN, "Aves," in the *Handbuch der Zoologie*, 8 vol. (1927–34), a classic general work (in German); J. VAN TYNE and A.J. BERGER, *Fundamentals of Ornithology* (1959), an advanced, college-level text, especially noteworthy for its accounts of avian families; J.C. WELTY, *The Life of Birds* (1962), a college-level text crammed with general information on birds; H.F. WITHERBY *et al.*, *The Handbook of British Birds*, 5 vol. (1938–41), a concise summary of knowledge of British birds; R.S. PALMER (ed.), *Handbook of North American Birds*, vol. 1 (1962), a compendium of information on North American loons, grebes, petrels, pelecaniforms, ciconiiforms, and flamingos, other volumes of which are planned for the remaining orders; P. BRODKORB, "Catalogue of Fossil Birds," pt. 1–4, *Bull. Florida State Mus.*, 7:179–293 (1963), 8:195–335 (1964), 11:99–220 (1967), 15:163–266 (1971), an up-to-date world list of fossil birds through the Piciformes; A. WETMORE, "A Check-List of the Fossil and Prehistoric Birds of North America and the West Indies," *Smithsonian Misc. Collns.*, vol. 131, no. 5 (1956), an account of North American fossil birds not included in the parts of Brodkorb's catalogue published up to the early 1970s; G.R. DE BEER, *Archaeopteryx lithographica* (1954), the most thorough account of this important fossil; O.C. MARSH, *Odontornithes* (1880), the best available figures and descriptions of *Hesperornis* and *Ichthyornis;* K.C. PARKES, "Speculations on the Origin of Feathers," *Living Bird*, 5:77–113 (1966), arguments favouring evolution of feathers for flight rather than thermoregulation; E. MAYR and D. AMADON, "A Classification of Recent Birds," *Amer. Mus. Novit.*, no. 1496 (1951), and A. WETMORE, "A Classification for the Birds of the World," *Smithsonian Misc. Collns.*, vol. 139, no. 11 (1960), two differing arrangements of the orders and families of birds.

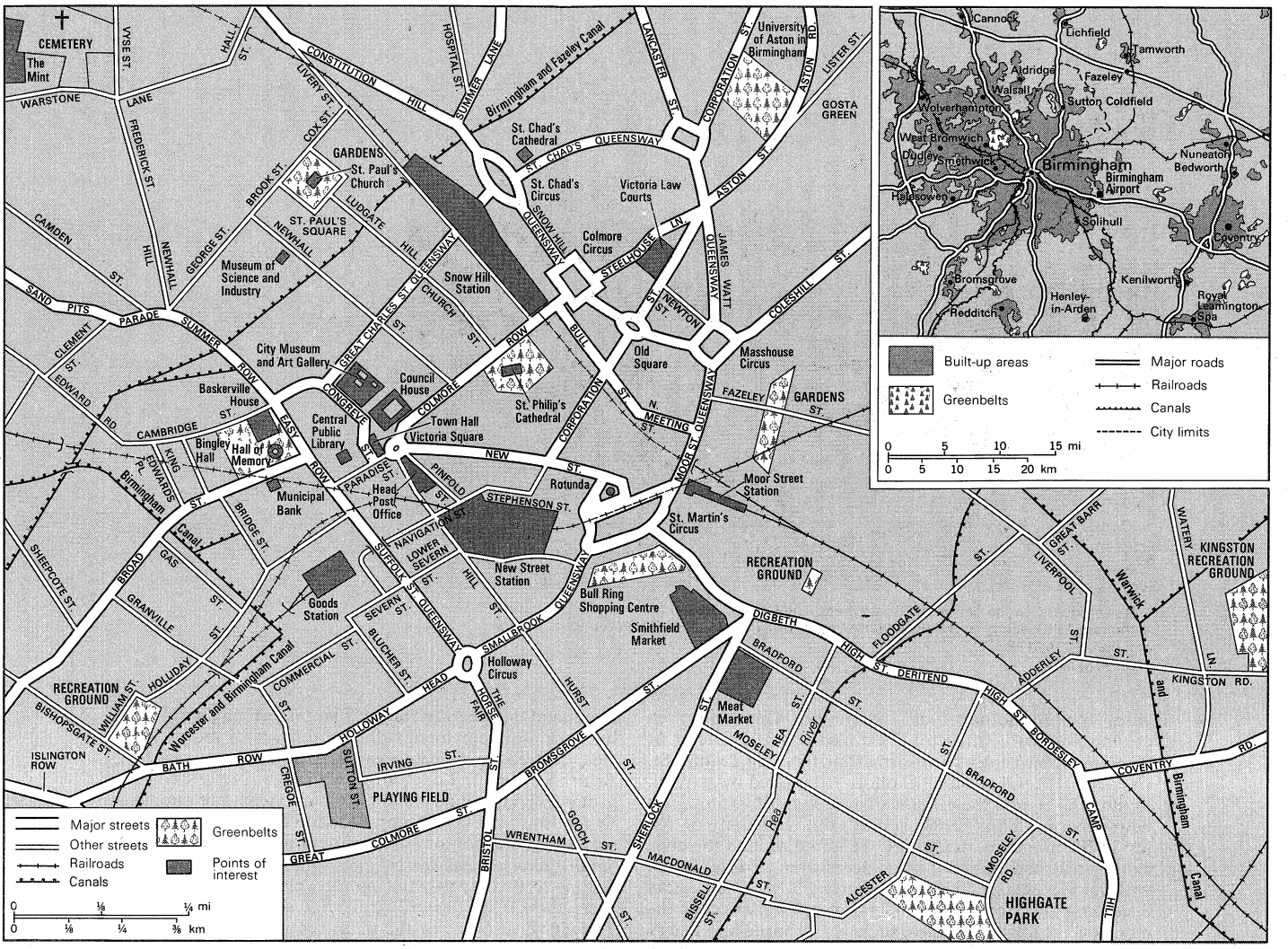
(R.W.St.)

Birmingham

A city and parliamentary borough (returning 12 members to Parliament) forming part of West Midlands metropolitan county, Birmingham is the second city in Great Britain in population (estimated in 1972 at 1,092,000) and one of the principal manufacturing and commercial regions of the United Kingdom. It lies near the geographical centre of England, at the crossing points of the national railway, motorway, and canal systems. With an area of 80.6 square miles (208.8 square kilometres) and a perimeter of 52 miles (84 kilometres), it is also the largest town in the West Midlands conurbation—itself over 20 miles (32 kilometres) across, with nearly 2,400,000 inhabitants—for which it acts as an administrative, recreational, and cultural centre.

In the late 18th century Birmingham became the leading nucleus of the Industrial Revolution in Britain, pioneering new methods of production in the metal and engineering trades. A group of men of genius lived in the city at that time, many of them members of the philosophical and scientific group known as the Lunar Society. They included the engineers James Watt (inventor of the steam engine), Matthew Boulton, and William Murdock (pioneers in steam engine development), the chemist Joseph Priestley, and the printer John Baskerville. The entire group deeply influenced the technological progress of Birmingham and the nation. Boulton's Soho Manufactory, which developed the steam engine for industrial use, became famous throughout Europe. Despite bad housing and the uncontrolled spread of industry, Birmingham became well-known in the later 19th century for its efficient and enlightened civic administration. During the 1950s and 1960s, a vast and often controversial redevelopment and road-building program transformed the Victorian city beyond recognition. Though dominated by industry, Birmingham has always been surrounded by pleasant suburbs, rich in trees and flowers, and has easy access to a beautiful countryside.

City growth. Birmingham's lack of river transport, by cutting it off from the maritime contacts important in the medieval period, impeded its development from a small manufacturing town to a large city until the late 18th century. The population of 15,000 at the end of the 17th century had become 70,000 a century later. The town



Central Birmingham and (inset) its metropolitan area.

then spread from its hilly centre, which, by 1800, contained St. Philip's Cathedral (1715) and many Georgian buildings in the St. Paul's Square area and elsewhere. The metal and gun trades expanded, fine jewelry was made alongside cheaper lines, and local buttons and trinkets, often in brass, served a world market. It was not until after the Reform Act of 1832 that Birmingham elected its own members to Parliament, and the city did not receive its charter of incorporation until 1838. In the same year, rail links to Liverpool and London were completed. In 1869 the city found the leader it needed when Joseph Chamberlain, a rich local industrialist, was elected to the council. In 1873 he was elected mayor, a position he held for three years. He initiated important reforms, among them sweeping slum and city-centre redevelopment schemes, incidental to which were the setting up of a school board and several major improvements to urban services.

Granting of a charter

Birmingham was a British pioneer in town-planning schemes (1911), one-way traffic experiments (1933), and municipal airports (1939). The Municipal Bank, still the only one in the country, opened in 1916. Wartime industrial activity and heavy bombing left the city temporarily exhausted in 1945, but it quickly recovered. Smokeless zones and the conversion of the railways to diesel oil and electricity eliminated much of the industrial haze and smog, while carefully preserved parks and open spaces kept the city and suburbs rich in greenery. A new inner ring road scheme, a rebuilt central station, and new shopping and commercial complexes began the city's transformation. The most striking visual features of this by the 1970s were the Rotunda, a circular office building 265-feet (81-metres) tall, the post office radio tower (500 feet [150 metres]), and the Bull Ring Shopping Centre, incorporating shops, a bus station, a car park, markets, restaurants, and gardens.

The small and cramped city centre has had to learn to build high, but at the same time much central industry and population have been moved out, first to five central redevelopment areas immediately adjoining the city centre, then farther out to 38 additional comprehensive redevelopment areas.

The contemporary landscape. Birmingham stands in the centre of an upland plateau shut in by the valleys of the rivers Trent, Severn, and Avon. The land rises from 267 feet (81 metres) above sea level in the east to 736 feet (224 metres) in the west. The northern suburbs adjoin Staffordshire, and the southern and southwestern suburbs border on Hereford and Worcester. On the northwest the city merges into the old industrial agglomeration of small towns known as the Black Country, a name referring to the effect of industrial pollutants on land, buildings, and the atmosphere. To the southeast it adjoins Solihull, a separate district and popular residential area, while the southern suburbs straggle out pleasantly into the Warwickshire countryside. A few medieval and Elizabethan buildings, often much altered, lie scattered through the inner and outer suburbs; a fine Jacobean house, Aston Hall (1618), is situated close to the city centre.

The Black Country

One central focus is Victoria Square with the classical Town Hall (1834) and the Renaissance-style Council House (1874–81) and City of Birmingham Museum and Art Gallery. St. Philip's Cathedral, in its green church-



Victoria Square, in the centre of Birmingham. The Council House (centre left) was built in the late 19th century in Italian Renaissance style. The statue of Queen Victoria (centre) was erected in 1899.
A. Devaney, Inc.

yard, forms another focus, while the Georgian area around St. Paul's Church (1779) also has a character of its own. Other centres are formed or forming around St. Chad's Cathedral (Roman Catholic), designed by A.W.N. Pugin (1841), and the elegant *Birmingham Post and Mail* newspaper building (1964), around the University of Aston in Birmingham, in Gosta Green, and around the civic centre in Broad Street. The adjacent area of Edgbaston, with its Georgian and early Victorian houses, is rapidly becoming an extension of the commercial districts of the city. Several former villages, such as Harborne and King's Norton, keep something of their rural aspect. More than 7,300 acres (3,000 hectares) of parks keep the district open and pleasant, though the inner circle of older former suburbs, among them Sparkbrook, Handsworth, and Saltley, are extremely dilapidated and overdue for redevelopment.

Population. The main demographic problem is the large immigrant population—white and nonwhite. The nonwhite ("New Commonwealth") population at the 1971 census in the former county borough numbered 68,000, of whom 25,000 were from the Caribbean, 38,000 from Asia (mostly India and Pakistan), and the rest from the continent of Africa and from other countries. There was an Irish population of 45,000, about 40,000 persons being Irish-born. The poorer immigrants settled in the older suburbs near the city centre. Despite some overcrowding, however, there had been no serious racial or immigrant trouble in Birmingham by the mid-1970s. The city as early as 1954 appointed an officer whose sole concern was community relations. Birmingham lost some population after the 1950s to residential districts beyond what was then the city boundary, to new towns, and to "overspill" areas. The five redevelopment areas close to the city centre, which will have their own shops, schools, doctors, and industries, and the tower, or high-rise, apartments in the city centre itself, are being constructed as part of a scheme to attract residents back to the centre. During the last years of the 1960s the former Birmingham Corporation built more than 7,250 new dwellings, and its total for the latter half of the decade was some 33,000, a record for municipal housing in Britain and possibly in the world.

The immigrants

The economy. Once known as "the workshop of the world" and still often described as "the city of 1,500 different trades," Birmingham is the centre of Britain's light and medium industry, principally concerned with the metal and engineering trades. The largest single industry in terms of employment is the production of motor vehicles, centred on the automated Austin Morris plant of British Leyland at Longbridge, with other factories of the group in and around the city. Components are supplied by firms as large as Dunlop Ltd. and the electrical accessories organization of Joseph Lucas and by hundreds of small factories and workshops, some employing only a few persons. Bicycles and motorcycles are made in the area, though the bicycle trade has somewhat declined. The city is also one of the main centres of the machine-tool industry. Local industry also includes heavy engineering, electroplating, and manufacture of aircraft components, fine and costume jewelry, silverware, plastics, sporting guns, buttons, and ornamental badges. Chemicals, furniture, and toys have their place in the city's economy, and Bournville, an early example of industrial town planning, is the site of Cadbury's chocolate factory.

The automobile industry

Administration. With the Local Government Act of 1972 coming into operation in 1974, the Birmingham City Council was replaced by a new Birmingham District Council, which took in the former borough of Sutton Coldfield. The council is also represented on the West Midlands Metropolitan County Council.

Birmingham's water supply comes from Wales, 73 miles (117 kilometres) away. Large-scale housing development took place between 1918 and 1939, and building since 1945 has included many multi-story apartment buildings, though by the 1970s there was a tendency to return to low-rise housing. The Birmingham District Council provides old people's and children's homes, residential and day nurseries, and child welfare centres. Local transport, formerly run by the corporation, is now controlled by the West Midlands Passenger Transport Executive. Snow Hill Railway Station closed its mainline services in 1967 but remained open for local services to Wolverhampton. New Street Station was completely rebuilt in the late 1960s, and a large shopping centre was erected above it. Electrification has reduced the train time between London and Birmingham to one hour and 35 minutes. The inner-ring road, completed in 1971, furthers the city's plan to segregate pedestrians and traffic and to keep unnecessary through and suburban traffic from crowding the city centre.

Cultural life. Birmingham is the cultural centre for a wide area, possessing three professional theatres and a

The Birmingham Repertory Theatre

number of amateur companies with their own theatres. The Birmingham Repertory Theatre was opened by Sir Barry Jackson in 1913 and fostered the early careers of, among others, such distinguished performers as Laurence Olivier, Ralph Richardson, Margaret Leighton, and Paul Scofield. A new theatre for the "Rep" is being built in the cultural complex in Broad Street. The Midlands Arts Centre for Young People, located in Cannon Hill Park, was an outstanding enterprise of the 1960s; when complete, the 15-acre site will house two theatres, a concert hall, a cinema, an art gallery, workshops, and studios. The City of Birmingham Symphony Orchestra plays in the Town Hall and in other Midland halls. Though subsidized by the Corporation, it faces the financial problems of all subsidized orchestras and is badly in need of a proper concert hall. It is hoped to include such a hall in the Broad Street development. The City Art Gallery is one of the largest outside London and is famous for its English watercolours and its pre-Raphaelite paintings. The Central Library contains more than 700,000 volumes, and work has started on a new library that will be among the largest public libraries in Europe. The city has two professional soccer clubs, and the Warwickshire County Cricket Club has its ground at Edgbaston.

The University of Birmingham, with over 6,750 students enrolled annually at the start of the 1970s, has an extensive campus at Edgbaston, with a medical school attached to the Queen Elizabeth Hospital; Birmingham was an early centre of medical research. The University of Aston, which opened in 1966 near the city centre, has over 3,000 full-time and more than 200 part-time students and specializes in technology. There are also educational and artistic facilities at the Birmingham and Midland Institute, now housed in the building of the Birmingham Library, a subscription library founded in 1779. The Selly Oak group of colleges provide education in religious and social subjects. King Edward's School, Edgbaston, founded in 1552, is the oldest educational institution in Birmingham; King Edward's High School for Girls stands on the same site, and other schools of the foundation are to be found elsewhere in the city.

The unusually small city centre, the lack of suitable buildings, and the short train journey to the theatres and galleries of London have hitherto adversely affected the city's native cultural life. But wide provision for a Birmingham-based culture is being made in the redevelopment. Despite controversy over details, the "new" Birmingham could ultimately rival the 19th-century city at its height, when Mendelssohn wrote *Elijah* for the city's music festival in 1846 and Joseph Chamberlain made the name Birmingham a European byword for civic progress.

BIBLIOGRAPHY. VIVIAN BIRD, *Portrait of Birmingham* (1970), a personal, controversial study of the city—does not supersede earlier books on Birmingham pre-1900, but useful on personalities and social changes since WWI; *Birmingham Post Year Book and Who's Who* (annual), an almanac of local detail, firms, personalities, and statistics; *Birmingham Public Works Planning* (1969), a booklet giving details of redevelopment, new road systems and buildings; *Birmingham and Its Regional Setting* (1950; reprinted with new foreword, 1970), an authoritative survey of the area; C. GILL and A. BRIGGS, *History of Birmingham*, 2 vol. (1952), an official history of the city up to 1938, scholarly and authoritative; ROBERT E. SCHOFIELD, *The Lunar Society Of Birmingham* (1963), an American study, detailed and scholarly, of a group of Birmingham-based manufacturers and professional men whose ideas had profound influence on the development of the Industrial Revolution; G.C. ALLEN, *The Industrial Development of Birmingham and the Black Country, 1860–1927* (1929; reprinted with corrections, 1966), standard work on its subject up to 1927; J.T. BUNCE *et al.*, *History of the Corporation of Birmingham*, 6 vol. (1878–1950), standard official work; WILLIAM HUTTON, *An History of Birmingham to the End of the Year 1780* (1781; 4th ed., 1808), an important early sourcebook.

(K.Br.)

Birth Control

The term birth control, coined by Margaret Sanger in 1914–15, denotes methods of preventing conception, whether involving the male or female, and now comprises all methods of fertility control, including abortion and sterilization. The term family planning came into use in the 1930s, followed by planned parenthood, denoting efforts (including the treatment of infertility) to regulate (whether by increase or decrease) the size of families. Other terms used by the early clinical movement have been contraceptive, contraception, voluntary parenthood, and family limitation and such Malthusian or neo-Malthusian terms as artificial checks, anticonception, anticonceptics, and prevenception.

HISTORY OF BIRTH CONTROL

The drive for fertility regulation by individuals has been evident in all cultures at all times, even in those societies in which social or religious rules favoured the abundant production of children. Until this century governments have more often been concerned to stimulate fertility than to encourage its limitation. The oldest recorded medical recipes to prevent conception, nevertheless, are contained in ancient Egyptian papyri, notably the Petri Papyrus, written about 1850 BC, and the Ebers Papyrus, written about 1550 BC, which describe many means of averting pregnancy. In ancient Greece and Rome there was much concern over fertility regulation. Soranus of Ephesus, a Greek gynecologist who worked in Rome, gave a lucid and detailed account of contraceptive methods in his writings during the 2nd century AD. He distinguished clearly between contraceptives and abortifacients (abortion-inducing agents) and observed that the prevention of conception is medically preferable to repeated abortions. This ancient knowledge was incorporated in the writings of Islāmic doctors and so made available throughout Europe, forming most of the scientific basis for contraception up to the late 17th century. It is evident from these works and other evidence that contraception had a place in medicine during this time—as a means either of helping to maintain the general health of women or of easing specific medical problems. Such sophisticated medical care was available to only a small percentage of the citizenry; widespread availability of medical birth control did not begin until the 20th century.

Malthusianism

Reformers and reformist groups. Since Thomas Malthus' *Essay on the Principle of Population* in 1798, birth control has increasingly been involved with concern for population size. Malthus, a clergyman, advocated the limiting of birth by sexual abstinence and late marriage. Both of these means were rejected by such reformers as Jeremy Bentham and Francis Place, who instead advised "such precautionary means as would, without being injurious to health and destructive of female delicacy, prevent conception" and "enable young men and women to remain chaste by making it possible for them to marry early." In 1831 Robert Dale Owen initiated discussion in the United States by publishing his book *Moral Physiology*. Probably the most detailed account of contraception techniques since the writings of Soranus was contained in *The Fruits of Philosophy: or, The Private Companion of Young Married People*, by Charles Knowlton, which appeared in the U.S. in 1832 (and in England in 1834); Knowlton himself was fined and briefly imprisoned as a result of his publication.

As evidence of industrial poverty and overcrowding increased in the 19th century, the secular Free Thought movement adopted neo-Malthusianism as part of its labour reform proposals. In the 1860s one of its members, George Drysdale, started the Malthusian League, in which he was joined in 1874 by Annie Besant. Under continual attack by the churches and the medical profession, however, it made little headway. In 1876 a Bristol publisher was imprisoned under the 1857 obscene publications act for selling "pornographic" books, including *The Fruits of Philosophy*, published by the freethinkers. Incensed, Charles Bradlaugh (leader of Britain's National Secular Society) and Annie Besant reissued the pamphlet themselves, notifying the police. Action was brought against them for corruption of youth, but the conviction was quashed on grounds of faulty indictment. The trial

publicity boosted neo-Malthusianism, and the Malthusian League, reformed, now aimed "to agitate for the abolition of all penalties for the public discussion of the population question" and "to spread among the people by all practicable means a knowledge of the law of population, of its consequences, and of its bearing upon human conduct and morals." A medical branch was formed to gain medical cooperation, and by agitating for discussion in the medical press, to obtain recognition of contraception's scientific basis. Its success, however, was small.

Malthusian leagues started in France, Germany, and Holland. The one in Holland started a medical service under Aletta Jacobs in 1882 and some clinics from 1890 on, and the concept of clinics spread elsewhere. Interest, despite powerful opposition, was also growing in the United States, largely involving sexual emancipation. An act of Congress penalizing contraception was passed in 1873 through the efforts of Anthony Comstock, who had founded the Society for the Suppression of Vice in New York. In 1912 Margaret Sanger began writing there in the Socialist newspaper *The Call*. After a visit to Europe to collect contraceptive information, she issued *Woman Rebel*, a monthly journal, which was barred from the U.S. mails and caused her to be indicted for "misuse of mail." In defiance she printed *Family Limitation* and left for Europe to meet her continental counterparts and see the Dutch clinic. She returned in October 1915 prepared to stand trial, but the charges were dropped. In 1916 Margaret Sanger and her sister opened a clinic in Brownsville, Brooklyn, after the police had closed it, and were arrested and convicted on charges of "maintaining a public nuisance." The appeal court's findings, however, enabled doctors to give contraceptive information "for prevention or cure of any disease." In 1915 the feminist Mary Ware Dennett formed the first U.S. birth control society. Battles against the Comstock laws ensued, but the American movement grew, becoming in 1942 the Planned Parenthood Federation of America; and in 1963, after merging with the World Population Emergency Campaign (a fund-raising organization for global birth control), it became Planned Parenthood-World Population.

Clinics and family planning associations

During and after World War I, feminist activities reinforced the neo-Malthusians, though anti-contraception legislation increased. In London in 1918 Marie Stopes published *Married Love*, followed by *Wise Parenthood*, linking birth control with marital harmony. In 1921 she started a clinic for poor women, followed by the Malthusian League demonstration clinic, modelled on a state-aided maternity centre. In 1922 she founded the Society for Constructive Birth Control and Racial Progress. Pressure on the minister of health to make advice available in maternity clinics was growing. In 1927 the medical and scientific Birth Control Investigation Committee was formed, producing after research a spermicide, "Volpar," and instituting testing of contraceptives for an "approved list." At last in 1930 the minister of health authorized local authorities to give advice to married women in medical need: and the National Birth Control Council was formed to encourage them. Progress was slow and the limitation to medical need restrictive, so voluntary clinics continued to grow, all of them, except that of Marie Stopes, uniting in 1938 to form the Family Planning Association.

By 1970, 935 British FPA clinics were advising 650,000 patients a year. In the United States, Planned Parenthood-World Population had 524 affiliated family planning centres in 38 states and sponsored research and family planning education. The Swedish association, founded in 1933, had government backing with such statutory measures as support for sex education in schools. The Dutch association received government aid and had over 50 centres. A new French association was organized in 1956, but adverse laws impeded it and similar voluntary movements in many countries. Such restrictions are being steadily reduced and are often replaced by laws and measures promoting birth control. Communist countries have introduced both abortion and contraception services for the well-being of individuals, at times adapting them to public policy. China officially approved birth control in 1956 with a countrywide campaign; in 1958 this policy changed, but birth control is now routinely facilitated. Indeed, China seems to be the developing country most successful in achieving motivation and making all methods available. In the developing countries of Asia, Africa, and Latin America, government programs are often linked with economic development and population policies.

By 1970, 103 countries throughout the world had voluntary family planning associations. There were some government provisions in 63 countries. Social surveys have indicated that about 90 percent of couples in the United States use contraceptives, about 80–90 percent in the United Kingdom, and about 53 percent in Japan.

Growth of a birth control trade. The opportunities for profitable manufacture and retailing of contraceptives have always been obvious. The proliferation of quack contraceptives and abortifacients prompted pioneer reformers to provide skilled advice for the poor. There were well-intentioned traders, too—for example, W.J. Rendell of London in 1885 marketed pessaries of quinine bisulfate in cocoa butter. As trade interest grew so did advertising, harassed by organizations like the 1851 British Union for the Discouragement of Vicious Advertisements. But with growing acceptability, advertising increased, appearing in pharmaceutical and professional journals and the press, so far as laws and trade conventions allowed. Consumer protection, however, until recently was limited. In 1962 the American Consumers' Union, and in 1963 the British Consumers' Association, began evaluations of contraceptives with criticisms of advertising, efficacy, instructions, and prices. Many countries now have official bodies that assess and regulate the quality and safety of contraceptives, particularly of drugs.

Traditional outlets for contraceptives for many years were barbers, back-street stores, tobacconists, and mail-order companies. Today in Sweden and Denmark registered pharmacists are required to stock them, and in such places as South Australia and France pharmacies are the *only* outlets. Excluding China, about 80 percent of the world distribution of contraceptives is through trade, 19 percent through governments, and 1 percent through voluntary outlets. The encouragement of trade outlets is therefore an important means of promoting birth control.

International developments. In 1900 British, Dutch, French, and German leagues formed the Fédération Universelle pour la Régénération Humaine. International work gathered momentum from the 1920s, with Margaret Sanger travelling widely to promote it. Periodic conferences took place. The International Committee on Planned Parenthood was formed in 1948 at a conference in Cheltenham (England). At a 1952 Bombay conference this committee became the International Planned Parenthood Federation. Cochairmen were Margaret Sanger and Lady Dhanvanthi Rama Rau (head of the Family Planning Association). By 1964, with regional subgroupings, it was arranging regional and world conferences; encouraging research and promotion schemes; training staff; producing handbooks, literature, and medical and research journals; devising method-testing criteria; and providing testing facilities, an information centre, library, and so forth. That year it gained consultative status with the Economic and Social Council of the United Nations, and later with the International Labour Organisation, the World Health Organization, UNICEF, UNESCO, and the Food and Agriculture Organization. By 1971 it had 70 member countries. It functions as a filter for funds to help projects and services throughout the world.

International Planned Parenthood Federation

After the mid-1960s, the United Nations Economic and Social Council and other specialized international agencies, progressively backed by the General Assembly and by the findings of the Population Commission, urged and supported measures to aid the adoption of voluntary birth control in countries around the world. In 1968 the ILO

decided to promote family planning in industrial medicine, welfare, and education; and UNESCO included a study of the interrelation of education and fertility in a ten-year mass education program. For the UN's activities 1968 had been designated the International Year for Human Rights, and the UN General Assembly and Tehrān Conference asserted that family planning was a basic human right. The secretary general started his own fund to aid UN work in the field and sponsored a declaration by 30 non-Communist world leaders that family planning enriched human life.

Contemporary views on birth control and population. In the world today there are four general views on birth control: (1) Man should intervene in natural increase and selection only to the extent possible by resorting to abstinence or "natural" contraceptive means (such as coitus interruptus or the rhythm method). (2) Man can achieve adequate numerical and eugenic control simply by individual, voluntary use of all medically acceptable contraceptives and techniques of abortion, sterilization, sex selection, and diagnosis of fetal abnormality. (3) Man should employ mass temporary sterilization with the granting of individual requests for reversal to bear children, in order to achieve more effective, though still essentially voluntary, individual control. (4) Man should resort to relatively strict governmental controls over population numbers, quality, and distribution by means of long-term sterilizing or fertility-reducing agents, compulsory sterilization of individuals, eugenic procedures, and the like—all because, with population outrunning world productivity and resources, voluntary cooperation is insufficient.

All these proposals require massive educational programs and increased administrative and technical efficiency. The entirely voluntary measures—(1) and (2)—have not yet been fully tried even with known contraceptive means. Their possible effects when universally applied can perhaps be gauged by experiences in Europe, North America, Japan, and other countries where individuals have resorted to birth control in response to general economic and social pressures, though not always in accord with government policies. In general, governments have been able to encourage trends more easily than reverse them in opposition to the social climate and individual experience. The involuntary measures—(3) and (4)—raise serious ethical, political, and administrative issues, such as deprivation of freedom. And, in any event, scientific means of mass fertility and eugenic control might take many years to develop. Some would require a degree of international consensus not yet apparent.

TYPES OF BIRTH CONTROL

Contraception. To satisfy individuals, contraceptive methods need to be sure and accurate (as well as physically and aesthetically satisfying), whereas to achieve social policies and satisfy society, only general effectiveness is needed. In any case, however, the side effects and dangers of new contraceptives need to be, and generally now are, systematically assessed by laboratory techniques and before human use and subsequently surveyed by epidemiological studies.

Mystical versus rational methods of birth control

There are two modes of approach to birth control: mystical methods (such as tying knots at bridal ceremonies, eating dead bees, and wearing amulets) have flourished in primitive societies, ancient and classical times, and the Middle Ages, and persist in folk practice today; and rational methods seek to influence known or surmised processes involved in reproduction. Rational methods subdivide into those that involve some particular behaviour only (nonappliance methods) and those requiring a special agent (appliance methods).

Nonappliance methods. Since ancient times, oral and anal intercourse, homosexuality, special postures, gestures merely suggestive of coitus, and a host of other techniques have been resorted to in order to achieve satisfaction but avoid conception. Soranus advocated holding the breath, and coughing, jumping, and sneezing after intercourse in order to expel semen; he also hypothesized infertile times in the menstrual cycle. Many peoples have also recognized that most, though not all, women do not conceive (because ovulation is suppressed) while suckling, and thus, as a means of birth control, suckling has been prolonged.

The so-called rhythm method requires sexual abstinence during the woman's "fertile period"—those days around the time of ovulation when conception is most likely to occur. The other days of her menstrual cycle are known as the "safe period." A "safe period" has been used in many societies, but until the 1920s there was not sufficient physiological understanding to determine when ovulation occurs in relation to menstruation. The methods now used entail recording, under medical guidance, dates of menstruation and the temperatures taken on daily awakening to ascertain the slight rise that occurs one to three days after ovulation. The "safest" period is after this has happened. Other safe days must be calculated in advance from the probable time of ovulation after menstruation. To do this, one must assume that the pattern is regular, which is not always the case. Even regular cycles can be altered, and temperature readings can be confused by pelvic inflammation, illness, emotional disturbance, climate, altitude, menopause, or conditions after pregnancy or miscarriage. Medication can sometimes be used to regularize cycles, but in at least 15 percent of women regularity cannot be achieved.

Coitus interruptus, withdrawal of the penis before ejaculation, has probably achieved more fertility control through the ages than any other method. Failure of the method relates to the timing of withdrawal and the deposit of some semen before the climax. It is criticized as unsatisfying to both partners and has been rejected in some Eastern cultures as unseemly. It is permitted to Muslims but forbidden to Orthodox Jews.

Coitus obstructus is full intercourse with the ejaculate suppressed by depressing the base of the urethra to force the semen into the bladder. Coitus reservatus (also called *Karezza*) is intercourse with ejaculation intentionally withheld.

Appliance methods

Condom. The condom, or sheath fitting over the penis, has long been used as protection, and as early as the 16th century the physician Gabriel Fallopius designed a medicated linen sheath for the glans or tip of the penis for the purpose of preventing venereal infection. By the 17th century, the condom was utilized as a contraceptive as well. Early condoms were generally made of animal gut, hemmed at one end, or fish membrane and were often inefficient. The marquise de Sévigné sarcastically described goldbeater's skin (intestinal membrane from cattle) as "armour against love, gossamer against infection." Casanova used condoms as contraceptives. Legend is confused on the origin of the term condom—one story telling of a man named Condom devising such a contraceptive for Charles II of England. Since the 1840s most condoms have been made of vulcanized rubber or, since the 1930s, of latex. At first they were usually washable but now are generally disposable and slightly lubricated. Efficient, convenient, but still disliked for its dulling of sensation, the condom fails mainly because of irregular use.

A comparable device, the barrier-spermicide "C-film," has been used in Hungary since 1968. A thin, translucent, pliable, water-soluble film four centimetres square, containing a highly active nontoxic spermicide, is placed over the moistened glans or in the vagina. It dissolves quickly and appears to make the cervical mucus impenetrable while destroying the sperm.

Methods per vaginum. Among the methods used by women, one of the oldest is "douching"—the practice of flushing out the vagina, generally with a liquid solution, after coitus—which is intended to impede the passage of spermatozoa or to remove them altogether. The Romans used oil and soft wool for this purpose, the Greeks vinegar. Charles Knowlton recommended various substances, including alum, hemlock, green tea, zinc, sugar of lead, and baking soda; other 19th-century writers advocated quinine, tannin, opium, prussic acid, iodine, strychnine,

alcohol, and carbolic acid. But no matter what type of solution is used, the method has low effectiveness.

Also ancient in origin are "spermicides," sperm-destroying and sperm-impeding substances prepared in the form of aerosol foams, jellies, creams, soluble suppositories, and foaming tablets, which are inserted into the vagina before each act of intercourse (often by means of a slender plastic applicator). In ancient times one ingredient was often a sticky substance such as honey or olive oil; the Egyptians also used crocodile dung, and the Jews used onion or peppermint juice. Spermicidal creams and jellies are used with a diaphragm or condom. Though, when used alone, some are theoretically in the upper range of effectiveness, as a group they protect only for short periods and are not generally reliable.

A number of other devices may also be placed in the vagina to cover the cervix, or neck of the uterus, and thus to prevent fertilization by blocking the passage of sperm to the ovum. An effective early method, subsequently promoted by neo-Malthusians, was the vaginal tampon or sponge (generally used in conjunction with a spermicide). Similar practices were known even earlier: Casanova reported inserting half a lemon in the vagina; the Japanese once used disks of bamboo tissue paper. In 1823 a German physician, F.A. Wilde, started taking wax impressions of the cervix and making caps of metal and subsequently of crepe rubber to fit; these were the first true caps. They were made to be inserted and changed regularly by a physician. Although today a physician or trained nurse is still required to fit the cap properly and give instruction in its use, latex cervical caps are available in several sizes and can be inserted by the user. In contrast to caps, diaphragms are larger in size and consist of latex or soft rubber mounted on a spring metal rim; like caps, however, they are placed in the vagina in front of the cervix and, carefully used, have a high rate of effectiveness. They were first produced in 1880 by W.P.J. Mensinga, professor of anatomy at Breslau, and because they were promoted by Dutch neo-Malthusians, they became known as the "Dutch caps." First made of vulcanized rubber in three sizes, they now appear in latex in 21 sizes. Highest protection is achieved when the diaphragm is used in combination with spermicidal jelly or cream.

To be included within this category are various means of preventing intercourse altogether—such as chastity belts (any of various belted devices to prevent access to the vagina) and infibulation (tying together the prepuce or labia with clasps or stitches). The former was used in medieval times and the latter in ancient Rome. Of minor importance (but of ancient usage) is artificial retroversion of the uterus—that is, turning the uterus backward on its axis so that the cervix and vagina no longer connect.

Intrauterine devices (IUD's). In the 19th century, physicians sometimes inserted intracervical devices (such as a stem pessary, collar buttons, or wishbone) into the cervical canal to be changed and inspected at intervals. About 1920 silkworm gut loops attached to a silver wire to aid removal were used; it was feared, however, that the silver link provided a route for infections. To avoid this, a German, E. Gräfenburg, designed various intrauterine devices—first a star of silkworm gut tied with silver wire, later a coiled silver wire ring. Although he excluded patients with pelvic inflammation and used aseptic procedures with good results, the American Medical Association reported 17 deaths and 176 serious conditions when the method was used by other physicians, and it was discredited.

From 1959 on, however, new reports and research on intrauterine devices (IUD's) made of silver, stainless steel, and molded plastics that caused no corrosion brought promising results. Within five years, a variety of such devices had been developed and subjected to intensive tests. When they were skillfully inserted and retained, their safety and effectiveness were found to be considerable. Molded in a number of simple shapes such as spirals, loops, and circles, they are about 20 to 30 millimetres in diameter.

The IUD is inserted into the uterine cavity in a simple medical procedure and may generally be left in place until a pregnancy is desired. The IUD's mode of action is not yet clear. Substantial research evidence indicates, however, that it does not interfere with ovulation but may prevent implantation, and that if a conception persists the fetus is not damaged. Some women experience cramping or bleeding after IUD insertions, but these side effects usually subside or can be controlled by medication. In some 25 percent of women the device will not remain in place or must be removed. Because the successful insertion of an IUD gives lengthy protection, the method has been promoted in densely populated nations of the developing world. Adverse side effects needing medical attention, as well as failure rates, however, make it less successful than was at first hoped.

Oral contraceptives. There have always been potions that it was hoped would prevent conception without inhibiting the enjoyment and spontaneity of coitus. Some had purely symbolic content—such as drinking water that had been used for washing a dead person. Some—such as infusions of willow bark, plantains, walnut leaves, saffron, potions of gunpowder, water in which smiths had quenched their forceps, and pills of oil and quicksilver—were intended to induce physiological change. Some herbal medicines may have had some effect: tribeswomen in central Paraguay took extracts of *Stevia rebaudiana*, now found to affect the fertility of rats; the barbasco root used in Mexico provides a basic ingredient of pills today. Many recipes, however, were not only ineffectual but harmful, and Soranus, St. Jerome, and 19th-century writers warned against them.

Although oral contraception on a scientific basis was forecast as early as the mid-19th century, the first effective "birth control pills" achieving physiological suppression of ovulation were not successfully developed until the 1950s. In 1955 Gregory Pincus, a research biologist at the Worcester Foundation for Experimental Biology (Shrewsbury, Massachusetts), assisted by the American pharmaceutical firm of G.D. Searle, found a group of orally effective compounds. In large-scale field studies conducted in Puerto Rico and Haiti the pills were found effective and acceptable.

The pill

The pills contain hormonelike substances that enter the bloodstream and set up a chain reaction of the nervous system that inhibits the production of ova (thus reproducing the biochemical action by which ovulation is stopped naturally during pregnancy). Despite early side effects such as nausea, dizziness, and headache, bleeding and spotting, and weight gain among some users, which can generally be controlled, oral contraceptives were found to be acceptable as well as effective. Further research has yielded improved compounds that seem to have reduced these difficulties. The pills, available only by prescription and under the supervision of a physician, are not recommended for use by women with any of a number of medical tendencies such as vascular, thromboembolic, liver, or diabetic disorders, or cancer.

Research suggests that there may be several hazards in the use of birth control pills. There appears to be a possible relation between their use and cancer of the breast and uterus. There is some evidence of increased hypertension, abnormal glucose tolerance, and other biochemical changes. There is a slight possibility of genetic damage to the ovarian egg. The pills can aggravate such allergies as asthma, eczema, and migraine and such other conditions as alopecia, psoriasis, epilepsy, multiple sclerosis and otosclerosis, and porphyria. The worst hazard seems to be a ninefold increase in thromboembolic disorders (involving clotting of blood or plasma). The publication of such findings has led to increased anxiety about the use of the drugs. Nevertheless, such dangers—or those also arising from the use of IUD's—may be less frequent than serious complications of unwanted pregnancies.

Sterilization. Permanent contraceptive sterilization by means of castration of men or hysterectomy and ovariectomy in women has been largely replaced by surgery that leaves the sex glands intact but severs or obstructs the

tubes through which the sperm or ova pass to accomplish fertilization. Known as tubal sterilization in the female and vasectomy in the male, this form of sterilization impairs neither sexual appetites nor performance. Although the operation is generally irreversible, fertility may sometimes be restored spontaneously or by subsequent surgery. Satisfactory pharmacological means of sterilization have not yet been found.

Sterilization can present various social and psychological problems. Voluntary sterilization, for instance, is rarely defined in law or is defined inadequately, and thus personnel performing such operations may risk being objects of civil or criminal action. The person accepting sterilization may subsequently experience regret and mental distress. Compulsory sterilization for eugenic reasons is legal in some countries, but implementation of the law usually presents problems. Nevertheless, voluntary sterilization is being promoted increasingly in the United States and the United Kingdom, as well as in Asia, where India has a major sterilization program.

Induced abortion. Abortion is the termination of pregnancy before the fetus has attained the ability to live independently. A fetus is usually considered viable when weighing 1,000 grams or more and after 20 to 28 weeks of pregnancy. (Medical and legal definitions vary as to the required number of weeks.) Induced or artificial abortion is a major birth control method, to be distinguished from spontaneous or accidental abortion, popularly known as miscarriage.

Methods used to induce abortion depend on the woman's condition and the duration of pregnancy. If possible, during the first 12 weeks, there is surgery consisting of dilation of the cervix and evacuation of the uterus by some means of scraping, suction, or vacuuming. After the first 12 weeks the size of the fetus and placenta usually demands a vaginal or abdominal hysterotomy (surgical incision of the uterus). Various chemical measures for emptying the uterus are also possible and may be combined with surgery. Safe and reliable oral abortifacients have yet to be found, but encouraging experiments are being carried out.

Although most religions have condemned induced abortion, the application of severe criminal sanctions to deter its practice became common only in the 19th century; induced abortion was usually made illegal but was sometimes permitted or tolerated to preserve the life of the mother. The first important departure from the 19th-century pattern came in the U.S.S.R. in 1920, when the post-revolutionary government authorized abortion at the request of the mother; in 1936 another decree restricted legal abortion practice, but in 1955 abortion on request was restored. Since the late 1940s a movement to legalize abortion on all or some grounds gained impetus, nourished by widespread support for women's rights, by official concern about rising birthrates and a rapidly expanding world population, and by unhappiness about very high maternal death rates associated with illegal abortions. First Japan and some of the eastern European countries authorized abortion at the request of the mother. The Scandinavian countries and Swiss cantons followed with rather more restrictive grounds and procedures for authorization. Great Britain, Haiti, and some states of the United States liberalized the laws during the late 1960s, though greater restrictions were imposed in Romania, Hungary, and East Germany. In Asian countries other than Japan and China, abortion is generally not covered by legislation, though this is under discussion in India. In different countries the grounds for abortion vary: they may be for preservation of the life of the mother, for other strictly medical reasons, for eugenic or some social reasons, or simply "on demand," without restriction; of course, there are no grounds where it is wholly banned. Where laws are liberalized, abortions tend to increase steeply, at least at first. The proportion of pregnancies ending in induced abortion, legal and illegal, is often very high; for example, in The Netherlands and West Germany it is estimated at 25 percent, in France 50 percent, and in Japan over 50 percent.

The move to legalize abortion

Legal abortion for private or social reasons, as granted by doctors and official committees or boards and carried out in hospitals or approved centres, is usually allowed only during the first three months of pregnancy; and some countries do not permit a second induced abortion within six months of the first. Nevertheless, some countries have machinery for appeal against refusal. And a woman refused legal means may turn to illegal ones; in Yugoslavia, for instance, it was found that 58 percent of those refused legal abortion resorted to illegal abortion. Criminal abortion is hazardous and is not wholly displaced by legal abortion, even under the most liberal laws. Mortality in criminal cases is estimated at 35 to 94 per 100,000; in legal, but nonmedical, cases performed in the first three months, it is about 1.2 per 100,000 (though it rises to about 40 per 100,000 in legal cases of all kinds). After abortion the likelihood of abnormal pregnancies, premature births, illness in the woman, or even sterility is increased. Of legal abortees a high proportion are married women; and of criminal abortees the largest proportion are the very young, the unmarried, previous abortees, and wives pregnant at marriage.

Characteristics of legal and illegal abortion

Abortion can be followed by feelings of guilt, disturbed sexual relations, and psychologically induced illness. Studies show a high proportion of abortees to be maladjusted; in a study in Yugoslavia only 24 percent were considered well balanced. In cases in which legal abortion has been denied, on the other hand, the children may suffer; one study revealed that children born after refused requests for abortion showed a high incidence of delinquency, social maladjustment, educational failure, and need for psychiatric help. For purposes of birth control, medical opinion strongly favours contraception over abortion.

Infanticide. Infanticide, once an important method of fertility control, was common in classical times and is practiced in some primitive or isolated cultures even today. The Christian emperor Constantine of Rome legislated against it in the 4th century AD, but with poverty, lack of facilities for preventing conception, and condemnation of illegitimacy, it continued in Europe, despite punitive laws, into the 19th century. With the spread of contraception and legal abortion, it is now rare and is generally associated with puerperal disorders or with parents' mistreatment of their children (the "battered baby" syndrome).

Although there are arguments for euthanasia for disabled babies who survive birth, infanticide is now generally illegal.

Abstinence and ceremonials. Both abstinence and late marriage are used by some individuals and members of religious groups whose conscience forbids contraception and abortion. Some governments promote such means as population curbs.

Some customs, ceremonials, and rites have a recognized contraceptive element; for example, the male pubertal rite of subincision (splitting the underside of the penis) among Australian aborigines reduces fertility, and there are cultures forbidding intercourse in child marriage or during menstruation, lactation, and periods of religious celebration.

Possible future methods. Because it is unlikely that a single contraceptive method can be universally applied, the goal of birth control technology is to develop a variety of simple, effective, inexpensive, and harmless new methods to meet the needs of all people of all cultures and faiths. Contraceptive research seeks new "appliance" and "nonappliance" techniques and improvements of existing ones in order to control or interrupt the sequence of events leading to pregnancy.

Basic research into reproductive physiology has begun to identify possible new approaches to controlling fertility. Formerly, birth control research was unfashionable and neglected; medical and scientific personnel were frequently hostile or indifferent, and money and facilities were not forthcoming. But with the acceptance of birth control, such research is now intensified. First, there is the research designed to improve present oral and other

methods in order to achieve long-term action (by oral dose, injection, or implants). Second, there is research dealing with the role of the glands and the central nervous system in all stages of reproduction. This research may lead to means of inhibiting glandular secretion, controlling the production and transport of ova and spermatozoa, or intervening in fertilization or placenta formation. To influence processes such as these, many agents and combinations of agents are being investigated. The chief problems are still the dangers of partial control, the maintenance of control, toxicity, and unwanted side effects.

SOCIAL AND PSYCHOLOGICAL ASPECTS OF BIRTH CONTROL

Throughout history and among most cultures, power and prosperity have been associated with growing populations. Social and economic status and advancement of families often depended on numbers of offspring. With high mortality from disease, war, and other natural hazards, the survival of society as a whole and of families was generally best served by pronatalist attitudes and institutions. Not infrequently, however, limitation of reproduction has been motivated by the contrary need to maintain or improve health and living standards. This happened when, with temporary freedom from war or epidemics, population rose and methods of cultivation could not provide enough food, work, or housing—for instance, in the Greek city-states, in medieval Europe prior to the 14th-century decimating plague called the Black Death, and in 19th-century industrial cities favoured by the control of disease. There have always been innumerable reasons why some individuals and groups should wish to control their fertility.

Population problems and birth control

The abiding personal, ethical, and economic reasons for birth control have been intensified in recent decades by the so-called population explosion. After World War II the dramatic lowering of death rates through mass public health measures resulted in the doubling of the growth rate of the world's population in about 15 years. It was less than 1,000,000,000 in the year 1800; it is now more than 3,500,000,000 and by the end of the 20th century could be twice that number. Furthermore, the most rapid population increases are occurring mainly in the countries with the lowest per capita economic production—those that have the greatest need to raise production and living standards but are least able to keep pace economically with rapid population growth. Many of these developing nations are in Asia and Latin America, and contain 70 percent of the world's population. To help solve this problem, birth control has been advocated in order to bring population growth rates to levels at which agricultural, industrial, and social developments can not only maintain but improve the quality of life. But developed countries are also experiencing problems of population imbalance, though less acute; and in most of them, birth control by one means or another is common. The major social factors affecting the objectives and practice of birth control are religious and ideological beliefs, biological urges, cultural and familial patterns, the health and status of women, the pressures of national class or group interests, ambition, education, occupational patterns, living standards, housing, and so on.

Social and ethical issues of induced abortion

Special moral and social questions attach to abortion. Those who object to abortion on ethical grounds begin with the premise that, except where the mother's life is at stake, there is no rational basis for distinguishing between the newborn infant and the fetus: the fetus grows, changes, and even reacts to its environment in the mother's womb; the infant is in every sense only a potential member of society, just as the fetus is. Proponents of abortion, on the other hand, make somewhat the following arguments: prior to the fourth month of pregnancy at least, the fetus has acquired so few of the characteristics commonly identified as human that most people do not equate destruction of the fetus with murder; population control demands (because contraceptive techniques will never be perfect) that abortion be available to those mothers who wish to terminate a pregnancy for good cause; it is inefficient, and socially unwise, to impose criminal sanction against behaviour acceptable to a significant body of opinion; and, finally, only if women are given complete and sure control of births can their complete emancipation be assured.

Religious attitudes and beliefs. The major religions of today, influenced by the condition of the times in which they took root, support marriage and reproduction and have been fearful of measures like birth control that might affect marital stability. They have all recognized the power of the sexual urge and have sought to direct and control it, some more fearfully than others. Birth control has seemed irrelevant or antagonistic to reproduction and to spiritual goals as they have been understood, and the churches' slow processes of evaluating new knowledge and other circumstances have retarded acceptance of it. Militant religious groups have swayed governments on the issue of birth control and inadvertently inspired the growth of independent services. In the West, religious acceptance of birth control came only when sexual pleasure and physical union were accepted as important to marriage and when problems of overpopulation were recognized. The newer secular ideologies of humanism and socialism were concerned not with sexual morality but with individual freedom, well-being, and justice; the control of fertility was accepted as an individual right and means to prosperity.

Religions are now generally agreed on the responsibility of parents, but grounds for control and methods are still disputed. Furthermore, although religious observance has declined, its precepts, sanctions, and attendant emotional reactions of fear, guilt, and shame, which have permeated social custom, continue often to exert influence.

Judaism. Historically, Jewish doctrines on marriage and procreation have been related to national ambition and struggle for survival; survival itself was identified with a close and disciplined community in which the individual was perpetuated through family and name. Celibacy has had little place in Judaism; there is an obligation to have children, though love and companionship are deemed equally important in marriage. The seed has traditionally been considered viable; thus, coitus interruptus (onanism) was banned as "spilling the seed," though abortion is allowed if the mother's life is threatened. There is a long tradition of the use of sterilizing potions, vaginal tampons, and even abstinence during famines. Orthodox sections condemn male methods of birth control and those interfering with spontaneity, but for the mother's health certain female methods are allowed, after consultation with medical and rabbinic authorities. Reformed and Conservative branches of Judaism, however, urge proper education in birth control as the means of enhancing spiritual values in marriage and achieving the welfare of mankind. Female methods, subject to medical considerations, are acceptable, and a high proportion of Jewish couples use birth control.

Early Christianity. In the early church, formed in a time of revulsion against the prevailing sexual excesses of the Roman Empire, celibacy was considered superior to marriage because the Second Coming of Christ was thought to be close at hand; it was not that reproduction was no longer needed, but that spiritual perfection was considered to be impeded by worldly distractions. While self-discipline and the denial of appetites were emphasized over the whole range of human behaviour, coitus as the vehicle of original sin was particularly restricted and came to be associated with guilt, shame, and uncleanness. Its primary rationale was childbearing, which was not to be frustrated. Such feelings still have force to inhibit the use of some birth control measures.

Eastern Orthodoxy. Although indulgence in sex is considered a handicap to spirituality in the Eastern Orthodox Church, parenthood is a duty, and economic pressures must ideally be met by frugality and by social aid for large families. Traditionally, only abstinence is permitted, because interference with any phase of reproduction is forbidden and the killing of sperm is considered no dif-

ferent from abortion and infanticide. However, the church, while considering the use of contraception to be the result of failure in spiritual focus, has not sought to hinder the distribution of contraceptives.

Roman Catholicism. Roman Catholic doctrine on birth control stems, in particular, from the teachings of St. Augustine, St. Thomas Aquinas, and legalistic and scholastic elements within the church, who held that procreation was the natural result of coitus and the primary purpose of marriage and that contraception was therefore a frustration of natural law. Throughout most of the history of the church, however, there have been a few theologians who have taken a more sympathetic view of sexual pleasure in marriage; and this concept, together with that of avoiding children because of poverty, gained ground by the 19th century, when the use of contraceptives began to be widespread.

During the 19th century there developed the idea that birth control by "natural" means—particularly the rhythm method—was not a positive exclusion of procreation and was therefore to be regarded as morally licit if properly used. Formal approval of the rhythm method was first given by Pope Pius XI in his 1930 encyclical *Casti Connubii.* It was repeated in 1951 by Pope Pius XII, who specified that Catholics might practice the rhythm method for a number of medical, eugenic, economic, and social reasons. Other methods, as well as abortion and sterilization, however, were still rejected. In 1958 he forbade the use of oral contraceptives or anovulants except for some therapeutic purposes.

A number of Catholic theologians, doctors, and laymen had begun to question whether coitus reservatus and the rhythm method were the only moral ways of family planning. Some claimed that oral contraceptives could be morally used, since they postpone ovulation without interfering with the sexual act. Finally, some sought full acceptance of contraception and contraceptives.

The debate among Roman Catholic theologians

Of the many reasons for the growing uncertainty among Catholic moralists about birth control, two reasons were prominent. First, increasing emphasis was being placed upon the "secondary" end of marriage—mutual love and fulfillment. This was the result of various factors, such as a better insight into the nature of interpersonal relations, the changing status of women, a concern for health, the pressures of increasing population, the spread of higher education, and the acceptance of contraception by other denominations and its widespread use. Second, many moralists felt that in the past undue emphasis had been placed on the physical element of the sexual act and that instead of being subordinated to man's total well-being, the biological aspect had acquired an undeserved value.

Against this background, leading theologians concluded that the entire problem required intensive reappraisal. To study the question from all possible angles, Pope John XXIII in 1963 appointed a commission of theologians, gynecologists, psychologists, demographers, and married couples. While the commission continued its deliberations, the Vatican Council in 1965 approved its "Pastoral Constitution on the Church in the World of Today." Although the council did not dispute previous bans on contraceptives, the constitution opened the door to future change by eliminating the distinction between "primary" and "secondary" purposes of marriage. Thus it placed the fostering of conjugal love on the same plane of importance as the procreation and education of children. It also established the primacy of individual conscience.

The international commission appointed in 1963 reported to Pope Paul VI in 1966, but its recommendations were not made public until April 1967, through an unauthorized press "leak." The majority, it was noted, urged the pope to define doctrine so that chemical and mechanical means of birth control would be morally licit. The minority recommended no change in the interpretation of teachings. In July 1968 Pope Paul VI delivered his encyclical *Humanae Vitae* ("Of Human Life"), in which he declined the advice of the commission and firmly restated the traditional view.

Humanae Vitae

The pope noted the social and economic difficulties caused by rapid population growth, but, he said, chemical and mechanical contraception are not permissible means of coping with these problems. He urged that, instead of sponsoring fertility control programs, governments encourage social and economic progress, and though noting the personal difficulties imposed on couples using the rhythm method, he said that rigorous self-discipline required to use this technique may enhance rather than harm the purity and value of conjugal love. He repeated Pope Pius XII's plea that medical science help make the rhythm method more reliable but yielded no ground to demands that the traditional views on birth control be modified. Following the traditional natural law position, he declared illicit "every action which, either in anticipation of the conjugal act or in its accomplishment, or in the development of its natural consequences, proposes, whether as an end or as a means, to render procreation impossible," including abortion and sterilization as well as birth control methods other than rhythm and coitus interruptus. The teaching in the encyclical carried the authority of the pope but was not infallible. It set off widespread controversy and a further polarization of Catholic views. Dutch bishops, for example, viewed the pronouncement as only "one of many factors" to be considered in making decisions on child spacing and family size, along with "mutual love" and "social circumstances." It is now widely accepted that individual conscience must be the final guide, though discussion of the issue continues.

Protestantism. Martin Luther and John Calvin held an Augustinian view of coitus even stricter than that of the Roman Catholic Church of their time, but they gave marriage a status equal to celibacy and did not consider children necessarily the primary motive of marriage and coitus. While their view of sex placed Protestantism in opposition to contraception, the emphasis on other aspects of marriage eventually aided acceptance of the concept of birth control. Contraception was accepted by some Protestant clergy by the end of the 18th century, though general consensus came only in the 1930s. Since then, most Protestant churches have evolved a concept of responsible parenthood that strongly sanctions the use of contraception within marriage. The shift from opposition to support of family limitation is exemplified by the Lambeth Conference of Anglican bishops, which specifically condemned contraception in 1908 and 1920, cautiously approved it in 1930 in cases in which there was a clearly felt moral obligation to limit or avoid parenthood, and finally in 1958 approved it strongly by declaring, "The responsibility for deciding upon the number and frequency of children has been laid by God upon the consciences of parents everywhere. . . . The *means* of family planning are in large measure matters of clinical and aesthetic choice, [provided] they [are] admissible to the Christian conscience" (Proceedings of the 1958 Lambeth Conference, Bishops of the Anglican Communion). In short, Christians might use the gifts of science for proper ends. The Lambeth Conference also joined the question of birth control to larger social and global population questions, particularly in developing countries; but it did not impose legislation on individuals, asserting that the family does not exist for the sake of society.

Methodists were early acceptors of birth control, and the Methodist Conference in 1939 considered the aim of contraception to be "the healthiest family in the healthiest possible way." In 1931 the U.S. Federal Council of Churches ruled that careful and restrained use of contraceptives by married people was valid and moral, and in 1961 the National Council of Churches in the United States affirmed that "The general Protestant conviction is that motives, rather than methods, form the primary moral issue, provided the methods are limited to the prevention of contraception" (National Council of Churches policy declaration, 1961). Eventually the World Council of Churches, through the Commission of the Churches on International Affairs, committed itself to a policy of cooperation with the United Nations in demographic

The modern Protestant position

tasks; in 1966 it stated that responsible parenthood must be accepted as an integral part of the ethic of today.

Islām. Islāmic marital and sexual morality has roots in both Judaism and Christianity. While the Qur'ānic advocation is to marry and generate, there is no clear objection to birth control; and because common sense may be applied where there is no Qur'ānic prohibition, liberal interpretation has been possible. Birth control was early advocated by the Prophet's friend Hazrat ibn As, in order to avoid poverty. Though there was disagreement among the disciples of the Prophet, permission for coitus interruptus has been attributed to the scholars among them and to subsequent scholars. A medieval theologian, al-Ghazālī, said that contraception by means of withdrawal was justified to protect one's property, to preserve the health and beauty of one's wife, and to preclude worries about overlarge families. Muslim doctors like ar-Rāzī and Avicenna in the 10th century described many contraceptives and offered medical reasons for avoiding childbirth. A more recent Islāmic *fatwā* (pronouncement) was made by the grand mufti of Egypt in 1937: "It is permissible for either husband or wife, by mutual consent, to take any measures to prevent semen entering the uterus, in order to prevent conception." In 1960 the Turkish government obtained a *fatwā* permitting coitus interruptus if the proper raising of children was made impossible by social conditions. Therapeutic abortion was accepted on the basis of the Qur'ānic precept that "necessities overrule prohibitions." There is, however, no clear consensus: although a *fatwā* in Jordan in 1964 ruled that a government birth control policy was binding on individuals, other authorities resisted limitation of family size and legislative action.

Hinduism and Buddhism. Among Hindus, the begetting of a son is regarded as a primary religious obligation —both to continue the family and to facilitate the salvation of one's father and forebears. According to doctrine, women were created to bring children into the world, and marriage ceremonies stress the obligation to produce many sons.

In Buddhism, religious doctrine does not directly stress procreation, since the essence of this religion is to break away from worldly passions in pursuit of serene detachment. The tradition of high fertility among Buddhists appears to be supported mainly by cultural mores rather than religious doctrine.

On the control of fertility, Hinduism and Buddhism both lack either specific prohibitions or obligations in the modern sense. Doctrinal opposition to bodily injury has been a religious deterrent to infanticide and abortion. And the general fatalism of both faiths, as embraced in the various concepts of *karma*, has tended to restrain direct interference with fertility.

At the same time, there has been enough doctrinal latitude in Eastern religions so that modern spokesmen have not found it too difficult to construct a theological foundation for the use of contraception as a means of family planning. As the Indian philosopher and statesman S. Radhakrishnan expressed it to the Third International Conference on Planned Parenthood at Bombay in 1952, "If the purpose is not wrong, there is no ethical or spiritual harm done, and it is the purpose which determines the use or abuse of these modern inventions."

Effects of economic conditions and social status

Social influences. In modern societies providing economic opportunity and social mobility there are far more incentives to birth control. Men have ambitions for social and economic position, higher living standards, and education, and parents desire that their children should fare better than they themselves in this world. Such individual fulfillment probably has a better chance of success if income is concentrated on fewer children. In poorer societies, on the other hand, or in poorer segments of affluent societies, men generally lack much confidence that they can alter their adverse circumstances, or they fear disappointment; in any case the motivation for contraception is low. Furthermore, birth control is not highly valued when infant mortality is high, when there is dependence on family or child labour, or when aggrandizement is achieved by marriage alliances or the socioeconomic distribution of sons and daughters.

For individuals there may be a variety of reasons for avoiding or neglecting contraception. Women in traditional or impoverished cultures are often pressured to believe that bearing children is their raison d'être, and the males, feeling fatalistic or ineffectual in relation to the world, tend to believe that fathering a string of children is a demonstration of manliness and power. Furthermore, men who desire submissive wives can fear that contraception tends to liberate women. Conversely, men are more likely to accept birth control when they see their husbandly role as good friends rather than good lovers and when women's status in the family and society is relatively high.

The problem of usable methods

General motivation is not the only determinant of birth control, however. Method is also important. Since contraception is essentially a voluntary procedure, the individual or couple must be motivated to use the method. Thus, for instance, the absence of privacy in most homes, and the lack of medical guidance, make techniques such as the diaphragm unsuitable in preindustrial societies. The expense of condoms usually puts them beyond the reach of the poor. And while the rhythm method involves no "devices," it does involve relatively sophisticated use of calendars and measurements of time and temperature that are alien to many cultures and difficult to introduce to barely literate populations. The newer intrauterine contraceptives (IUD's), however, show substantial promise in helping curb population growth, because their use requires only a single insertion for longtime protection against pregnancy and requires only momentary medical attention and occasional checkups.

With increased investment in contraceptive research, the perfection of other simple new methods is anticipated. Meanwhile, educational efforts under government auspices are seeking to increase motivation for fertility control in a growing number of nations. (See also POPULATION.)

BIBLIOGRAPHY

Historical writings: THOMAS MALTHUS, *An Essay on Principles of Population* (1798, reprinted 1958), the classic work that prompted the social crusade; F.H.A. MICKLEWRIGHT, "Rise and Decline of English Neo-Malthusianism," *Population Studies,* vol. 15, no. 1 (1961), a short history of the Neo-Malthusians. Autobiographical and biographical works on pioneers include MARIE STOPES, *Early Days of Birth Control* (1923); MARGARET SANGER, *My Fight for Birth Control* (1932); K. BRIANT, *Marie Stopes: A Biography* (1962); and P. FRYER, *The Birth-Controllers* (1965).

Methods: N. HIMES, *Medical History of Contraception* (1963), the classic and still only comprehensive history, with an up-to-date preface by A. Guttmacher; J. PEEL and D.M. POTTS, *Textbook of Contraceptive Practice* (1969), a medical textbook with short expositions of social, legal, and religious considerations.

Religious beliefs: M.A.C. WARREN *et al., The Family in Contemporary Society* (1958), the Anglican viewpoint; R.M. FAGLEY, *The Population Explosion and Christian Responsibility* (1960), a review of the position of the major religions by a Protestant clergyman; J.T. NOONAN, *Contraception: A History of Its Treatment by the Catholic Theologians and Canonists* (1965), the major history of Roman Catholic doctrine; P. HARRIS *et al., On Human Life* (1968), an examination of the papal encyclical *Humanae Vitae.*

Social factors: E. DRAPER, *Birth Control in the Modern World* (1965), a short review of all facets of the birth control–population problem affecting individual decision; L. RAINWATER and K.K. WEINSTEIN, *And the Poor Get Children* (1967), a discussion of what motivates the less privileged in the U.S.; M.G. SCHOFIELD *et al., The Sexual Behaviour of Young People* (1965), attitudes and practices of the adolescent.

Family planning: P.K. WHELPTON *et al., Fertility and Family Planning in the US* (1966), the latest major U.S. demographic survey; B. BERELSON (comp.), *Family Planning Programs: An International Survey* (1969), and S.J. BEHRMAN *et al.* (eds.), *Fertility and Family Planning: A World View* (1969), reports of international conferences on fertility trends and population control. A great range of continuing discussions and research reports on many aspects of family plan-

ning methods may be found in *Studies in Family Planning*, a quarterly of the Population Council of New York.

(E.E.N.D./Ed.)

Birth Defects and Congenital Disorders

The term birth defect includes both abnormalities of structure and deficiencies of function that are present at the time of birth. These defects may be inherited (genetic), or they may be the result of an accident encountered by the child while in the mother's womb.

Congenital disorders include birth defects but also include diseases, such as syphilis, that can attack the child before birth and damage structures already formed. Among infections possibly only rubella (German measles) affects the embryo during its period of development, producing true malformations.

When a defect develops

The unborn human infant acquires its essential form, and its organs and tissues are all laid down and defined, within the first eight weeks after conception. This is the period during which the child is often, but not always, described as an embryo; only afterward is it called a fetus. After the first eight weeks there is differentiation and growth, and some anomalies may arise, especially of the brain, eye, and inner ear; but all gross disturbances of form will already have occurred. The distinction between embryo and fetus is useful if it emphasizes the early period of development critical in the production of congenital defects. Many biochemical defects may not be manifest until the metabolism is separated at birth from that of the mother; all such defects, however, have their inception in an earlier failure to develop some enzyme system.

CAUSES AND INCIDENCE OF CONGENITAL DEFECTS

Causes. Grossly malformed children have at different times and in different cultures inspired either awe or revulsion. They have been regarded as the playthings of the gods, and some gods have been modelled on human or animal malformations. The defects have been regarded as signs and portents or as punishment for sin. The ancient belief that they are produced by maternal emotional impressions or shock still lingers today. In its more absurd form it is expressed by such things as the belief that if a mother is frightened by a frog or a rabbit, the child in consequence will lack the top of its head or may have a harelip. There is no evidence in favour of such beliefs. Nor is there evidence that maternal emotional stress or anxiety contributes in any nonspecific way to the production of malformations.

It must be emphasized that birth defects do not all have the same basis, and it is even possible for apparently identical defects in different individuals to be due to different causes. Though the basis of most defects is still uncertain, almost all are due to genetic factors, environmental influences, or a combination of these two. Genetic factors must include not only inherited familial defects but also spontaneous genetic mutations and chromosomal anomalies arising during division of the cells.

Genetic factors

Some birth defects can be recognized as the direct outcome of Mendelian dominant or recessive inheritance. Two genes at the same location—these genes are called alleles—on a set of two chromosomes—homologous chromosomes—represent alternative characteristics, such as yellowness or greenness of peas. In dominant inheritance only one of the alleles need be for the characteristic; in recessive inheritance the pair of alleles must be identical: that is, if one of the two genes for colour of peas is for yellow and one is for green, the yellow is inherited; if both are for green—a recessive trait—greenness is inherited. Such direct inheritance may be modified by the environment; the manifestation of most of these defects may be modified in such a way that the pattern of Mendelian inheritance is obscured—some defects may be difficult to recognize in individual cases, and large and detailed pedigrees may be necessary to determine patterns of their recurrence. A high incidence of defects occurring in the offspring of cousin marriages may point to recessive inheritance of conditions rare in the community. The skeletal anomalies of achondroplasia (abnormal conversion of cartilage into bone with resultant dwarfism) and osteogenesis imperfecta (a disease marked by fragility of the bones) exemplify defects with dominant inheritance. Albinism (absence of pigment in skin, hair, and eyes), microcephaly (possession of an abnormally small or imperfectly developed brain), and many inborn errors of metabolism are determined by recessive inheritance.

Such a relatively simple pattern of inheritance fails to explain many common anomalies, such as harelip (with or without cleft palate) and congenital dislocation of the hip, and yet these conditions do show some tendency to recur in families. It is thought possible that several sets of alleles—polygenic inheritance—may be involved, in which each defect would represent the summation of the operation of these several independent sets of alleles; and only when these come together, as they are more likely to do in offspring of kindred, does the defect appear. This pattern of inheritance is difficult to prove, but from mathematical study of many families, supporting evidence is accumulating. When the pattern operates, the occurrence of two defective children increases the probability of a third, in contrast to the constant ratio of risk when only one set of alleles is involved. A severe form of any anomaly, or an anomaly present in the sex in which it does not usually occur, carries a higher risk to members of the family. The relative risk of the anomaly within a kindred is proportionately greater for those anomalies that are rare in the community. In many other anomalies, such as heart defects, complete obstruction or narrowing of the gut, defects of the brain and of the kidneys, bladder, and urinary passages, there is little or no increased incidence in particular families, and polygenic inheritance appears an inadequate explanation.

In chromosomal defects, there is defective sharing of the chromosomes among the cells, usually during division of the sex cells; this results in a gross genetic error. Such gross genetic disturbances probably account for 25 to 30 percent of early abortions; but few affected offspring apart from those with Down's syndrome (also called mongolism) survive to show birth defects. In inherited conditions, such as achondroplasia and the blood disease hemophilia, the death rate before reproduction is so high that the gene concerned would be eliminated if it were not renewed by spontaneous mutation. In achondroplasia only one out of five individuals receives the dominant gene from the family. In the other four it appears as a spontaneous mutation. If a mutant, dominant gene produces a condition so severe that the affected individuals fail to reproduce, the condition cannot be recognized. This is probably a rare occurrence, but it cannot be proven or disproven. By increasing the mutation rate, nuclear and other high-energy radiation can increase the incidence of congenital defects.

Environmental factors

The association of fetal anomalies with maternal rubella (German measles) infections and more recently with thalidomide has focussed attention on the possible effects of environmental factors. These had been extensively studied in free-swimming embryos, but the mammalian embryo was thought to be protected within the uterus. The incidence of fetal malformations due to maternal German measles varies in different epidemics and with the stage of the pregnancy at which the infection occurs. The greatest number of defects result when the infection occurs between the fifth and eighth week following the last menstrual period, but even then only rarely do as many as 20 percent of the infections cause defects. Even in epidemic periods the disease can be the cause of only 2 or 3 percent of all congenital defects. Such defects as deafness and cataracts result from infections during the fetal period; in earlier infections cardiovascular anomalies predominate. Thalidomide tends to involve the arms and legs. These are shortened, intermediate segments are missing, and the hands and feet are deformed. The embryo exposed to radiation may suffer damage to cells most actively dividing at that time, but this is very rarely the cause of human malformations. Despite observations

in experimental animals, it is doubtful that shortage of oxygen or nutrient substance produces defects in man. When an abnormality is produced in an animal by drugs or other agents, the incidence and nature of the abnormality are often dependent on the species and strain of animal used; and the effect of the agent may be to cause a pre-existing and latent genetic defect to become manifest rather than to act directly on the embryo and cause the defect.

The embryo floats cushioned in amniotic fluid and shielded in the maternal uterus, and it is extremely doubtful that physical injury ever determines a congenital defect. If the amniotic fluid is scanty near the end of the pregnancy, the fetus may be molded by continued pressure within or upon the uterus. Ears may be distorted or limbs bowed, but these defects of later fetal life often correct themselves.

A genetic factor can only express itself in a suitable environment. Much of the environment of the developing embryo is dependent on the operation of other genetic factors, and this is in part the basis of polygenic inheritance. Unknown and presumably chemical influences from the mother do, however, operate. Variations in the implantation and subsequent development of the placenta must also be important. Even when genetic factors are established, the genetically identical eggs of "identical" (monozygotic) twins may produce one affected and one unaffected child. This conditioning of the genetic inheritance by the environment is of the greatest importance. Its future understanding provides the best hope of limiting the occurrence of some abnormalities, in some cases by preventing the unmasking of harmful genetic factors by drugs, toxic agents, and maternal deficiencies. Unfortunately little is yet known of this interaction.

Interplay of environmental and genetic factors

Incidence. The incidence of birth defects depends on what is classified as a defect. About 20 percent of all babies born dead, or dying in the first week of life, die because of some serious malformation. This is about six per thousand births. Significant malformations will usually be detectable in the first two weeks of life in just under 20 per thousand total births. By the age of five years, another five or six per thousand will have been recognized. When serious metabolic errors such as fibrocystic disease of the pancreas and defects undetected or appearing later are added, a figure of 30 per thousand births is probably a conservative estimate for significant birth defects. If all classifiable structural anomalies found on careful dissection are included, nearly half the population will show some anomaly and, if all local tissue malformations such as skin blemishes and moles are included, there will be few unaffected.

Some conditions, such as anencephaly (the absence of all or most of the brain), account for up to 4 percent of stillbirths and early neonatal deaths in Ireland and western Britain but may have an incidence as low as one-twentieth of this elsewhere. There is wide variation in the incidence of different defects in different racial groups. It is much less certain that the total incidence of congenital defects varies significantly.

Offspring with severe malformations do not survive the embryonic period, and all but a few chromosomal defects are eliminated as early abortions. Most defects represent a local failure of growth often allied to a failure of differentiation. This can include a failure to canalize a passage or failure of a septum (thin dividing wall) or ridge to fuse. When there is imbalance in the constituents present in an area, the part may be enlarged, and the useful functioning tissue reduced. Failure to differentiate and arrange cells properly is often also local, with an imbalance in their proportion in tissues and organs. Such areas of faulty development may range from skin nevi ("birthmarks") to totally malformed kidneys and disorganized brains. Cells not taking part in normal development may later proliferate, forming aggressive and malignant tumours (embryomas) or tumours, especially in testes and ovaries, that show a mixture of elements of the various embryonic layers (teratomas). Whether areas once formed degenerate or fail to grow because of disturbed blood supply or some degenerative disturbance is uncertain.

General nature of defects

DEFECTS OF THE DIFFERENT BODY SYSTEMS

Musculoskeletal systems. Achondroplasia and osteogenesis imperfecta are among the more important general disturbances. In achondroplasia there is an inherited defect in long bones that develop through cartilage, and the arms and legs are stunted. There are also deformities of the chest and skull. About 80 percent of infants with achondroplasia die soon after birth, but those that survive are strong and sturdy dwarfs with unimpaired intellect. In osteogenesis imperfecta (brittle bones) there is an inherited and obscure defect of bony tissue, and bones may be broken even before birth. Multiple fractures occur on the slightest injury until puberty, after which there is a progressive improvement.

A great variety of local anomalies occur. Congenital dislocation of the hip and the talipes equinovarus form of clubfoot, in which the foot is extended and turned inward, show strong familial incidence. Fusion and webbing together of the digits of the hands and feet (syndactyly) or the presence of extra digits (polydactyly) are often inherited but are often variable in their form even in the same kindred. Imperfect extremities include hemimelia, in which hands and feet are imperfect or deformed, and phocomelia, in which defects in the long bones result in attachment of hands and feet, often also imperfectly developed, close to the body like the flippers of a seal. Apart from the numerous cases resulting from thalidomide, phocomelia is rare, and there is considerable detailed variation.

Cervical ribs, or additional tiny ribs arising from the lower neck vertebrae, are common anomalies that are unrecognized except in a few individuals in later life, when, probably because of changes in posture, the cervical ribs may press on nerves going to the arm and produce pain and other symptoms. Other anomalies may be complex; thus in cleidocranial dysostosis—"imperfect development of collarbones and skull"—the clavicles, or collarbones, are absent or rudimentary, there are complex deformities of the skull, and teeth are absent or poorly developed.

Central nervous system. A most important defect involves the spine and allows protrusion of the membranes of the spinal cord in the midline of the back (meningocele) or, worse, exposes the spinal cord itself in the floor of the defect (myelomeningocele or myelocele). The defect may be associated with anencephaly. The defect of the cord sometimes extends the whole length of the vertebral column. When nerves and especially when cord structures protrude, attempts at repair nearly always result in some loss of control of the lower part of the body. Often, when only the cord seems to be involved and for reasons that are still debatable, the ventricles, or cavities, normally present within the brain dilate (hydrocephalus). The head enlarges, and, with thinning of the cerebral cortex (outer layer of the brain), there is variable impairment of brain function. Congenital hydrocephalus may also result from a defect that significantly narrows the passage between the ventricles.

Defects in spinal cord and brain

Many complex anomalies of the brain can occur, often with deformities of the folds or convolutions (gyri) of its surface. Small and large gyri (microgyria and macrogyria) occur. They may be absent (agyria), reduced (oligogyria), or increased (polygyria) in number, and multiple areas of imbalance between nerve cells and supporting elements (tuberous sclerosis) may occur. Sometimes there is an overall reduction in brain size (microcephaly), often associated with a disturbed gyral pattern; this is usually inherited as a Mendelian recessive. Such anomalies will be found in only a minority of those with mental defects and with spastic and other disturbances; debate continues on what proportion of these defects result from less evident developmental errors and how many from damage done by lack of oxygen and by other deficiencies or by physical injury at birth.

Sense organs. Congenital deafness may be due to a maternal infection with German measles (rubella) or to

inheritance. Both rubella and genetic factors can also produce cataracts. Small eyes (microphthalmia) and absence of eyes (anophthalmia) may be associated with grave brain defects but can exist alone. Injury by rubella and infection with the protozoan *Toxoplasma gondii* within the uterus may produce small eyes. The organs of smell and taste are very rarely disturbed and usually only with associated complex brain anomalies. Congenital indifference to pain must depend on a rare defect at a high level in the brain.

Digestive tract. Anomalies of the digestive tract occur at all levels. Harelip, especially when combined with cleft palate, creates a serious feeding difficulty in early infancy; it is due to a failure of structures that are growing forward to form the nose and mouth to unite. A break in the continuity of the esophagus (esophageal atresia) is almost always further complicated by an opening between the esophagus and the trachea. Spasm and a thickened ring of muscle at the exit from the stomach (hypertrophic pyloric stenosis), appearing and progressing in the first month after birth, may cause persistent vomiting. It is five times more common in boys than in girls and tends to occur in more than one member of a family. The incidence may be as great as five per thousand boys.

Defects in the bowel

A segment of intestine, usually near the anus, may lack nerve cells and networks concerned with peristaltic movements and may act as an obstruction, leading to intermittent but progressive constipation and gross dilation of the intestine above the defect (Hirschsprung's disease). As in hypertrophic pyloric stenosis, there is a high familial incidence. A stenosis or area of structural narrowing may occur anywhere in the alimentary canal; the passageway, or lumen, may be totally occluded. Parts of the bowel may be duplicated, usually forming cystlike structures leading to obstruction. Atresia (lack of an opening to the outside) of the anus is often due only to a thin membrane, but fistulas or narrow passages from the obstructed bowel may open into the urethra, the vagina, and the area between the vagina and the anus. Less often the rectum ends at a higher level. Here again there may be fistulas and even wider openings into the urinary tract; rarely, the arrangement may approach the common excretory cloaca of lower animals.

In cystic fibrosis of the pancreas there is an abnormality of mucus secretion affecting the whole body in varying degrees. In intrauterine life, abnormal secretions in the bowel and the lack of pancreatic enzymes to digest protein may result in a mass of material in the bowel that distends and blocks it (meconium ileus). In infancy there can be deficiency of digestive secretions and, often, disturbed fat absorption. Thick viscid secretions in the bronchi later lead to emphysema (overdistention of the lungs because of dilation of the air sacs) and repeated lung infections. This occurs in about one in a thousand births and is inherited as a Mendelian recessive.

Cardiovascular system. A multitude of anomalies exist, and some are discussed elsewhere. The essential consideration is whether or not abnormal openings between cavities exist. Such shunts may allow oxygenated blood from the lungs to mix with venous blood from the body. This occurs in small defects of the septa, or dividing walls, of the heart, and in a ductus arteriosus (shunt between the pulmonary artery and the aorta) that fails to close. More serious are anomalies in which blood returning from the body mixes with oxygenated blood from the lungs, producing cyanosis (blueness of skin due to inadequate oxygen supply). Extreme cyanosis, or blue baby disease, occurs in the tetralogy of Fallot—an interrelated set of defects in which there is a wider defect of the septum and obstruction to the outlet from the right ventricle to the lungs—and in other complex anomalies. In less serious imperfections, however, there may be no admixture of blood; sometimes, as in transposition of the heart to the right side, which may or may not be associated with a similar transposition of the abdominal viscera, anomalies cause no circulatory impairment.

Blood. Defects of the blood might be used to illustrate inherited or inborn metabolic anomalies. The life and shape of the red blood cell depend on enzyme systems, which may be defective. The hemoglobin that the cell carries may exist in numerous variant forms. These variants are usually less useful and even harmful, but some have probably had survival value. Thus the hemoglobin of the sickle cell trait and sickle cell anemia increases resistance to malaria. Many defects are so widespread in populations that they can scarcely be regarded as abnormal.

Excretory system. Absence of one kidney may occur in one in a thousand. There is enlargement of the other, and the condition is usually unsuspected. There is a high association with conditions such as, in females, a bicornuate uterus (a uterus with two partially united compartments) and vaginal atresia (lack of a channel in the vagina) and, in males, with absence of a testicle. Absence of both kidneys is fatal soon after birth and is accompanied by defects in the sexual organs and defective lung development. Disorganized development of one or both kidneys is often associated with renal (kidney) cysts and abnormal structures. The poorly developed kidneys with cysts in the newborn infant differ from the inherited cystic kidneys that develop in later life. The ureters and the bladder itself may be dilated, either behind an obstruction (for which congenital folds of mucous membrane in the urethra may be responsible) or without visible obstruction—the so-called megalo-ureter. Gross distention of the urinary passages is sometimes associated with the absence of abdominal muscles. In all these conditions there is a serious risk of kidney infection.

Kidney defects

In ectopia vesicae (displacement of the bladder), a serious defect, the bladder lies open on the lower part of the abdominal wall with the ureters discharging there. This urinary defect presents a long-term problem in surgical management. In the male a lesser degree of the anomaly is epispadias, in which the urethra opens on the abdominal side of the penis rather than at the tip; bladder control is absent. Openings on the opposite side of the penis, in hypospadias, are entirely different, since bladder control is retained.

Reproductive system. The anomalies of hypospadias and epispadias affect the reproductive system as well as the excretory system. In the female a lack of an opening in the vagina or in the cervix will allow mucus from the cervix (hydrocolpos) and later blood from the uterus (hematocolpos) to accumulate and progressively enlarge the uterus.

In pseudohermaphrodites the sex organs are so developed that they appear not to correspond to the gonads (testes or ovaries) present. In female infants a congenital defect, manifest before birth and inherited as a recessive, disturbs the production of adrenal hydrocortisone and results in overproduction of the male sex hormone. In consequence the clitoris enlarges and at birth may resemble a penis with hypospadias. The vagina appears to open into the urethra and is not visible, and the labial folds resemble the scrotum. The condition is progressive after birth, but both it and the associated salt imbalance respond dramatically to treatment with cortisone. Somewhat similar sex organs may be produced by steroids, often progesterone, given for threatened abortion but also exerting a male hormonal influence.

Pseudohermaphrodites and true hermaphrodites

Persons with a wide range of abnormalities have been described as male pseudohermaphrodites. The group includes persons with hypospadias and undescended testes, those with a clitoris and a vagina ending in a mass connected to testes, and persons with an imperfect uterus, a single fallopian tube, and no sex gland on that side. The correct management of all these anomalies demands most expert care.

In a true hermaphrodite both ovarian and testicular tissues are present on either one or both sides. The arrangement of other genital structures is extremely variable in the hundred or so reported cases, and the whole problem is extremely complex.

Respiratory system. Cleft palate, an excessively small lower jaw, and esophageal atresia (imperfectly developed

passageway in the esophagus) with an opening between esophagus and trachea all cause respiratory difficulty. A defect in the diaphragm allowing abdominal contents to enter the chest is an important cause of failure of respiration in the newborn. Anomalies of the lungs range from absence of one or both lungs to insignificant variation in the size and shape of the lobes.

Endocrine system. Anomalies of the ductless glands are rare. Deficiency of the thyroid will, soon after birth, produce signs of cretinism with slow metabolism, coarse skin, and mental retardation. Overdevelopment of the adrenal cortex affects metabolism of the cortical hormones with disturbance of genital development and often of salt balance.

Skin. In albinism there is inability to produce melanin pigment, and the resulting white skin and hair and sometimes red eyes and intolerance to light are due to a metabolic defect inherited as a Mendelian recessive. It occurs in both white and coloured races and varies greatly in degree of severity. Many rare skin diseases that well illustrate Mendelian inheritance have been studied. The dry, scaly skin of ichthyosis occurs in different forms, illustrating different patterns of inheritance. Birthmarks, such as port wine stains and strawberry nevi (small pigmented areas on the skin), are a local and abnormal development of blood vessels. These and pigmented moles are so common that they are often not regarded as malformations.

CONDITIONS ARISING FROM CHROMOSOMAL DEFECTS

During division of the sex cells the number of chromosomes is halved, one of each pair going to each cell. It sometimes happens that both chromosomes of a pair go to one sex cell (nondisjunction). When two sex cells unite, there can then be not two paired chromosomes but three. The cell is then called trisomic. The unpaired chromosome may be fused with another chromosome and can then be carried through to succeeding generations. When neither chromosome of a pair goes to a germ cell, the zygote or fertilized egg is short of a chromosome and has one chromosome unpaired. This is called a monosomic anomaly. More complex anomalies occur but disrupt development so early that abortion occurs. This happens also with trisomies of all but a few small chromosomes and with all monosomies except those involving the sex chromosomes.

Monosomy and trisomy

Turner's syndrome. In a form of monosomy called Turner's syndrome, only one female sex chromosome (X-chromosome) is usually present, but more complex arrangements can occur. The external genitalia are feminine but remain infantile after puberty. The ovaries are rudimentary and do not contain eggs, body growth is impaired, there is often a distinctive webbing of the neck, and there is a characteristic facial appearance.

Klinefelter syndrome. In the Klinefelter syndrome the genitalia are normal and masculine before puberty. Later development is usually retarded, enlargement of the breast often occurs by the 20th year, and the incidence of mental retardation is increased. Sperm are absent from the testes. The number of female (X) chromosomes is increased, usually giving the pattern XXY. Other patterns —XXXY, XXXXY, and XXYY—which occur much less frequently—raise interesting problems as to the time at which separation of the chromosome pairs fails. The Klinefelter syndrome may have an incidence as high as one in 500 males.

Down's syndrome (mongolism). This is the only numerically important example of a trisomy involving autosomal (nonsex) chromosomes, and it has an incidence of one to two per thousand. Babies with this defect are small and have distinctive arrangement of skin folds around the eye, which is not unlike that of the Mongolian race; short fingers with incurved little finger; and diagnostically useful abnormalities in the skin creases of the palm of the hand. These children are retarded in physical and mental development and often have more serious anomalies such as congenital heart lesions and intestinal atresia (imperforate intestine). When the extra chromosome becomes fused with another chromosome, the anomaly may recur in other members of the family. If the mother of the child is young, chromosome studies of the parents are advisable. The defect is most common among mothers aged 35 or more; in these instances the defect is usually not familial.

BIBLIOGRAPHY. J.E. MORISON, *Foetal and Neonatal Pathology*, 3rd ed. (1970), descriptions of congential defects and other diseases, with illustrations and many references; A.P. NORMAN (ed.), *Congenital Abnormalities in Infancy*, 2nd ed. (1971), good descriptions with illustrations of both localized and generalized anomalies; E.L. POTTER, *Pathology of the Fetus and Infant*, 2nd ed. (1961), especially valuable for the many fine illustrations; J.A.F. ROBERTS, *An Introduction to Medical Genetics*, 5th ed. (1970), a useful account of the genetic factors in human anomalies and disease; A. RUBIN (ed.), *Handbook of Congenital Malformations* (1967), systemic descriptions of anomalies but not illustrated.

(J.E.M.)

Biscay, Bay of

Noted for its rough seas, squalls, and storms, the Bay of Biscay, a roughly triangular-shaped extension of the North Atlantic Ocean, is bounded on the east by the west coast of France south of Île d'Ouessant (Ushant, off the western tip of Brittany) and on the south by the north coast of Spain out to Cabo (Cape) Ortegal at the extreme northwest. The name is a corruption of the Spanish Golfo de Vizcaya (after the province of Vizcaya, around Bilbao), which is restricted to the angle between the two countries. The French name is Golfe de Gascogne. The sea is crisscrossed by important shipping routes and, since prehistoric times, has witnessed the movements of maritime people along the fringe of Atlantic Europe.

Its hypothetical boundary with the Atlantic extends about 357 miles from Île d'Ouessant to Cabo Ortegal. The area of the bay is about 86,000 square miles (223,000 square kilometres), and its maximum depth, a little south of its centre, is 15,525 feet.

Physiography. The continental shelf is up to about 100 miles (160 kilometres) wide off Brittany but narrows to less than 40 miles off the Spanish shore. The edge of the shelf and the continental slope are dissected by numerous submarine canyons of which that of Cap Breton, in the southeastern corner of the bay, is one of the largest. The bottom sediments are largely sand and mud.

Beyond the continental slope lies the Biscay Abyssal Plain, with depths of about 15,000 feet, which occupies about half the area of the bay. Much of it has very flat topography. The principal rivers flowing into the bay are the Loire, the Adour, and the Dordogne and Garonne, which form the Gironde Estuary.

Geological history

The continental shelf and slope and the abyssal plain are underlain by thick sediments of Mesozoic and Tertiary age (*i.e.*, formed between about 225,000,000 and 2,500,000 years ago) that appear to form a seaward continuation of rocks of similar age in the Aquitaine Basin of southwestern France. The shelves are believed to have been formed by sedimentation keeping pace with subsidence, and the French shelf has subsided by about 10,000 feet since Lower Cretaceous time (about 136,000,000 years ago). The shelf areas are underlain by continental crust, but the abyssal area is underlain by thinner oceanic crust, the boundary between the two being a faulted zone beneath the continental slope. It has been suggested that the abyssal part of the bay was formed by the anticlockwise rotation (drift) of Spain relative to France, leaving a triangular gap in which oceanic crust was formed as the gap opened. The hypothesis has received some support from paleomagnetic data and has been popular with workers attempting to reconstruct the geometric fit of the continents prior to the formation of the Atlantic Ocean. Since the hypothesis envisages the beginning of the rotation in the Late Cretaceous, however, it is rendered more difficult to accept by the probability that Mesozoic and Tertiary sediments exist beneath the abyssal plain and by the proved existence of Early Mesozoic and earlier rocks beneath Aquitaine. The rotation, if it occurred, must have been of much earlier date.

Currents, tides, and salinity. Surface currents are influenced by the clockwise circulation in the North Atlantic that produces a clockwise circulation in the bay. This pattern is frequently disturbed by wind, so that, on the shipping route across the mouth of the bay, currents in any direction may be encountered. Westerly gales in winter give rise to an eastward-flowing current along the north coast of Spain, which reaches five miles per hour. The main outflow of water is believed to be a subsurface current flowing westward off the north coast of Spain.

The range of mean spring tides is about 20 feet on the French coast at the northern end of the bay near Île d'Ouessant, decreasing southward to about 12 feet in the southeastern angle near Biarritz. Strong westerly winds increase the normal height of tide by several feet. Easterly winds depress the tidal levels.

Salinity of surface waters is about 35 parts per thousand, slightly higher than average ocean water. This is the result of the northward spread of hypersaline water leaving the Mediterranean through the Strait of Gibraltar.

Climate. The Bay of Biscay is noted among sailors for rough seas. The frequency of gales is in fact less than in more northerly parts of the eastern Atlantic, but they can be severe and may exceed 70 miles per hour. Squalls are also a hazard to navigation and may occur at any time of year. The north coast of Spain is noteworthy for dangerous squalls and capricious winds. The climate on shore is maritime, with mild winters and cool summers.

Principal ports

Economic activity. The principal ports are Brest, Nantes, La Rochelle, Bordeaux, and Bayonne in France, and Bilbao, Santander, Gijón, and Avilés in Spain; none are able to take large vessels. Resorts include La Baule, Biarritz, and Saint-Jean-de-Luz, all on the French coast. Fishing is a principal industry. Catches include sardines, tuna, anchovies, hake, cod, bream, and crustaceans, including lobster. Oyster culture is practiced in shallow lagoons and estuaries along the French coast. Offshore drilling has taken place off the Aquitaine coast in the hope of finding possible future sources of petroleum and natural gas.

BIBLIOGRAPHY. Recent geological and geophysical work is described by L. MONTADERT *et al.*, "The Continental Margin in the Bay of Biscay," in *Report of the Institute of Geological Sciences*, no. 70/15, pp. 43–74 (1971). Sediments and recent history of the French continental shelf are discussed by G. BOILLOT *et al.*, "Morphology, Sediments and Quaternary History of the Continental Shelf between the Straits of Dover and Cape Finisterre," *ibid.*, pp. 75–90. There is no comprehensive work on oceanography and climate, but the general background is given by R.W. FAIRBRIDGE, ARNOLD GORDON, and ERIC OLAUSSEN, "Atlantic Ocean," in R.W. FAIRBRIDGE (ed.), *The Encyclopedia of Oceanography*, pp. 56–85 (1966); and some information on tides, currents, and climate in GREAT BRITAIN, HYDROGRAPHIC OFFICE, *Bay of Biscay Pilot*, 4th ed. (1956).

(D.T.D.)

Bismarck, Otto von

Otto von Bismarck, the Prussian statesman who founded and became first chancellor of the German Empire, was born on April 1, 1815, at Schönhausen in Brandenburg. His father, Ferdinand von Bismarck-Schönhausen, was a Junker landowner who had served in the Prussian Army; his mother, Wilhelmine Mencken, was the daughter of an untitled bureaucrat who had risen high under Frederick William III. The contrast between his parents did much to cause the conflict in Bismarck's own nature. In physical appearance and avowed tastes he was a Junker: powerfully built, devoted to country pursuits, and with an enormous appetite for food and drink. Mentally and emotionally, however, he was sophisticated and highly bred, sensitive to the point of hysteria, with a subtle intellect and a gift of expression that put him in the highest class of German writers.

LIFE

Early years. Bismarck went to school in Berlin, living with his mother and developing a romantic nostalgia for the family estates, which he seldom visited. He read law first at Göttingen University and then at Berlin; he was



Bismarck. 1880.
Achiv fur Kunst und Geschichte

a disorderly student, though he passed his examination. He then entered the Prussian service and became a judicial administrator at Aachen. There he was drawn into the international society that stopped at the spa on the way to other watering places; and at one time absented himself without leave for some months in order to follow across Germany an English girl with whom he had fallen in love. The adventure came to nothing and the criticism of his official superiors drove Bismarck to resign from the service at the age of 24. He declared, "I will play music as I like it or none at all." The family estates were running into difficulties, and Bismarck took over the management of them from his father and to some extent from his elder brother.

Under his mother's influence, Bismarck as a youth had been a deist, indifferent to Christianity. In the country, he came to know his Pietist neighbour Adolf von Thadden and was deeply affected by the death of Thadden's daughter Marie. Soon after he fell in love with another girl of this Pietist circle, Johanna von Puttkamer, whom he married in July 1847. To win her hand Bismarck avowed his own conversion, though he was inclined to lay down to God the conditions on which he would acknowledge his existence.

Early career. In 1847 Bismarck became a member of the United Diet, the quasi-representative body that Frederick William IV summoned in order to authorize a loan for the building of a railway to East Prussia. Bismarck stood out as the provocative spokesman of absolutism and reaction, dismissing the mildest liberal measures with contempt. The Diet demanded that it should be summoned regularly; when the King refused this it rejected the railway loan and was prorogued. Before it could meet again the Revolution of 1848 swept Germany, and Frederick William IV himself capitulated to the revolution in Berlin on March 18. Bismarck wished to organize a military repression and civil war. When the King pleaded that he had been unable to sleep for worry, Bismarck replied, "A king must be able to sleep." He attempted to force the King's abdication and suggested to Princess Augusta, wife of the King's brother William, that her husband, too, should withdraw in favour of his young son, who would then become the helpless figurehead of the reaction. Augusta indignantly refused and remained the mortal enemy of Bismarck throughout her life. In later years he launched the version that Augusta had made the proposal to him and that he had rejected it out of loyalty to the King—one of many instances in which he put the best appearance on failure and sought to mislead posterity.

The 1848 revolution

Bismarck was in eclipse so long as liberalism prevailed in Germany. He was elected neither to the National Assembly at Frankfurt nor to the first Prussian Parliament at Berlin. In October the army occupied Berlin, and Frederick William IV issued a more monarchist constitution.

Bismarck was elected to the new Second Chamber. In April 1849 the Frankfurt National Assembly offered the imperial crown to Frederick William. He refused it; and Bismarck was one of the few who welcomed this refusal. Frederick William next attempted to gain the leadership of northern Germany with the cooperation of the princes. This challenge to Austria was opposed by Bismarck, and when open conflict seemed to be approaching in October 1850, he advocated return to Austro-Prussian cooperation in German affairs. He welcomed enthusiastically the Prussian surrender at Olmütz (now Olomouc, Czechoslovakia), when Frederick William's schemes were abandoned, and praised Austria as "a German power that is fortunate enough to rule over foreign peoples." He hoped that, with the defeat of the revolution, Prince Metternich's "system" with its Holy Alliance of Austria, Russia, and Prussia would be restored. These views made him an obvious choice to represent Prussia at the federal Diet at Frankfurt. He arrived there on May 11, 1851, renewing a career in the service of the state that was to last without interruption for 39 years.

Prussian representative at Frankfurt

Bismarck went to Frankfurt as the leading advocate of cooperation with Austria in German affairs. Within a fortnight, his outlook was revolutionized, and he became convinced that the Austrian statesmen would never treat Prussia as an equal. The men who had defeated the revolution in Austria no longer thought in Metternich's cautious terms. They believed that they could hold the revolution in check without any assistance from their former conservative partners Russia and Prussia. But if the challenge came from Austria, Bismarck was not slow to take it up. He disputed every formal sign of Austria's leadership, though Austria was in fact "the presiding power" at the Diet by virtue of the federal act; and he carried on a relentless personal feud with successive Austrian representatives. Hitherto only the Austrian had smoked at meetings. Now Bismarck, too, pulled out his cigar case. Such trivialities were to reshape the destinies of central Europe. Bismarck soon came to speak with contempt of Frederick William IV and of Otto von Manteuffel, his foreign minister. He believed that they were sacrificing Prussia's greatness to their romantic conception of conservative solidarity. In his opinion "the Holy Alliance was dead," and Prussia should follow a policy of self-centred realism. It should claim the headship of Germany and make alliances with foreign powers against Austria in order to achieve it; it should even ally itself with "the revolution."

During the Crimean War (1853–56) Bismarck feared that Prussia would support Austria or even join the Western powers in war against Russia. He reached the conviction that the Balkans and the lower Danube, however vital to Austria, were no concern of Prussia's, and this view did much to shape his later policy. At the end of the war he believed that Russia and France would soon come together in a "revisionist" alliance—Russia to undo the results of the Crimean War, France to expel Austria from Italy—and he urged that Prussia should make a third in this restless combination. He overrated the dynamism of both Russia and France and insisted that Prussia could gain the headship of Germany only as part of a general European upheaval. His greatest state paper dismissed as impossible the limited reordering of German affairs that he himself carried through 14 years later.

Bismarck grew increasingly critical of Prussia's official policy in his last years at Frankfurt. He was over 40 and seemed to have achieved nothing. In 1858 Frederick William IV became insane, and his brother William took his place as regent. This began the short-lived "new era," when Prussian policy took a liberal turn. Bismarck seemed out of place at Frankfurt with his reactionary reputation, and early in 1859 he was sent as ambassador to St. Petersburg—"put in cold storage" was his own phrase.

In January 1861 Frederick William IV died and the prince regent became King William I. A soldier by training and profession, he wished to increase the army in order to be able to train all those liable to military service. This increase would, in addition, enable him to cut down the *Landwehr*, or militia, which Prussian conservatives regarded as a dangerously democratic institution. The liberal majority in the Prussian Parliament was ready to increase the grant for the army, but they defended the *Landwehr* and sought to reduce the period of military service from three to two years. In 1860 Parliament authorized the increased grant for one year only. Albrecht von Roon, the minister of war, at once organized new regiments, implying that the grant was permanent. In November 1861 a general election strengthened the more advanced liberals, who were now organized as the German Progressive Party. Roon urged William to make Bismarck prime minister, but the King was alarmed by Bismarck's advocacy of alliance with France and his plans for sweeping away the German princes.

Ambassador at St. Petersburg and Paris

In March 1862 Bismarck was recalled from St. Petersburg. He expected to be made prime minister (minister president); but though William was running into increasing difficulties with the Chamber of Deputies, he still shrank from an open breach of the constitution, and Bismarck on his side would not take office until the King could be forced to accept a revolutionary foreign policy. He was therefore sent as ambassador to Paris. There he and Napoleon III dangled before each other the idea of alliance against Austria.

Appointment as prime minister. Meanwhile events in Berlin had reached a crisis. The Chamber of Deputies would authorize the increased military expenditure only if the period of service was reduced to two years. On September 17 the ministers, including Roon, agreed to this compromise; it was rejected by William, and Roon swung around to his support. Immediately after the meeting of the council Roon sent to Bismarck the famous telegram, "Periculum in mora. Dépêchez-vous" ("There is danger in delay. Hurry up"), meaning that the King's reluctance had been broken. On Sept. 22, 1862, Bismarck arrived in Berlin and was at once made prime minister. It was an uneasy partnership. William wished to defend his independent conduct of military affairs but still rejected an adventurous foreign policy; Bismarck encouraged the domestic conflict in order that the King should be dependent on him and then be drawn into a foreign policy by no means to his taste.

The constitution laid down that the budget must be agreed upon by the Chamber of Deputies, the upper house, and the king. Bismarck argued that if one of the three rejected the budget there was "a gap in the constitution," and that the government must collect the existing taxes (and of course spend them) until agreement was reached. This theory served to promote a conflict. He told the Chamber in his first speech, "The great questions of our day cannot be solved by speeches and majority votes—that was the great mistake of 1848 and 1849—but by blood and iron." Time and again the liberal majority offered a compromise, but Bismarck knew that this would destroy his hold over the King and always evaded it. Dissolution failed to shake the liberal majority; this, too, was an advantage for Bismarck, for a conservative Chamber would have made him unnecessary.

Bismarck's main interest lay elsewhere—in foreign policy. He had long announced his intention of settling the German question. When he took office a conflict had already started on Austrian initiative. The Austrians proposed to the federal Diet a delegate conference at Frankfurt to strengthen the confederation. Bismarck answered by proposing the direct election of a German Parliament, and he asked the French government what it would do "if things grew hot in Germany." Napoleon III was in a conservative mood and gave a timorous answer. In February 1863 the Diet rejected the Austrian proposal and the alarm blew over.

The Polish revolt

Russia and France had worked together with increasing intimacy since the end of the Crimean War. This intimacy was disturbed by a revolt in the part of Poland ruled by Russia, which broke out in January 1863. Napoleon III tried to turn his back on this embarrassing affair and pretended that it was a question of Russian domestic

politics. Bismarck, however, sent Count Gustav von Alvensleben to conclude a convention with Russia (Feb. 8, 1863), for cooperation against the Polish insurgents. Bismarck always regarded Polish national ambitions as the greatest threat to Prussia's existence, and his motive in proposing the Alvensleben Convention was principally to silence the advocates of concession to Poland at the Russian court. Certainly the convention did not win Russian gratitude: Prince A.M. Gorchakov, the chief minister, resented the implication that Russia was in any need of help. Moreover, Napoleon III, who was impotent to act against Russia, thought of making a demonstration in favour of Poland by threatening Prussia; and at the beginning of March, Bismarck had to ask Gorchakov to allow him to withdraw the convention. Far from being a great stroke of policy, the convention was an impulsive gesture against the Poles that led Bismarck into a position of some danger: he had almost put Prussia as a hostage between Russia and France.

Meeting of the German princes at Frankfurt

In August 1863 the emperor Francis Joseph I invited the German princes to a meeting at Frankfurt to discuss the reform of the confederation. It was the high-water mark of the Austrian attempt to unite Germany by agreement. Its principal element was a "directory" of princes. Bismarck might have been willing to compromise with Austria at the expense of the princes; he would certainly not give the princes authority over Prussia. William I was tempted to attend the Frankfurt meeting, especially when King John of Saxony brought an invitation in person. Bismarck threw all his weight on the other side. It was his first struggle for ascendancy over William, and he won it, though at great strain. When the argument was over, Bismarck smashed a large jug of water on the ground and broke into hysterical sobs. William's refusal wrecked the Frankfurt meeting. Though the princes approved the Austrian proposals, they added the proviso that Prussia must approve them too. The princes had, in fact, no interest in German unity and welcomed the conflict between Austria and Prussia as giving them an easy excuse.

In answer to the Austrian policy of working with the princes, Bismarck emphasized his friendship with Russia and with France. In the autumn of 1863 he concluded a commercial treaty with France and imposed this treaty on the Zollverein, the German customs union, which Prussia controlled. There was no room for Austria in this arrangement, and the economic division of Germany was thus made final, even while the political confederation still existed.

War over Schleswig-Holstein

In November 1863 Frederick VII, the last Danish king of the male line, died, and therewith the Schleswig-Holstein question, which had played a vital part in the German Revolution of 1848, was renewed. The two duchies, Schleswig and Holstein, were in union with Denmark; Holstein alone was a member of the German confederation, though there was also a German majority in southern Schleswig. In 1848 the German liberals had attempted to "liberate" the two duchies with the assistance of the Prussian Army but had been defeated by the protests of the great powers, and the Treaty of London (1852) had reaffirmed personal union. The Danes had subsequently tried to extend their constitution to include the two duchies, and German national sentiment countered by supporting a rival claimant to the duchies, the Duke of Augustenburg. Bismarck had no interest in creating another petty German state, particularly on Prussia's border; on the other hand, he would not let Austria steal a march on him by appealing to German national feeling.

Johann Bernhard von Rechberg-Rothenlöwen, the Austrian foreign minister, was a conservative who hated German liberalism and therefore fell in with Bismarck's plans. On Jan. 16, 1864, he signed a treaty of alliance for war against Denmark, not to liberate the duchies but to enforce personal union as laid down by the Treaty of London; German nationalism was repudiated. The Danes were soon defeated and Schleswig overrun. A conference of the great powers then met to try to save the settlement of 1852, but the coalition of non-German powers no longer existed. On May 28 Austria and Prussia repudiated the Treaty of London and were able to renew the war against Denmark without intervention or serious protest from any foreign power. In August Denmark made peace, ceding the duchies to Prussia and Austria jointly (the definitive treaty was signed at Vienna on Oct. 30, 1864). Later in August, Bismarck and William visited Francis Joseph at Schönbrunn. Rechberg was willing to surrender the duchies to Prussia but asked in return for a guarantee of Venetia and help in recovering Lombardy from Italy; Francis Joseph also wanted compensation in German territory. The terms were logical enough, but they were too high for Bismarck. He might compromise with Austria; he would hardly go to war with Italy, and perhaps France as well, for Austria's sake. Austria and Prussia therefore agreed to rule the duchies jointly until something turned up that would make a lasting solution possible.

Conflict with Austria

Bismarck still hoped to reach an agreement with the Austrians when their difficulties increased elsewhere—either in Italy or the Near East. Instead they concentrated on German affairs and put forward the claim of the duke of Augustenburg to the duchies. On May 29, 1865, a Prussian Crown Council discussed whether to meet the Austrian challenge by war. The generals favoured war; Bismarck insisted on alliance with France and an appeal to German nationalism if war was decided on. He said: "If war against Austria in alliance is struck out of the vocabulary of diplomacy, it is impossible for Prussia to have a policy." This idea was still too revolutionary for William, and compromise followed once more. By the Convention of Gastein (Aug. 14, 1865), Austria was given the administration of Holstein, Prussia of Schleswig. Bismarck, rewarded with the Prussian title of Count von Bismarck-Schönhausen (September 1865), soon moved again toward an alliance. In October he visited Napoleon III at Biarritz. Hitherto he had assumed that Napoleon would repeat in Germany the policy that he had followed with Count Camillo di Cavour in Italy: he would join in the war against Austria and expect to receive a territorial reward for his assistance. At Biarritz Napoleon made it clear that he planned to remain neutral and that his reward should be the cession of Venetia to Italy. The Biarritz bargain was negative: Bismarck would not guarantee Venetia to Austria, and Napoleon would not give Austria support.

After promising not to defend Venetia, it was an easy development to promise it to Italy. This was the essential clause of the alliance with Italy, which Bismarck signed on April 8, 1866; in return Italy promised to join in war against Austria if it broke out in the next three months. In May there was a last attempt at compromise, organized by the brothers Anton and Ludwig von Gablenz—one a Prussian, the other an Austrian general. Bismarck would have been satisfied with the military headship of northern Germany. Austria was still unyielding and at the beginning of June repudiated the alliance with Prussia by bringing the question of the duchies before the Diet. Prussian troops invaded Holstein, and when a majority of the Diet condemned Prussia, Bismarck declared the German confederation at an end (June 14). War followed against Austria and those German states that took the Austrian side.

Two Prussian armies invaded Bohemia, and on July 3 the Austrians were decisively defeated at Sadowa, near Königgrätz (Sadova, near Hradec Králové). They appealed for French help, but Napoleon III would only mediate. A believer in the national principle, he sympathized with the Prussian purpose and himself urged Bismarck to annex the whole of northern Germany. Bismarck's real struggle was with his own king. William had been gradually taught to believe that the Austrians were the aggressors, and he wished to punish them by annexing some of their territory. Bismarck got his way only after a bitter dispute, in which he called the Crown Prince to his aid. By the preliminary peace of Nikolsburg (July 26), he excluded Austria altogether from Germany, instead of di-

viding Germany between Prussian and Austrian zones at the line of the Main, as he had proposed before the war broke out. The southern German states were to have "an international independent existence." Bismarck's moderation in 1866 has been much praised, and certainly he wished to preserve Austria as a great power—"we shall need its strength in future for ourselves." But Austria remained intent on revenge until the events of 1870 made it impossible. Moreover, in preserving Austria, Bismarck saddled himself and his successors with "the Austrian problem." The War of 1866 defeated the empire that Metternich had served, but within a few years Bismarck re-created the conservative system of Metternich in a modified form.

A general election took place in Prussia on the day of the Battle of Sadowa (Königgrätz). The Progressive Party was much weakened and soon afterward split. The majority formed a new party, the National Liberals, that would support Bismarck in exchange for concessions in home affairs. He agreed to admit that he had acted illegally in collecting the taxes without parliamentary authorization, and the Diet passed an act of indemnity.

All the states north of the Main that had fought against Prussia in 1866 were annexed except Saxony; the others had to join a federation under Prussian control. The King of Prussia became its president and commander in chief. Bismarck set up a Bundestag elected by universal suffrage, but he did not propose to allow it any say in the military budget, which was to be written into the constitution; nor was there to be a responsible ministry—the real decisions were to be taken by a federal council composed of delegates from the states. The National Liberals revolted against this sham, and Bismarck met them halfway, perhaps with one eye on liberal feeling in southern Germany. The chancellor became a responsible minister, though the only one, and the military budget was authorized only until the end of 1871. Then, after further dispute, Bismarck agreed in 1874 to authorization only for seven years (the Septennate), and this produced a crisis every seven years while he remained in office.

Bismarck and the National Liberals made a genuine compromise. Certainly the chancellor was appointed by the king, or emperor, as the prime minister had been in 18th-century England; but neither Bismarck nor his successors ever infringed the constitution, though Bismarck was often tempted to do so. There was in imperial Germany a true rule of law, and the will of the people could have prevailed if they had known what to will. The Bundestag won the essential point of any constitution: it controlled supply for the military forces, though it could exercise this control only every seven years instead of annually, as in a fully liberal country.

Conflict with France: the Luxembourg crisis

At the time of making peace with Austria, the French ministers—though not Napoleon himself—demanded compensation for Prussia's aggrandizement. Bismarck answered with vague talk of Belgium, but when the French asked for an alliance and the cession of German territory on the left bank of the Rhine he evaded them. In September 1866, Napoleon III, who was wiser than his advisers, declared himself content with the new order in Germany, but French public opinion insisted on some concrete satisfaction. Luxembourg seemed the solution. It had been a member of the old confederation and was garrisoned by Prussian troops, but it was ruled by the King of The Netherlands, and Bismarck did not claim it for his new federation. Indeed he advised Napoleon III to buy it from the King of The Netherlands. The French dallied over these negotiations, and the German public began to protest against the loss of this "old German land." Bismarck himself joined in the protest (April 1, 1867), though finally he accepted compromise at an international conference: the Prussian garrison was withdrawn and Luxembourg was neutralized instead of being annexed to France.

The Luxembourg crisis is one of the most disputed episodes in Bismarck's career. Some writers have seen in it a trap by which he meant to involve France in war; others claim that he was surprised by the strength of German feeling and had to go along with it. Certainly the crisis put an end to all hopes of a Franco-Prussian alliance, such as Bismarck had often discussed. On the other hand, although Bismarck regarded war with Austria as one of the only two possible alternatives before 1866, he does not seem to have worked deliberately for war with France in the same way. Even in 1869 he suggested that France might acquire Belgium in return for the addition of the southern German states to the federation.

These states were the point about which Bismarck's policy revolved. In 1866 his renunciation of them had been genuine: he did not want to add Roman Catholic states to a predominantly Protestant federation. But he counted on their liberalism to keep them from alliance with either France or Austria. Instead the liberal parties in southern Germany lost ground to clericalists who were strongly anti-Prussian. Bismarck was faced with a dilemma. If he rallied the southern liberals by proposals of unification, this would strengthen liberalism throughout Germany; if he did nothing, Austrian and even French troops might soon be on the Main. His solution was to prepare for unification by increasing the prestige of the Prussian monarchy.

One device to achieve this was the proposal, which he put forward unsuccessfully early in 1870, to give the king of Prussia the title of German emperor. Another was to encourage members of the House of Hohenzollern to accept any vacant thrones that might be offered them. In 1866 Prince Charles of Hohenzollern-Sigmaringen went to Romania on Bismarck's advice to become prince and later king of that country. In 1869 his brother Leopold was offered the throne of Spain, vacant after a revolution in 1868. Though Bismarck later established the version that he had no connection with this offer, there is, in fact, no doubt that he repeatedly urged it on the Spaniards until they wore down Leopold's reluctance in June 1870. This does not imply that the Hohenzollern candidature was intended to provoke a war with France. France was to be presented with a *fait accompli*, and Leopold was to become king before the French learned of it.

Franco-German War

On July 3 the news leaked out prematurely. The French government demanded Leopold's withdrawal, and William, who had always disliked the affair, seconded them. On July 12 Leopold withdrew. Bismarck, though taken by surprise, had wished to reject the French demands and to score a victory of prestige, if not actually to provoke a war. Instead he seemed to have been humiliated. At the last minute he turned the humiliation against the French. He transformed a message that William sent him from Ems from a surrender into a defiance, and the French declared war a week later. Bismarck soon claimed that he had caused the war by means of this "Ems telegram," but it is likely that the Bonapartists, who dominated Napoleon III, would have insisted on war in any case.

The war of 1870, unlike the war against Austria, had no practical purpose; Bismarck asked nothing of France except to be left alone. But from the beginning of the war he accepted the military proposal to annex Alsace-Lorraine. The generals urged the strategical advantage; Bismarck thought rather of giving German national feeling some concrete symbol of victory, and, as well, he was glad to have some cause for lasting estrangement between France and the German liberals. Later he regretted this estrangement and assured the French that he had been opposed to the annexation of Metz. There is no contemporary evidence of this. His diplomacy during the war was limited to preventing the French from finding allies or provoking a European mediation. He was successful and in February 1871 imposed upon them an indemnity of 5,000,000,000 francs, as well as the loss of the territory. This became the Treaty of Frankfurt in May 1871.

The German Empire

All the southern German states joined the war against France, but they were still reluctant to enter an empire controlled by Prussia. It would have been easy to drive them on by evoking popular feeling, but Bismarck was determined that Germany should not be made by demo-

cratic means—it must be made by the princes, not by the Bundestag. He gave some political concessions to the southern German states and, more effectively, bribed the king of Bavaria with large sums from his secret fund. The Bundestag merely asked William to deign to accept the imperial crown; the actual offer was made by the princes, and William I was proclaimed German emperor on Jan. 18, 1871. Two things stand out in this supreme achievement of Bismarck's life. Germany was made without the German people, almost against them: though Germany was technically united, 8,000,000 German Austrians were excluded from the German national state.

Imperial chancellor: Prince von Bismarck. Nevertheless, Bismarck was the hero of the German people in 1871, and he was created Prince von Bismarck (March 21, 1871) and appointed imperial chancellor. The National Liberals acted almost as a government party in the Reichstag. Bismarck discussed every legislative measure with their leaders before introducing it and left the conduct of measures through the Reichstag to them. In addition, the conservatives also split, and one section, the Free Conservatives, accepted liberal measures so long as Bismarck advocated them. Between 1871 and 1878 Bismarck had a majority in the Reichstag, almost in the Western parliamentary sense; and he carried through fundamental changes, comparable with the work of the Revolution in France or of the Benthamites in England. Germany was given a common currency and a central bank. A single code of commercial and civil law was created, and a high court of justice set up at Leipzig. The tangle of medieval survivals was swept away in less than a decade, and Germany became the most modern and liberal of states, so far as laws could make it so.

Distrust of political Roman Catholicism was common doctrine among liberals in the 19th century. Bismarck shared this distrust, and it grew stronger in 1871 when a confessional party, the Centre, gained 58 seats in the Reichstag. Though the Centre claimed to defend strictly Roman Catholic interests, it drew its support from all the elements that had opposed Bismarck's work; for instance, Protestant Hanoverians supported it, while Bavarian nationalists, though Roman Catholic, did not. Bismarck regarded the empire as his own creation, and he branded any opponent of his as a *Reichsfeind* (enemy of the empire). He believed, too, that unity was most easily created if there was some object to attack, and the Centre provided that object when Germany had no foreign enemies. He said of Ludwig Windthorst, leader of the Centre, "Everyone needs somebody to love and somebody to hate. I have my wife to love and Windthorst to hate."

The struggle began over the state control of education. When the Roman Catholics claimed the right to choose and license their own teachers, Bismarck attacked the teaching orders and then insisted that the state should train and license priests. Priests and bishops were imprisoned and sees were left vacant. By 1876, when the struggle reached its height, Bismarck seemed to have committed himself irrevocably to the liberals; even the Prussian conservatives, though Protestant, were turning against him. It was the climax of the liberal era.

Bismarck assumed in 1871 that, with the defeat of France, a long period of peace would follow. The only danger seemed to be the French desire for revenge. Hence he encouraged republicanism in France in order to create an ideological gulf between the republic and the two eastern empires. With these empires he revived the conservative partnership of the Holy Alliance in a new form. This was the League of the Three Emperors, or Dreikaiserbund, constituted in 1873. Impressive in title, it was never more than a vague expression of monarchical solidarity in fact. The league implied that the three empires, and especially Russia and Austria-Hungary, would settle their disputes peacefully. In reality it held together only so long as there were no disputes.

League of the Three Emperors

Its weakness was shown in 1875, even before the Eastern question raised its head. Bismarck was on bad terms with the French government, whom he accused of patronizing the German Roman Catholics; in addition, he was offended at the reorganization of their army. In April a German newspaper published, on Bismarck's instructions, an article entitled "Is War in Sight?" Probably he wished only to shake the French nerve and to underline the Centre's association with the "foreign enemy." The French, however, raised the alarm and appealed to the great powers. Austria-Hungary did not respond—the first open symbol of Vienna's new dependence on Berlin. Both Russia and Great Britain protested in Berlin, and Bismarck had to repudiate his own words. The Eastern crisis, which began in July 1875, raised more serious difficulties. Germany had no direct interests in the Near East, and during the Crimean War Bismarck had been the most emphatic advocate of neutrality, even to the extent of refusing Russian diplomatic support. But then the main conflict had been between Russia and the western powers. Now France kept an attitude of reserve; Great Britain took the lead against Russia, and Austria-Hungary showed an increasing inclination to go over to the British side. Bismarck, at first, followed the French example and turned his back on the crisis. In December 1876 he declared that the whole of the Near East "was not worth the bones of a Pomeranian grenadier." His only concern was to keep Russia and Austria-Hungary on good terms, and he succeeded until the autumn of 1876. Then the Russians foresaw that they would be driven to war against Turkey for the sake of the Balkan Slavs; and they appealed to Bismarck to keep Austria-Hungary neutral, as they claimed (with some exaggeration) to have kept that monarchy neutral in 1870. Bismarck refused: he would not allow either Russia or Austria-Hungary to be destroyed as a great power.

The crisis reached its height in March 1878 when the Russians abandoned their moderate plans and imposed the Treaty of San Stefano on the Turks. Great Britain threatened war; Austria-Hungary would not promise peace. Bismarck acted as "honest broker" and brought Russia and Great Britain to a compromise. He could rightly claim to be the main architect of the peaceful settlement between them, which was confirmed by the Congress of Berlin (June 1878). He presided and this seemed to symbolize his paramount position as mediator between the great powers. All the same, the Congress of Berlin marked the turning point in Bismarck's diplomacy. Hitherto he had believed that Germany could secure peace by keeping out of the combinations of the great powers; after 1878, he tried to achieve the same aim by having a finger in every pie.

The Austro-German Alliance

The Eastern crisis of 1875–78 destroyed Bismarck's faith in monarchical solidarity. In addition it made him regard Russia as too unstable to make a reliable ally. He decided, with great reluctance, to commit himself to the defense of Austria-Hungary. William held out against the alliance with Austria-Hungary until Bismarck forced his hand by threatening his resignation and that of all the Prussian ministers. The alliance followed logically enough from Bismarck's policy in 1866. Then he had refused to destroy the Habsburg monarchy; now, therefore, he had to guarantee it against foreign danger. At the same time he still hoped to keep out of Austria-Hungary's Balkan adventures. The alliance was limited to a direct attack by Russia, and Bismarck even made out that it brought advantage to Russia, in that Austria-Hungary would no longer be tempted to work with Great Britain. He said to the Russian ambassador, "I wanted to dig a ditch between her and the western powers."

The alliance was something of a conjuring trick: it preserved Austria-Hungary without supporting Austro-Hungarian policies. But suppose Austria-Hungary should take the risk of an aggressive Balkan policy, of which Germany disapproved? This was the insoluble problem that Bismarck left to his successors. Previously alliances had only been concluded as preparation for a war or on its outbreak; now Bismarck tried to make international relations rigid. He himself regretted this later and advised his successors not to take their treaty obligations literally; every treaty, he said, contains the unwritten clause, *rebus sic stantibus* ("unless circumstances change").

The Austro-German alliance was signed on Oct. 7, 1879. It marked the beginning of a new period of conservatism in Bismarck's foreign policy. The same change had already taken place in his policy at home. Until 1877 he continued to work with the National Liberals. Then fiscal, economic, and political factors turned him against them. The existing taxes did not cover the cost of the armed forces, and the deficit had to be met by "matricular contributions" from the states. Bismarck wished to escape this dependence; on the other hand, he did not wish to increase the power of the Reichstag by introducing direct taxation in the empire. Indirect taxes were voted for an indefinite period; therefore, once granted, the Reichstag would have no more control over the military expenditure than the Prussian Parliament had had after 1862. The National Liberals foresaw this danger and they would agree to more indirect taxation only if they received "constitutional guarantees." In December 1877 Bismarck invited Rudolf von Bennigsen, the leader of the National Liberals, to become a Prussian minister and virtually his deputy. Bennigsen insisted that the invitation must be extended to two other National Liberals. Bismarck had planned to take Bennigsen prisoner. Instead the National Liberals threatened to turn Bismarck into the figurehead of a parliamentary ministry. He broke off the negotiations and on Feb. 22, 1878, announced his intention (never in fact achieved) of introducing a tobacco monopoly. This was his first breach with the National Liberal policy, and it ended the plan of giving the German Empire a parliamentary ministry.

Breach with the National Liberals

The breach with the National Liberals soon widened. In June 1878 an attempt was made to assassinate William I. Bismarck at once dissolved the Reichstag and launched an anti-Socialist campaign designed to hit the National Liberals on the rebound. They either had to renounce their liberal principles or be branded as the supporters of assassins. They lost 30 seats in the election and then tried to get back into Bismarck's favour by voting for the anti-Socialist bill after all. He was, however, intent on destroying them as a political force. In March 1879 he made his peace with Windthorst, the Centre leader, and at once began to undo the measures against the Roman Catholics, despite his early boast that he would never "go to Canossa" (where the emperor Henry IV submitted to the pope in 1077). His basic reason for the *Kulturkampf* (his conflict with Roman Catholicism) had been to create some target for national resentment. Now the Socialists filled the role even better. Furthermore, the Centre drew its main strength from the peasant farmers of western Germany. They sympathized neither with liberalism nor —what was more important—with free trade. Reconciliation with the Centre was the preliminary to protection, and this was the essential change that took place in Bismarck's policy in 1879.

The initial impulse was the decline in agricultural prices that hit all Europe at the end of the decade. Bismarck was determined to maintain the predominance of the agrarian classes—primarily of the Prussian Junkers, but also of the land workers who were supposed to make the best soldiers. Agrarian protection could not be carried by itself. Bismarck also introduced tariffs on manufactured goods, particularly on iron and steel. This satisfied the great industrialists of western Germany, who were henceforth ready to accept the Junker landowners as partners. The return to protection was more than a change of policy. Hitherto protection had been the defensive weapon of backward countries. Germany, the most progressive industrial nation, took it up for aggressive purposes.

Protection completed the estrangement of the National Liberals. They had always cared more for laissez-faire than for individual freedom, and Bismarck challenged them at their most sensitive point. In 1880 the party split, the more advanced members gradually drawing toward the Progressives. The National Liberal remnant became purely an interest group representing the needs of heavy industry.

Bismarck's manoeuvre was not altogether successful. He had destroyed the National Liberal majority without putting anything else in its place. The remaining National Liberals were in a minority even when the two conservative groups joined them. In fact the Centre held the decisive position, and it could defeat Bismarck's measures if it voted with the Progressives and the Social Democrats. His Reichstag policy between 1879 and 1890 all turned on this point: he had no solid party on which to rely, yet would not accept the open partnership with the Centre that was the only alternative. Hence he had to invent alarm cries, imaginary panics with which to stampede the electorate. The "social peril" had been the cry of 1878; it failed to work satisfactorily in 1881. Bismarck had to call in, first, colonial disputes in 1884 and then foreign dangers in 1887. Both were risky expedients that could not be repeated, and in 1890 Bismarck found himself once more without a majority.

Social security

Bismarck did not rely only on repression to defeat the Social Democrats. He was the first statesman in Europe to devise a comprehensive scheme of social security, offering the worker insurance against accident, sickness, and old age. This Bismarckian "socialism" later became a model for every other country in Europe. It represented in part the paternalist function of the state that Bismarck, as a conservative, had always held. But no doubt its prime function was as a weapon against the Social Democrats. It is commonly held that it failed of this purpose. The Social Democrats continued to increase despite both persecution and state insurance. Bismarck himself confessed in 1890 that he had failed and wished to establish a military dictatorship in order to crush the Social Democrats. But in a deeper sense he succeeded. The Social Democrats ceased to be a revolutionary party, and, when a mortal crisis came in 1918, they preserved the empire as Bismarck had created it. This was his last and most unexpected achievement.

Bismarck had not meant to commit himself exclusively to Austria-Hungary when he made the alliance of 1879. Rather he desired to renew good relations with Russia on a conservative basis. This was not difficult. The Russians asked only for security in the Near East, and the Austrians gave up hope of alliance with Great Britain when the Liberals came to power there in 1880. The League of the Three Emperors was renewed (June 1881). This was both a conservative partnership, echoing the phrases of the Holy Alliance, and a pact of practical cooperation in the Near East, guaranteeing Russia security at the Straits. Austria-Hungary assented to the league unwillingly; Vienna had a deep distrust of Russia that could never be surmounted and would have preferred an anti-Russian alliance with Italy and Great Britain. Bismarck, to satisfy the Austrians, agreed to a Triple Alliance with Italy (May 1882). This gave Austria-Hungary a guarantee at least of Italian neutrality in the event of war with Russia; in exchange Bismarck had to promise to defend Italy against France. This system of alliances—further extended to Romania in 1883—was strictly defensive and aimed at preserving the general peace of Europe. But, like the earlier League of the Three Emperors, it depended on Russia and Austria-Hungary remaining pacific and conservative in the Balkans, and it threatened to break down when new difficulties occurred there in 1885.

Colonial ventures

Bismarck's colonial ventures were closely related to his other policies. He had always repudiated interest in colonies. Pointing to Russia and France, with Germany in the middle, he said, "Here is my map of Africa." In 1884, however, he flung himself into colonial disputes with Great Britain and in the course of a single year acquired for Germany the Cameroons, South West Africa, East Africa, and part of New Guinea. His motives have been much debated. He certainly had close connections with the great trading firms of Hamburg and welcomed an occasion to estrange them from England and English liberal ideas. Moreover, he needed a cry with which to fight the election of 1884, and hostility to England seemed as good as any. He himself, as he grew older, clung ever more persistently to office. The old Emperor could not last much longer, and Bismarck feared that when the Crown Prince succeeded a liberal ministry

would be set up. He was constantly on the watch for this imaginary "German Gladstone ministry" and sought to drive out of public life anyone who might qualify as a member of it. But the Crown Prince could not himself be driven out; therefore he had to be isolated and preparations had to be made to discredit him. Since he had an English wife, England had to be presented as hostile to Germany.

Colonial ambitions also served the needs of Bismarck's foreign policy. He made a determined effort to be reconciled to France while the Eastern question was quiescent. He would not, of course, return Alsace-Lorraine, though he made misleading apologies for the annexation, but he offered to support the French anywhere else in the world. They suspected, however, that he was pushing them into conflict with Great Britain, and to lull these suspicions Bismarck trumped up disputes of his own with the British in order to be able to convince the French that he had common interests with them. This policy had a certain success so long as Jules Ferry, the great advocate of colonial expansion, was prime minister in France. But even Ferry did not intend a permanent reconciliation with Germany, though he used Bismarck's cooperation to further his colonial plans, and the policy broke down altogether when he fell from power on March 31, 1885. Bismarck then patched up his disputes with Great Britain and soon repudiated all interest in colonies. In 1889 he declared, "I am not a colonial man."

The army bill

The calm that had followed the Congress of Berlin was broken in 1886. In the Near East a new crisis was caused by the unification of Bulgaria. In France Georges Boulanger appeared as the leader of a patriotic movement agitating for revenge. Bismarck used this crisis both to strengthen his position in the Reichstag and to give his foreign "system" a new form. The Septennate, passed in 1880, was running out. Bismarck exaggerated the danger from France in order to justify an increased grant; moreover, by emphasizing the French danger, he could distract attention from Russia, which was his real anxiety. Finally he hoped to secure a new Reichstag more favourable to himself and thus get additional security against the accession of the Crown Prince. The parties in the Reichstag, including even the Progressives, were ready to agree to the proposed army increases but would grant them only for three years. This gave Bismarck the excuse for a conflict, which he deliberately provoked. In November 1886 the Reichstag authorized the army bill for three years only. It was at once dissolved. Bismarck persuaded the Conservatives, the Free Conservatives, and the National Liberals to form an electoral cartel on a so-called patriotic basis. Even Bennigsen, who had retired from politics in despair, was induced to return in order to defend the national cause. These tactics were successful. The Progressives lost half their seats, although the Centre was not shaken. The army grant was duly passed for seven years, and Bismarck had a Reichstag ready to support him in anything. As soon as the election was over, Bismarck minimized the danger from France and worked to remove the tension that he himself had largely created.

The Reinsurance Treaty

Bismarck's real anxiety was not an attack from France, which Germany could easily defeat, but a conflict between Russia and Austria-Hungary in the Near East. He urged the two countries to partition the Balkans, but Austria-Hungary would never agree to this. Bismarck was determined not to fight a war for Austria-Hungary's Balkan interests, let alone for the preservation of Turkey. He therefore provided Vienna with other allies who would bear the responsibility for him. In 1879 he had worked to separate Austria-Hungary from Great Britain; now he worked to bring them together. The two Mediterranean agreements (March and December 1887) between Austria-Hungary, Great Britain, and Italy were made under his auspices. By them the three powers agreed to maintain the status quo, particularly in the eastern Mediterranean.

Yet at the same time Bismarck himself made a secret agreement with Russia that deprived this coalition of much of its effect. The League of the Three Emperors, renewed for three years in 1884, ran out in 1887. The Russians, on bad terms with Austria-Hungary, would not renew it again. They suggested a new agreement with Germany alone. Bismarck accepted this, but revealed to them the Austro-German treaty of 1879, by which he would have to aid Austria-Hungary if Russia attacked it. The Russians countered by refusing to remain neutral if Germany attacked France. With these two exceptions, a treaty promising neutrality in the event of war was concluded in June 1887. This "Reinsurance" Treaty had a further clause by which Germany promised Russia diplomatic support in Bulgaria and at the Straits—the points on which the powers of the Mediterranean agreements were combined against Russia.

The Reinsurance Treaty may have made the Russians feel less isolated in the Near East and so deterred them from any violent impulse, but there is no good evidence that the Russians ever contemplated war there in 1887. Bismarck later made a much greater claim: the Reinsurance Treaty, he insisted, had kept Russia away from France, and the failure to renew it in 1890 caused the Franco-Russian alliance. This was an unjustified exaggeration. The Russians had little confidence in German policy even after the treaty, and it was the French who held back from alliance in 1887. Bismarck's economic policy itself helped to defeat his diplomacy. The rising German tariffs on Russian grain estranged the great Russian landowners, and when Bismarck closed the German market to Russian bonds in November 1887, he drove the Russians to look to Paris for money instead.

Frederick III

In March 1888 there came the catastrophe that Bismarck had long dreaded: William I died, and the Crown Prince succeeded as Frederick III. The liberal era seemed to have arrived. But the new emperor was already a dying man, and Bismarck had taken his precautions. Frederick III and his wife were surrounded by Bismarck's creatures, who cut them off from contact with the remaining liberal politicians. Bismarck remained securely in office. Only one reactionary minister, Robert von Puttkamer, was dismissed. Yet Bismarck was not satisfied. He launched a violent press campaign against Alexander of Battenberg, the former prince of Bulgaria, in the quite unfounded belief that the Emperor was planning to make him chancellor. Bismarck even appealed to the Russian government, though in vain, to threaten war if Alexander were nominated. These pathological alarms were ended only on June 15 when Frederick III died.

William II and Bismarck's fall

Bismarck had taken no precautions to make the new ruler, William II, amenable; on the contrary, he had encouraged him in subordination, as a preliminary to playing him off against his father. Still, Bismarck's position seemed unshakable. He had no political rivals, and in 1886 had made his son Herbert (born 1849; Count von Bismarck-Schönhausen from 1871) secretary of state in order to gain unchecked control of foreign affairs. Herbert had all his father's violence without his skill and encouraged him in the worst courses. Bismarck set out to take William II prisoner, as he had taken William I prisoner in 1862. His only method of governing was to raise the cry "the *Reich* in danger," and he intended to provoke a crisis so grave that a dictatorship would have to be set up, with himself as dictator and William II as a helpless puppet. But William II was a more skillful politician than his grandfather, and he understood the new Germany better than Bismarck. He was not prepared to be estranged from the mass of his subjects; he believed, not altogether wrongly, that the Social Democrats would cease to be dangerous if they were treated with conciliation. He preached a policy of social reform, including factory legislation and recognition of the trade unions. In his own words, he wanted to be a *roi des gueux* ("beggars' king"). He also advocated a demagogic foreign policy. He repudiated Bismarck's cautious policy of keeping Russian friendship and refusing to support Austria-Hungary. Instead he wanted alliance with Great Britain, unrestricted backing of Austria-Hungary, and a large-scale promotion of German economic interests in the Near East. Bismarck did not openly oppose this policy.

The crisis centred on the anti-Socialist laws, which were due to expire in 1890. The Conservatives wished to renew them without alteration; the National Liberals demanded a slight modification. Bismarck made no attempt to reconcile them. In fact, as he admitted, he wanted the bill to be rejected and then "the waves would mount higher and higher until a catastrophe occurred." The National Liberals insisted on their amendment. The Conservatives thereupon voted against the bill as a whole, and it was defeated with the aid of the radical and Social Democratic opposition on Jan. 25, 1890. The Reichstag was at once dissolved. The general election of February 20 brought defeat to the Bismarckian parties. The Progressives, the Social Democrats, and the Centre together had a strong majority. Bismarck welcomed the crisis. He proposed to carry out a coup d'etat. The German princes, who had made the empire, should now declare that it was dissolved and a new constitution should be drafted, abolishing universal suffrage and reducing the powers of the Reichstag. William II refused to follow this path; he would not, he said, stain the first years of his reign with the blood of his subjects. Bismarck attempted to isolate the Emperor. He revived a royal order of 1852, which forbade the ministers to advise the Emperor except in the presence of the prime minister, and he appealed to the ministers to threaten a joint resignation, as they had done in 1879. But times had changed. Bismarck was old, William young, and the ministers had their careers before them. Only Herbert Bismarck stood by his father.

Resignation

At the last moment Bismarck tried to turn foreign policy to his advantage. He claimed that the Russians were willing to renew the Reinsurance Treaty only if he remained in office. William II refused to be taken prisoner. On March 18 he demanded Bismarck's resignation. Bismarck drafted a letter of resignation, emphasizing only the differences in foreign policy and blaming William for the estrangement from Russia. He arranged that this letter should be published on the day of his death.

Last years. No one but Herbert went with Bismarck into retirement. Bismarck was an implacable foe. He never forgave an injury, even an imaginary one, and refused to use the personal title of duke of Lauenburg that the Emperor conferred on him when he retired. His last years were devoted solely to discrediting William, though the Emperor made many efforts at reconciliation. Bismarck used every weapon in this last campaign. In a speech at Leipzig in 1892, he criticized imperial influence and urged the German people to make the empire more democratic, although he had spent his life thwarting their democratic wishes. This demagogic appeal did not work. The German masses rightly regarded Bismarck as their enemy, and the left-wing parties even managed to prevent an address of congratulation to Bismarck from the Reichstag on his 80th birthday in 1895. Bismarck then swung back to the extreme right and ended up, as he had begun, as a violent Junker. He emphasized his claim that he had built and maintained good relations with Russia that his successors had destroyed. In 1896 he even published the text of the Reinsurance Treaty—a breach of official secrecy for which any lesser man would have been prosecuted.

His main energy in these last years went into the composition of his *Reflections and Reminiscences* (*Gedanken und Erinnerungen*), a work of great literary genius though of doubtful historical value. He increased the drama of every event and always presented himself in a favourable light. Bismarck died at Friedrichsruh on July 30, 1898, about three years after the death of his wife. He carried his feud with William II to the grave, for the inscription on his tombstone, which he devised himself, read: "A True German Servant of the Emperor William I."

Herbert Bismarck had entered the Reichstag in 1893, where he acted as a member of the extreme right. He advocated high agrarian protection and alliance with Russia, opposed colonial expansion, and became increasingly anti-British. During the South African War he was one of the most obdurate pro-Boers. He died in 1904 and with his death the challenge of the "Bismarckians" to William II came to an end.

ASSESSMENT

Bismarck was a political genius of the highest rank, but he lacked one essential quality of the constructive statesman: he had no faith in the future. The revolutions of 1848 convinced him that the old order could not be preserved unchanged, and all his later policy was shaped by this conviction. He went with the modern forces of liberalism and democracy solely to draw their sting. Like Metternich he regarded them as evil; unlike Metternich he turned them to his own purposes. He is sometimes compared with the leaders of the English governing classes, who under Sir Robert Peel also made a compromise with liberalism and democracy. But there was a basic difference. In England there was a genuine compromise, in Germany only a trick. The German people were defrauded, given a shadow instead of the substance.

Greatness in foreign policy

Bismarck was at his greatest in foreign policy. There he understood, as no one else did, "the art of the possible." He never aspired to dominate Europe; he was content to balance between the great powers. Though he had no moral objection to war, he preferred to get his way by diplomacy and went to war only for limited aims when it was necessary to his policy. The system of alliances that he built up was designed to secure the peace of Europe, and he played the powers off against each other with matchless skill. In fact, though no believer in eternal peace, he was the principal architect of the halcyon age that gave Europe 26 years of peace after the Congress of Berlin.

In domestic affairs his record is less inspiring. He had a lust for power that grew on him with the years. He wanted Prussia to be supreme in Germany. He wanted the king to be supreme in Prussia; but most of all he wanted to be supreme over the king. His boasted loyalty to the crown vanished as soon as William II showed signs of independence, but he was equally ruthless, though more subtle, with William I. He spoke contemptuously of the old Emperor's intelligence and did not shrink from the most unscrupulous tricks in order to keep his hold. His suspicion of possible rivals was unbounded, and he persecuted them out of public life one after the other. In his latter years all his energies went into the search for the "German Gladstone cabinet," which he was convinced was being prepared against him, a search all the more degrading in that this cabinet was always a creation of his imagination. He battered down any politician who dared to cross him even in trifles and refused to allow the Reichstag to pay a posthumous tribute to Eduard Lasker, a sincere National Liberal who had done much for the empire but had shown some independence. Yet Bismarck himself changed course whenever it suited him. He repudiated old friends and old policies without scruple and often showed the disloyalty that he denounced in others. He lived in the age of democracy and German power, and he devoted his life to making these two forces as harmless as possible. Despite his ringing, self-confident phrases, he was at heart a despairing conservative, caring only for the past, dreading the future, and trying to retard its arrival. Gladstone said of him, "He made Germany great and Germans small."

BIBLIOGRAPHY. There is no standard biography of Bismarck even in German. The more recent works in English include ERICH EYCK, *Bismarck and the German Empire*, 2nd ed. (1964), which is in part the belated revenge of a German Liberal; A.J.P. TAYLOR, *Bismarck: The Man and the Statesman* (1955), largely an expansion of the present article; and WERNER RICHTER, *Bismarck* (1962; Eng. trans., 1964). The first two volumes of Bismarck's autobiography were translated as *Reflections and Reminiscences* (1968), and the third volume has not been translated. There is a full discussion of books on Bismarck in G.P. GOOCH, "The Study of Bismarck," in *Studies in German History* (1948, reprinted 1969).

Sources: The collected edition of Bismarck's works, *Die gesammelten Werke*, ed. by HERMANN VON PETERSDORFF *et al.*, 15 vol. (1924–32), though claiming to be complete, must be supplemented by the following: (*Political correspondence*):

RITTER VON POSCHINGER, *Preussen im Bundestag, 1851–59*, 3 vol. (1882–84); LUDWIG RASCHDAU (ed.), *Die politische Berichte des Fürsten Bismarck aus Petersburg und Paris, 1859–62*, 2 vol. (1920); J. LEPSIUS, A. MENDELSSOHN-BARTHOLDY, and F. THIMME, *Die grosse Politik der europäischen Kabinette 1871–1914*, 16 vol. (1922–27). (*Speeches*): HORST KOHL (ed.), *Die politischen Reden des Fürsten Bismarck*, 14 vol. (1892–1903). (*Letters*): HORST KOHL (ed.), *Briefe* (1900; Eng. trans., 1903); HERBERT BISMARCK (ed.), *Fürst Bismarcks Briefe an seine Braut und Gattin*, 2 vol. (1900; 2nd ed., 1906; Eng. trans., *The Love-Letters of Prince Bismarck*, 2 vol., 1901); WOLFGANG WINDELBAND (ed.), *Briefe an seinem Sohn Wilhelm*, 2nd ed. (1922); HORST KOHL (ed.), *Briefe an seine Schwester und seinen Schwager 1843–97* (1915), *Briefe an den General Leopold v. Gerlach* (1896); A. ZEISING (ed.), *Briefwechsel mit Gustav Scharlach* (1912), *Briefwechsel mit dem Minister Freiherrn von Schleinitz, 1858–1861* (1905); HEINRICH VON POSCHINGER (ed.), *Fürst Bismarck und der Bundesrat, 1867–90*, 5 vol. (1897–1901). (*Reminiscences*): JULIUS BUSCH, *Graf Bismarck und seine Leute während des Krieges mit Frankreich*, 2 vol. (1878); *Tagebuchblätter*, 3 vol. (1898; Eng. trans., *Bismarck: Some Secret Pages of His History*, 1898); HEINRICH VON POSCHINGER, *Fürst Bismarck und die Parlamentarier*, 3 vol. (1894–96); R. LUCIUS VON BALLHAUSEN, *Bismarck-Erinnerungen*, 3rd ed. (1921); R. VON KEUDELL, *Fürst und Fürstin Bismarck: Erinnerungen aus den Jahren 1846–1872*, 3rd ed. (1902); HERMANN VON MITTNACHT, *Erinnerungen an Bismarck* (1904–05); ADOLF VON SCHOLZ, *Erlebnisse und Gespräche mit Bismarck* (1922); KURD VON SCHLOZER, *Petersburger Briefe* (1922).

(A.J.P.T.)

Bivalvia

The Bivalvia, known commonly as bivalves, constitute one of the major classes of the invertebrate phylum Mollusca. The group is characterized by a bivalved shell, *i.e.*, one with two separate sections; oysters, mussels, cockles, scallops, and clams are typical members. The bivalve shell is compressed toward the closure line between the valves, which are connected by an elastic ligament (valves and ligament jointly constituting the shell); with but few exceptions, the animal is entirely enclosed within the shell. Valves and ligament are secreted by the mantle (a fleshy or membranous outgrowth of the outer body wall) and are equivalent to the univalve shell of a snail. The two valves are usually bilaterally symmetrical, *i.e.*, they are mirror images of one another. Enclosure of the head with consequent withdrawal of the mouth from contact with the bottom, or other solid surface, has resulted in the evolutionary loss of the head. The mantle margins are responsible for sensory functions. Feeding is by the labial palps (lips), or, more usually, by large paired gills (ctenidia); a water current is created by cilia (*i.e.*, tiny hairlike structures). Minute organic particles in the water are retained in the lattice work of the ctenidia.

GENERAL FEATURES

Distribution and size range. Most bivalves are marine, occurring in all latitudes and at all ocean depths. Many species occur in freshwater. The size of bivalves ranges from that of certain Condylocardia—*i.e.*, about one millimetre (about .04 inch) long—to the giant clam of the South Pacific, *Tridacna gigas*, which may be over 120 centimetres (about four feet) long and weigh up to 300 kilograms (about 700 pounds).

Economic importance. No group of aquatic invertebrates is of greater economic importance than the bivalves. The animals have been used as food, and their shells have been used as objects of barter, as tools, and as ornaments from the earliest recorded times; they continue to be so used by primitive peoples. Generally, all bivalves—with the possible exception of the thorny oyster, *Spondylus*—are edible, although there is occasional danger from poison accumulated by the animals during so-called red tides, in which swarms of organisms known as dinoflagellates (usually *Gonyaulax*) are taken in with the normal planktonic food (microscopic plants and animals). In polluted areas, typhoid bacilli may be similarly accumulated without harm to the bivalve.

The most important edible bivalves are oysters of the genera *Ostrea* and *Crassostrea*. *Ostrea* includes the European flat oyster, *O. edulis*, and the small olympic oyster, *C. lurida*, of the Pacific. *Crassostrea* includes the American, Portuguese, and Japanese oysters, *C. virginica*, *C. angulata*, and *C. gigas*, respectively—all have been, or are, cultivated. *O. edulis* was cultivated in Roman times; in Japan, *C. gigas* has been cultivated at least since the 18th century. The common mussel (*Mytilus edulis*) is now widely cultivated in Europe, north of La Rochelle, on the Bay of Biscay, in France, where it is grown on stakes interwoven with brushwood—a method that originated in the 13th century. In northwestern Spain, where *Mytilus* is cultivated on a large scale, the mussels are grown on ropes suspended from floats. In the United States various species of oysters are cultivated; the soft-shell clam *Mya arenaria*, the hard-shell clam *Mercenaria* (*Venus*) *mercenaria*, and other clams are also cultivated on both the Atlantic and Pacific coasts. The scallops (*Pecten* and *Chlamys*) are also highly prized as food. The extent to which prehistoric man depended on bivalves has been indicated by discoveries of large man-made scrap piles composed almost exclusively of bivalve shells. Although bivalves of the genera *Pinctada* and *Pteria*, the source of pearls, have been collected by divers in many tropical seas since early times, many pearl-oyster fisheries have been discontinued since the commercial development in Japan of methods for artificial stimulation of pearl formation. In artificial stimulation, small beads (made from the shells of freshwater bivalves) are enclosed in a small bag of mantle (shell-forming) tissue and, by a simple operation, are introduced into the reproductive region of the body of living pearl oysters (*Pinctada*). In most cases the tissue graft survives, and layers of nacreous (mother-of-pearl) material are laid down, forming what is—apart from the centre—a true pearl. The oysters are kept in baskets slung from rafts in sheltered waters in which phytoplanktonic (*i.e.*, microscopic chlorophyll-containing organisms) food is abundant. This procedure is now the basis of a large industry in the shallow waters off the east coast of Japan and elsewhere in tropical regions in which species of *Pteria* are used. Smaller, less regular pearls are also obtained from various freshwater bivalves of the family Unionidae.

Oyster cultivation

The large gold-lip pearl shell (*Pinctada maxima*) is a source of mother-of-pearl, taken from the nacreous inner layer of the shell. The windowpane shell (*Placuna placenta*), common in the Indian Ocean and the west Pacific Ocean and living free on the surface of mud, has long been collected for its flat translucent valves, which are used for glazing windows. In the Philippine Islands, the shells are made into lampshades, trays, bowls, and other articles for export.

Damage by bivalves

Wooden boats, wharves, and other timber structures exposed to seawater require protection from attacks of various species of shipworms (Teredinidae). *Teredo navalis*, the best known, was once a serious menace in The Netherlands, where dikes were built partly of wood. The use of other materials and the development of modern methods of protecting timber, usually involving poisonous impregnations, have largely removed this danger. Piddocks (Pholadidae), the date mussels (*Lithophaga*), and other bivalves are able to bore into rock and sometimes damage man-made concrete structures in the sea.

NATURAL HISTORY

Reproduction and life cycle. Most bivalves are of separate sexes; *i.e.*, an animal is either male or female. The paired gonads open primitively into the pericardium, a membranous sac surrounding the heart. In most cases, however, the gonoducts (*i.e.*, the tubes leading from the gonads) open separately, and the eggs or sperm pass into renal (kidney) ducts before being discharged into the water. The eggs are usually small. Large eggs occur in cases of direct development (*i.e.*, without a larval stage); and in cases where development occurs within the gills (*e.g.*, marine species of Carditacea and Erycinacea, some species of *Teredo*, in freshwater species of Sphaeriacea and Unionacea).

Hermaphroditism (*i.e.*, a state in which functional re-

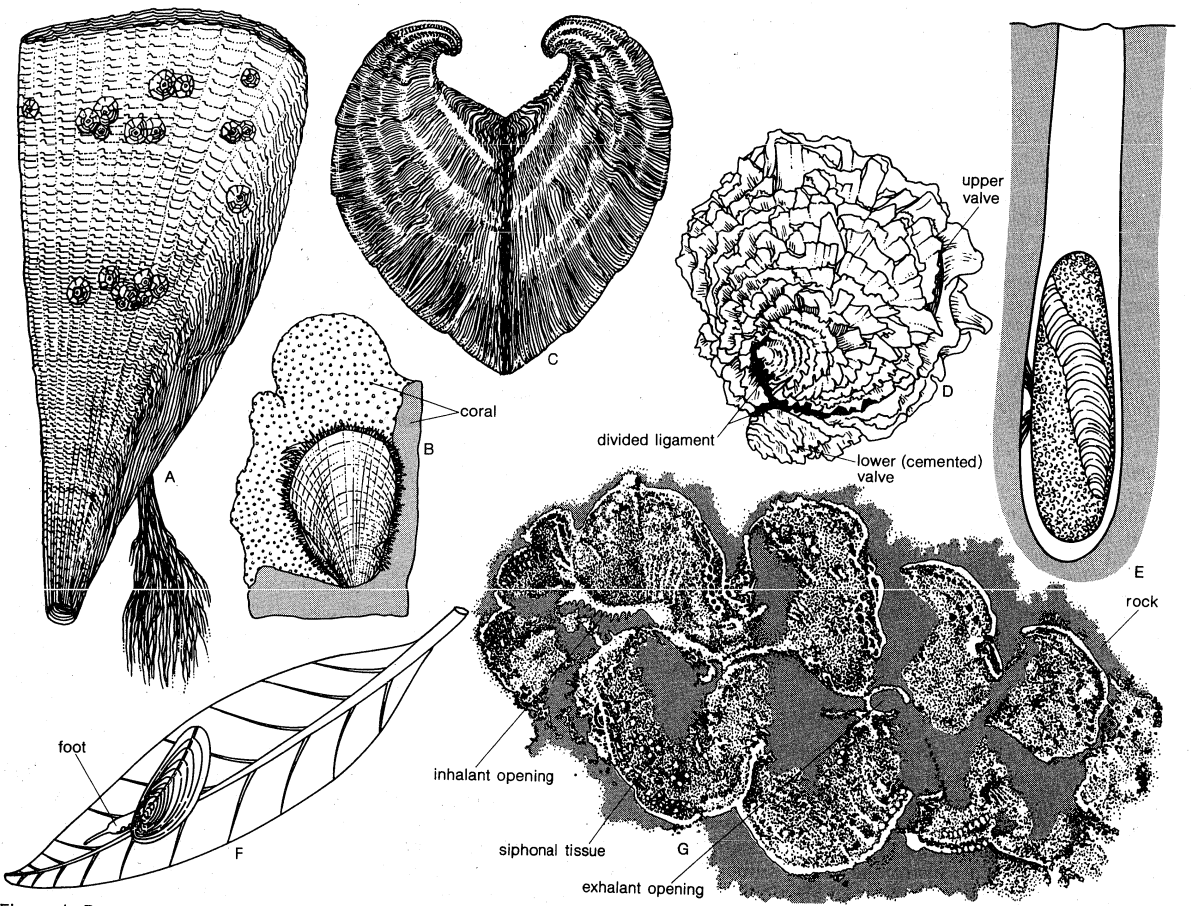


Figure 1: *Representative bivalves* in situ.
(A) *Pinna fragilis*, a fan shell showing attaching byssus threads; barnacles dot the shell. (B) *Pedum spondyloideum*, coral scallop in depression within coral. (C) *Arca senilis*, a West African "cockle," one of the ark shells. (D) *Chama pellucida*, the agate chama, cemented to rock. (E) *Lithophaga cumingiana*, a date mussel in excavated boring in calcareous rock. (F) Indo-Pacific *Enigmonia aenigmatica (anomid)* on a mangrove leaf. (G) Tropical *Tridacna crocea* in rock boring, with siphonal tissues containing symbiotic plant cells (zooxanthellae) exposed to light.
(G) By courtesy of the Geological Society of America

productive organs of both sexes occur in the same individual) takes various forms; the gonad may be divided into separate ovary and testis, as in certain scallops (most scallops have separate sexes), or sperm-producing follicles and egg-producing follicles may occur among one another in the gonad, as in the Tridacnidae. All such animals, which are represented in many superfamilies, are known as simultaneous hermaphrodites. Other bivalves known as consecutive hermaphrodites, because they alternate sexual roles, usually function first as males; there are exceptions, however, in which the initial phase is female. In the former type there may be a period of simultaneous hermaphroditism during which both types of gametes, or sex cells, are produced—*e.g.*, in the common *Mercenaria mercenaria* and the woodboring *Bankia setacea*. Rhythmical consecutive hermaphroditism occurs in the European flat oyster, *Ostrea edulis*, which usually is first male then female and continues to alternate from one sex to the other at a rate conditioned by the water temperature. It functions as male and female in alternate years in waters around Great Britain, but functions in both capacities in one season in the Mediterranean. This type of hermaphroditism occurs in all species of *Ostrea* in which the eggs are incubated and then must pass out the inhalant chamber, against the flow of water, and through the gills after spawning. *Teredo navalis* also alternates in sex, but its life-span is usually too brief to permit more than one change. In oysters of the genus *Crassostrea* (including common American, Portuguese, and Japanese species), in which both egg and sperm are freely discharged, most spawn first as males; later, the proportion of sexes becomes about equal; among the older individuals there are more females. Most bivalves appear to be true males or females, never changing sex.

Veliger larvae

In the majority of marine species and in the freshwater Dreissenacea, following the initial embryonic stages, a trochophore larva (*i.e.*, a free-swimming form) and then a veliger larva (which has the beginnings of a shell, mantle, and foot) develop. The latter also has the characteristic ciliated velum, or sail, which can be drawn within the already functional bivalve shell. The veliger larva swims by means of the long cilia around the velum, with the hinge beneath. Small cilia around the base of the velum carry food particles—microscopic plants—to the mouth. When the larva permanently attaches itself to the substrate (*i.e.*, any solid surface), radical changes in form occur. These changes may be delayed if the larva does not immediately find an appropriate surface. During this process the velum is lost by partial conversion into the labial palps. Several structures develop: the foot, which is used by the oyster for temporary crawling; the byssus gland, which is used for temporary attachment of the larva and then usually lost; and the first gill filaments. The fully adult form and habit are then assumed.

Notable exceptions to this course of development occur in the Protobranchia (*e.g.*, *Nucula*, *Yoldia*), in which the large velum, consisting of rows of large ciliated cells, gives the larva a barrel-like appearance; direct development occurs in protobranchs either within or outside the mantle cavity (*i.e.*, the space between the mantle and the body). In the freshwater Unionidae, the glochidium, or larva, with a toothed bivalve shell, is stimulated by shade to attach itself to the gills or fins of a passing fish; here it becomes encysted and temporarily parasitic, until it eventually escapes by rupture of its cyst, emerging as a free-living adult. The so-called haustorius larva of the Mutelidae is still further modified; that of the African *Mutela bourguignati*, for example, develops a filament

which it uses to attach itself to the fins of the cyprinid fish, *Barbus altianalis radcliffi*. The bivalve form gradually develops as the animal feeds parasitically. Sexual dimorphism (*i.e.*, secondary differences in form between male and female) is rare. In the protobranch *Nucula delphinodonta*, the female carries a brood chamber plastered on the outside of the posterior, or rear, end of the shell. In the Carditacea (*Thecalia* and *Milneria*) the shell of the female is modified as a brood pouch. The initial male phase of the wood borer *Xylophaga dorsalis* possesses a sperm-retaining organ that is absent in the later female phase. There is similar evidence of a secondary sexual organ in males of the septibranch *Cuspidaria*. In the freshwater Lamsilinae, the development of a marsupium, a structure for the containment of early embryos, from the gills results in a change in the shape of the female shell.

Behaviour and ecology. Compression of the valves toward one another has affected the foot, which, in contrast to the creeping organ probably possessed by primitive mollusks and certainly by existing snails and chitons (Polyplacophora), has become well adapted for penetration into and movement through soft substrates. By virtue of the lateral compression of shell and foot and the ability of these structures to collect suspended or deposited material (by means of water currents created by action of cilia on the gill filaments), the bivalves may be regarded as the most successful of all infaunal (*i.e.*, burrowing into the sea bottom) animals that feed by means of cilia. Many (*e.g.*, mussels), however, have secondarily become attached to a hard surface either by horny byssus threads secreted by a gland at the base of the foot or as a result of cementation by one valve (*e.g.*, edible oysters) to a surface. In this way, bivalves have also become important constituents of the epifauna (the animals, collectively, living on the surface of the sea bottom).

Attachment of sessile forms

From both infaunal and epifaunal origins, bivalves have become adapted for rock boring. Other bivalves are highly efficient wood borers, notably the shipworms (Teredinidae), which both bore into and digest wood. There are also instances of commensalism (*i.e.*, close association between two organisms in which one derives an advantage and the other is unaffected), for example, in the Erycinacea, and one example of parasitism. Although usually slow moving through soft substances or permanently attached by byssus threads or cementation, certain bivalves, notably certain scallops (Pectinidae) and species of *Lima* and *Solemya*, can swim for short distances.

Nucula, *Mactra*, and *Poromya* are infaunal. (*Nucula* is disposed parallel to the surface, the others are vertical.) This is the primitive habit of the Bivalvia—that is, unattached and slow moving. Changes in external form involving the relationship between the mantle–shell unit and the combined mass of foot and viscera (visceropedal mass), however, have permitted alteration in habit and the exploitation of different surfaces. Deep burrowers occur in the superfamily Tellinacea, in which the siphonal tubes (which draw in and expel water) are not united; the inhalant one is extensile and searches over the surface for organic deposits that are actively drawn in (*e.g.*, *Tellina*, *Abra*, *Macoma*, *Scrobicularia*). The animal may be up to 15 centimetres (about six inches) below the surface, in some cases lying on one side. Such tellinaceans are deposit feeders, as distinct from suspension feeders. Tellinaceans are largely successful because their tubular-shelled, vertically disposed bivalves are capable of extremely rapid upward and downward movement through soft sand or mud. The siphons are very short but the foot is enlarged to form an efficient organ for vertical movement. Other types of deep burrowing bivalves are the gapers—*e.g.*, species of *Mya*, *Lutraria*, and *Panope;* each belongs to a different superfamily, but the fused siphons of all are elongated. The animal penetrates deeper as it grows larger and may attain depths to 60 centimetres (about two feet)—*e.g.*, the geoduck *Panope generosa*. The siphons contract when the animal is exposed or endangered.

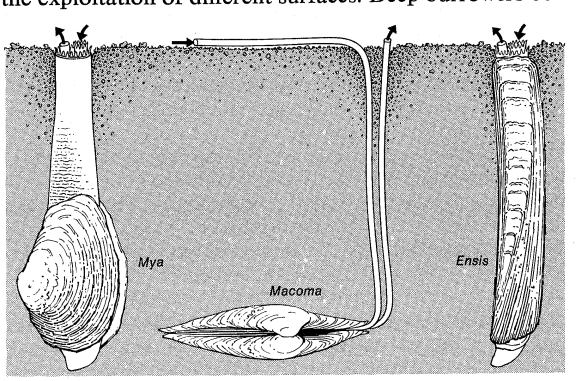


Figure 2: Deep-burrowing bivalves.

Mussels maintain themselves most effectively on an exposed rocky surface. At the enlarged posterior end the inhalant opening is raised clear of the bottom. This condition, which is always associated with byssal attachment, appears sporadically throughout the Bivalvia. It is present in the related Pinnidae, or fan shells, including the genera *Pinna* and *Atrina*, which live vertically embedded in soft material (a return to a fixed infaunal life); in the freshwater Dreissenacea (freshwater mussels superficially like the marine *Mytilus*); and members of other groups including *Cardita suborbicularis* (Carditacea).

Byssal attachment is fully retained in the pearl oysters and in the greatly modified saddle oysters (Anomiacea); it is also retained in some scallops, but many of these (species of *Pecten*, *Amussium*, *Chlamys*) become free and, horizontally disposed, are able to swim hinge hindmost. They also execute sudden escape movements with hinge foremost, water being suddenly ejected between the shell valves. If placed on the left side they can also turn over. These swimming movements appear to represent developments of mechanisms evolved in connection with cleansing action in the mantle cavity.

Other Pectinidae, for example, the Pacific rock scallop, *Hinnites*, and particularly the tropical thorny oysters, *Spondylus*, become cemented to a surface by way of the mantle and shell. The foot is retained to aid in cleansing the mantle cavity except in the Ostreidae and the Plicatulidae, which apparently became cemented at a more remote period in their evolutionary history and so have lost all trace of the foot.

Although most widespread among members of the somewhat miscellaneous order Anisomyaria, monomyarianism (*i.e.*, reduction to one adductor muscle for closing the valves) occurs elsewhere, notably in some freshwater Etheriidae and in the Tridacnidae. All members of the former (a family of the Unionacea) are adapted for life in rushing water; one South American genus of this group, *Acostea*, which attaches by either valve, becomes monomyarian by loss of its anterior adductor muscle during growth of the already reduced anterior regions. Monomyarianism has been attained in a unique manner in the tropical Tridacnidae. All species are attached by a massive byssus, although this is lost in the adult giant clam, which remains relatively fixed by virtue of its great weight. The structure of the tridacnids is unique; the hinge and ligament have moved from middorsal (as in *Cardium*) to midventral, lying beside the aperture through which the byssus passes. The free margins of the valves, now on the upper surface, are separated by greatly extended siphonal tissues. Originally (as in *Cardium*) confined to the posterior end, these margins are pulled out, occupying the entire upper surface; they are also laterally extended so that their scalloped margins extend, when expanded, over the edges of the shell. A long, broad expanse of highly pigmented tissue is thus exposed to sunlight, which is intense because the tridacnids are confined to shallow midtropical waters, usually on coral reefs. The movement of the mantle in relation to the visceropedal mass—which itself is little affected apart from the reduction of the anterior retractor of the foot and byssus—is explained by the presence in the exposed siphonal tissues of large populations of unicellular algae (zooxanthellae), which form a substantial portion of the

animals' food; the usual feeding and digestive organs are retained, however. Possibly the additional source of nutrition explains the exceptional size of the giant species, *Tridacna deresa* and *T. gigas*. One consequence of the turning movement of the mantle is the loss of the anterior adductor and the increased size and central position of the posterior adductor. *Lima* and *Tridacna* are the only vertically disposed monomyarians.

Cementation to a surface also occurs in other groups and may not always occur in monomyarians. In Etheriidae, cementation (by either valve) occurs in the genus *Etheria*. *Myochama* and *Cleidothaerus* (Pandoracea) are cemented by the right valves, but species of the largely tropical Chamidae are cemented by either valve. Here the shell grows in a turning movement involving a splitting of the ligament from the anterior end. The process of cementation influences the form of the mantle and shell, but not that of the visceropedal mass—a conclusive demonstration of the fundamental difference between their growth processes.

Locomotion. As noted, the laterally compressed foot is an efficient organ for penetration of soft materials. It is extended by the pressure of blood contained within extensive sinuses, or hemocoeles (*i.e.*, spongy tissues) under the control of various muscles. These muscles include persisting members of certain shell muscles present in the primitive mollusks and retained in that form in *Neopilina*.

Burrowing

Burrowing takes place by a series of digging cycles. First, the foot is extended and used as a probe with the siphons open. The siphons then close while the adductor muscles contract. This closure and contraction has the dual effect of compressing the water in the mantle cavity as well as the blood in the hemocoeles. The former leaves the shell by the only remaining opening, the pedal gape (through which the foot is extended), thereby loosening the substrate in that region. Blood is forced into the tip of the foot, which dilates to form an anchor; the pedal, or foot, retractors contract, pulling the animal forward; the adductors then relax, and the siphons reopen either permanently or during a brief resting period prior to the beginning of the next digging cycle.

The swimming movements of the scallops (Pectinidae) involve jet propulsion. *Lima hians* also swims, partly by the aid of long, vertically disposed tentacles.

FORM AND FUNCTION

Structural features. The Bivalvia consist of an inner body made up of a dorsal (*i.e.*, on the upper side) visceral mass and a ventral (*i.e.*, on the lower side) muscular foot, the two constituting the visceropedal mass, which is completely enclosed within a mantle, or pallium, that secretes the shell. The foot and the mantle margins can be protruded, and in most bivalves all tissues can be withdrawn within the protection of the shell. The internal cavity formed by the overhang of the mantle lobes constitutes the mantle cavity, which extends on either side of and behind the visceropedal mass.

There is much greater variety of habit in the Bivalvia than might at first appear. Members of the two major subdivisions, Protobranchia and Lamellibranchia, exemplify wide differences in habits while exhibiting little difference in outward form.

Gill structure

The protobranchs, so named because of the relatively simple structure of the gills, or ctenidia, have many undoubtedly primitive characters and have, if shell form is a reliable indication, remained unchanged since the Early Paleozoic Era (570,000,000 years ago). The rounded, nutlike shell of *Nucula* is bilaterally symmetrical (*i.e.*, the right and left valves are similar); they are also largely symmetrical anteroposteriorly. This results from reduction of the anterior region consequent to the loss of the head. The two adductor muscles are of similar size (isomyarian), and the two pairs of anterior and posterior pedal (foot) retractor muscles, slightly more centrally placed, are similarly symmetrical. Although laterally compressed as it protrudes to cut through a substance, the margins of the large foot open out widely to grip the surface as the pedal muscles contract, thus drawing the shell forward. The gills have a central axis with alternately placed filaments extending outward on either side, the axis being attached to the roof of the mantle cavity by a membrane. As in the gills of the most primitive existing gastropods (snails and slugs), lateral cilia on the opposed faces of the filaments create an upward respiratory current. Water is drawn anteriorly into a ventral inhalant chamber and passes between the gill filaments into the smaller, dorsal exhalant chamber; into the latter open the anus and the reno- (kidney) reproductive pore, and all of their products thus are discharged posteriorly by way of the exhalant current. Adjacent filaments are united by ciliary junctions. Particles filtered out as the current passes through the gill are collected on frontal cilia and may contribute to the food supply.

The largest organs in the mantle cavity are the labial palps, enlarged upper and lower lips concealing the mouth, which lies in the midline behind the anterior adductor. Only bivalves possess labial palps, and, in *Nucula*, they may be regarded as indicating the manner in which—during the evolution of the bivalves—the mouth maintained contact with the environment as it was gradually enclosed by the mantle and shell. The outer palp on each side bears an elongated, extensile proboscis with a ciliated groove on its inner face. The proboscis extends outside the shell and actively collects organic deposits from the material in which *Nucula* superficially burrows. This material is passed between the inner grooved faces of the opposed palp lamellae and exposed to the sorting action of complex ciliated tracts, which control the type of material that enters the mouth. Excess material is ejected onto the mantle surface, where it collects in mucus-laden masses known as pseudofeces. This material is periodically expelled by sudden contractions of the adductor muscles. These are divided into a catch muscle for sustained contraction and a quick muscle for periodic sudden contractions.

Pseudofeces

The form of the shell in *Spisula*, one of the commonest of sand-burrowing lamellibranch bivalves, and about twice the size of *Nucula*, is similar to that of the smaller bivalve. It also is equivalve (*i.e.*, the two valves are similar), equilateral (anteroposteriorly symmetrical), and isomyarian; the gills are very large, extending the length of the mantle cavity from the anterior adductor and consisting of many more filaments than in *Nucula*. Water is drawn past the filaments by the action of cilia. All intercepted particles are conveyed by oralward currents between the palps. Gills are the sole means of feeding; the palps are concerned with food selection. Pseudofeces collect in the inhalant cavity in *Spisula* and are rejected as in *Nucula*. Here, as in the great majority of lamellibranchs (and in some protobranchs), the inhalant as well as the exhalant current enters posteriorly. This permits the animals to burrow, anterior end downward; water is both drawn in and expelled from the posterior end. In *Spisula*, both openings lie at the end of a short united, retractable siphon formed by posterior extension of the mantle margins. An important distinction from *Nucula* is the fusion of the mantle margins, separating the ventral, inhalant opening from the dorsal, exhalant opening.

Poromya, a representative of the small group, the septibranchs (which are usually subsumed under the subclass Lamellibranchia), is similar in form to the protobranchs and lamellibranchs. It is equivalve, largely equilateral and isomyarian, and has a well-developed foot and anterior and posterior retractors. It is different in that the gills are replaced by a septal partition. This structure is slung like a hammock across the mantle cavity and attached by muscles at either end. It is connected laterally with the mantle wall and internally with the sides of the foot (or posterior to this with the other half of the septum). This septum is pierced in two places on each side by a small ciliated lattice through which water flows upward. Dead or motionless animals, small crustaceans, and worms, are drawn into the mantle cavity and pushed into the large mouth by muscular action of small, nonciliated, palps. No mechanism is needed for expulsion of waste particles. Unlike other bivalves, the septibranchs are carnivorous (*i.e.*, they feed on living animals), the gut being suitably modified. The stomach forms a crushing gizzard.

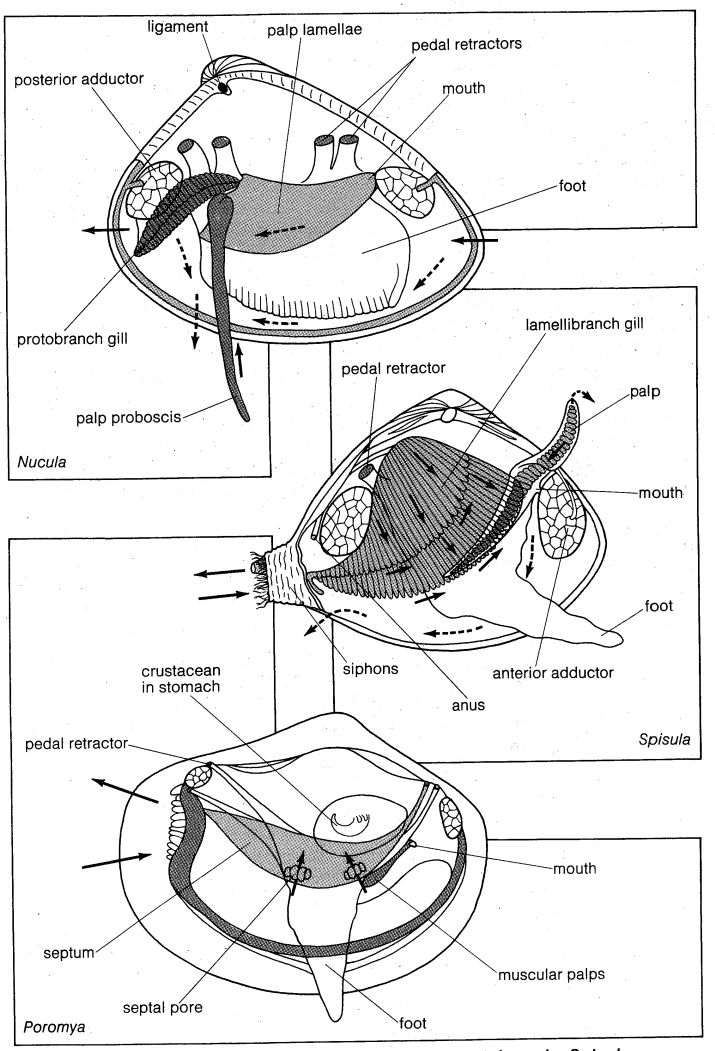


Figure 3: Internal views of *Nucula*, a protobranch; *Spisula*, a lamellibranch; and *Poromya*, a septibranch. Solid arrows indicate position of inhalant and exhalant water currents; broken arrows, currents collecting excess material (pseudofeces).

Change in shell form in *Nucula*, *Mactra*, and *Poromya* is a consequence of changes in rates of growth around the mantle margin. If these are more or less equally diminished on either side of the midventral region, then the rounded to oval shell found in *Nucula*, *Spisula*, and *Poromya* is produced. If growth gradients are higher at the two ends than in the middle, then an elongated shell is formed, with the hinge in the middle of the dorsal surface, as in the jackknife clam *Tagelus* (Tellinacea). More often only the posterior end is extended, with the hinge very near the anterior end as in the well-known razor shells, species of *Siliqua*, *Solen*, and *Ensis* (Solenacea). *Tagelus* is equilateral, but the razor shells are not.

Significance of the byssal attachment

The appearance of a byssus in some adult forms may be an instance of neoteny or paedogenesis, *i.e.*, the retention of larval characters. Such attachment in the postlarval stage is widespread—indeed, probably universal among bivalves. By the persistence of the byssal attachment the Bivalvia are able to colonize hard surfaces—that is, to become epifaunal instead of infaunal.

Such permanent attachment may have profound effects on form, hence, on habit. Throughout the evolutionary changes that have occurred, the foot and the byssus it secretes represent a fixed point in relation to which the mantle and shell and the anterior and posterior regions of the visceropedal mass may alter in their proportions. With the valves vertically disposed, the first effect of byssal attachment is a flattening of their ventral margins, for example, in attached ark shells (Arcacea) and in the mussels (*e.g.*, *Modiolus*, *Mytilus*). The ventral intake of water at the posterior end is affected by this flattening, with its accompanying pressure against a hard surface. In attached species of *Arca*, the problem is met by anterior as well as posterior intake of water. The shell remains equilateral with adductors of similar size, and the hinge line is parallel to the ventral surface. But in the mussels (Mytilidae) there is great posterior enlargement of the mantle and shell and of the visceropedal mass; there is also a corresponding anterior reduction. The shell is not equilateral, and the hinge line is inclined at an angle of up to 45°. A heteromyarian condition results: the anterior adductor is reduced and the posterior one enlarged. The anterior pedal (largely byssal) retractor muscle is also reduced and the posterior ones enlarged. The resultant bivalve is secured by radiating byssal threads, planted by movements of the elongated foot.

Further change along these lines leads to greater anterior reduction and loss of the anterior adductor. Bivalves with this monomyarian condition are horizontally disposed, all but the Ostreidae lying on the right side and —in species in which the byssus is retained as it is in the pearl oyster (*Pinctada*)—with the hinge line at right angles to the ventral surface. The anterior part of the mantle and shell and the anterior region of the visceropedal mass are confined to the small area between the centre of the hinge line and midline of the reduced and now very anteriorly placed foot. There is bilateral asymmetry, the lower valve being the more concave. The mantle and shell, and to a large extent the visceropedal mass also, become reorganized around the central adductor muscle so that bivalves such as *Pecten* acquire a secondary radial symmetry. They may be considered equilateral with the hinge line dorsal and the middle of the free margin of the valves ventral; but the foot, where it persists, indicates the morphological midventral point.

The mantle lobes secrete the calcareous valves; the mantle isthmus secretes the elastic ligament. Growth occurs at the margin. The mantle has three marginal folds; the outer is concerned with shell secretion, the middle—which carries tentacles and occasionally eyes (in scallops)—is sensory, and the innermost is muscular, controlling the inflow of water created by cilia on the gills.

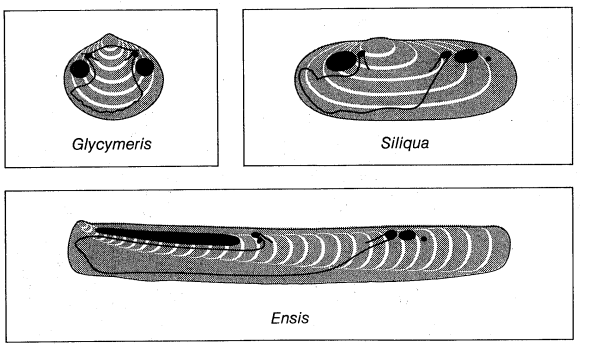


Figure 4: Representative bivalves, *Glycymeris*, *Siliqua*, and *Ensis*, showing direction of marginal growth and muscle attachments.

Shell layers

The shell has three layers. The superficial periostracum —a horny outer coating on the valve—is secreted by the inner surface of the outer mantle fold. The outer calcareous layer is formed by the outer surface of this fold, and the inner calcareous layer is formed by the general surface of the mantle. Usually this layer is bounded by a so-called pallial line where the pallial, or orbital, muscles are attached to the shell. Only this inner calcareous layer can thicken with age. The ligament consists of corresponding layers, but the innermost two are uncalcified, and the periostracum is often worn off.

The periostracum largely consists of protein; similarly, the other layers have a matrix of protein (conchiolin) impregnated with crystals of calcium carbonate that may be deposited in one of three forms: calcite, aragonite, or vaterite. Both the organic matrix and the crystalline calcium carbonate are formed in the extrapallial fluid, which lies between the mantle and the inner surface of the shell.

In this fluid are determined the pattern of the matrix and the nature of the crystal growth. Shell structure varies, but generally the outer layer is of prism-shaped sections, the inner in leaflike sections or layers. In the latter, nacre is formed if the crystals are of aragonite, calcitostracum if the crystals are of calcite.

The mid-dorsal hinge consists of elastic ligament, the opening thrust of which operates against the closing action of the contracting adductor muscles; sometimes the ligament is lost. In addition, there are usually teeth that interlock to ensure close approximation of the valves when closed. The teeth, with the corresponding sockets on the other valve, are secreted by a thin layer of mantle below the mantle isthmus, which forms the pallial crest and separates the calcareous surfaces throughout life. Various types of dentition occur and are of major taxonomic value, especially in fossil shells.

Functional features. *Circulation.* The heart, enclosed in the so-called pericardium, consists of one muscular ventricle with an auricle opening into it on each side. Blood enters the auricles by way of vessels from the gills and is driven into anterior and posterior vessels (aortae) by contractions of the ventricle. The blood then enters cavities called sinuses and usually returns to the heart by way of either the gills or the kidneys. There are no capillaries. Except in a few species in which the oxygen-carrying pigment hemoglobin occurs, there is no respiratory pigment. The blood corpuscles are amoeboid (*i.e.*, move by means of temporary extensions of the body known as pseudopodia), and some possess digestive powers.

Respiration and feeding. Blood flows through spaces within the gill filaments in the direction opposite to the upward flow of water between the filaments. In *Nucula* and similar protobranch bivalves, the gills retain an almost exclusively respiratory function. The palp proboscides are the feeding organs. In all lamellibranch bivalves the enlarged gills become, as in *Spisula*, the exclusive organs of feeding. There is great variation in the form of the gill filaments and in the arrangement of ciliary tracts upon them. Special straining cilia, the laterofrontals, are developed between the lateral cilia responsible for the formation of the water current and the frontal cilia now concerned with food collection and its passage, mixed with mucus, to the palps and mouth.

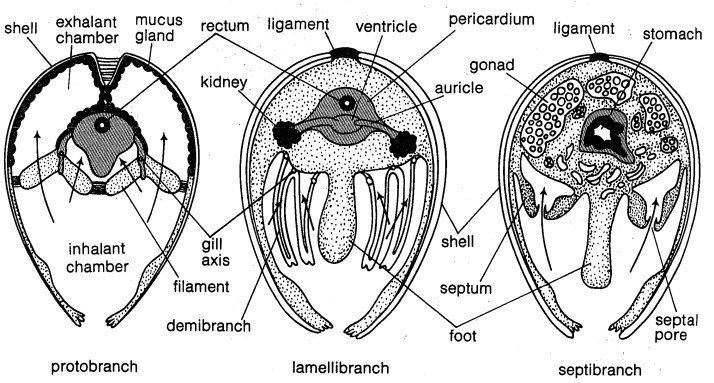


Figure 5: Transverse sections through bodies of a protobranch, a lamellibranch, and a septibranch. Arrows indicate direction of respiratory (and feeding) current in protobranch and lamellibranch; in septibranch they indicate ciliary-produced current through septal pores, which are closed when septum is drawn upward.

The lamellibranch ctenidium may be filibranch (W-shaped in cross section, with long filaments) or eulamellibranch (W-shaped in cross section and with two well-developed adductor muscles). In the former, adjacent filaments are connected by ciliary junctions (as in the protobranch condition), but in the eulamellibranch gill these are firmly united by tissue junctions of great complexity. Both types of gill are highly efficient in the collection of finely divided plant life. Their respiratory role is of minor importance. This is evident in the septibranchs, in which respiration must take place through the mantle surface. But the respiratory needs of these slowly moving, often attached bivalves are slight.

Digestion. The mouth leads by way of a short esophagus into a stomach, which communicates by ciliated ducts with the surrounding mass of digestive diverticula (formerly hepatopancreas or digestive gland). Associated with the midgut is a sac containing the crystalline style. The head of the style bears against the gastric shield, a hardened area of the stomach wall and the only non-ciliated surface in the gut. Rotation of the style helps to draw food into the stomach and to mix it; in addition, it slowly dissolves, liberating enzymes (biological catalysts) that digest starch, cellulose, and possibly fats. The diverticula may secrete some digestive enzymes. Elaborate ciliated sorting areas permit only finer particles and digested material to pass into the digestive diverticula for absorption or intracellular (*i.e.*, within the cells) digestion. Larger particles are conveyed by way of a deep groove into the coiled midgut, which passes dorsally into the hindgut, or rectum. The anus opens on the hindface of the posterior adductor. Food absorption occurs only in the extensive tubules of the digestive diverticula; the function of the midgut and hindgut is to consolidate the feces into firm pellets that will not foul the exhalant chamber. In the septibranchs, in which this procedure is unnecessary, the midgut is straight.

Digestive enzymes

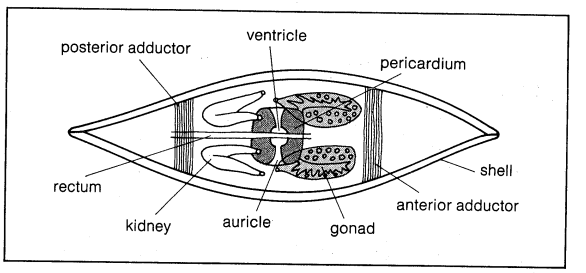


Figure 6: Pericardial (coelomic) complex in typical eulamellibranch bivalve.

Excretion. The kidneys, or organs of Bojanus, which open internally into the pericardium and externally into the exhalant chamber, consist of looped, usually branched tubes. Fluid forced through the walls of the heart into the pericardium flows through them and receives wastes. The kidneys may be united, as in the Tridacnidae, in which they are also greatly enlarged. Most bivalves also possess pericardial glands, which consist of excretory tissue lining the outer surface of the auricles or extending into the mantle (Keber's glands); in both cases these glands discharge into the pericardium.

Nervous system and sense organs. The nervous system is simple. The primitive molluscan system of four pairs of ganglia (*i.e.*, nerve centres)—cerebral, pleural (pallial), pedal, and visceral—occurs only in protobranchs such as *Nucula;* in all other bivalves, the two first ganglia unite, forming the cerebropleural ganglia, which lie above the esophagus. The pedal ganglia are in the foot, and the visceral ganglia are on the underside of the posterior adductor. Connective nerves unite the ganglia, and nerves extend through the body and into the siphons. In Pectinidae (scallops), which have the anterior regions of the body reduced, the cerebropleural and pedal ganglia are displaced in the direction of the visceral ganglia; the latter are enlarged to form a kind of visceral brain, a development even more evident in *Lima.*

With the loss of the head and the general inactivity of the bivalves, sense organs are poorly developed. Organs responding to the stimuli of vibration and gravity (otocysts) occur in the foot, except in the Ostreidae, which cement themselves to a surface and in which the foot is lost. In cemented bivalves that retain the foot, the otocysts persist. Vestigial (degenerate) eyes occur at the anterior end of the ctenidia in some primitive genera, but well developed pallial eyes are present around the siphonal openings in the cockles (Cardiacea). In the Pectinidae they are very numerous around the mantle margins. They are just as numerous and just as well developed in byssally attached forms (*Pedum*) and cemented forms (*Spondylus*) as in free-swimming forms (*Pecten*,

Sight

Chlamys). The pallial eyes are situated on the middle fold of the mantle margin. Bivalves without eyes also react to light, indicating the presence of photosensitive cells or else the direct effect of light on the nervous system. The numerous tentacles on the middle mantle fold are sensitive to chemical substances.

EVOLUTION AND CLASSIFICATION

Evolution. Existing Mollusca consist of a number of well defined classes that do not form an ascending series in terms of evolutionary development but rather have evolved radially from a common ancestral condition. From consideration of common structures, a primitive mollusk has been postulated from which existing classes, as diverse in external appearance and habit as gastropods, bivalves, and cephalopods, can be derived. The precise characteristics of this hypothetical animal have been questioned since the discovery of the first living monoplacophoran, *Neopilina galatheae*, with its possible, although by no means certain, indication of a primitive segmentation. This, however, is more of a concern of the origins of the Mollusca; for the Bivalvia there is no reason to alter previous postulates that the class arose from somewhat limpet-like animals with an anterior head and a ventral foot (the head foot) and above this the concentrated viscera covered with a mantle that secreted a domelike shell. A posterior mantle cavity housed one pair of gills with associated pallial organs. A series of paired shell muscles connected the foot with the shell, which initially was not calcified.

Classification. No satisfactory system of classification of bivalves has so far been devised. Paleontologists are concerned with the microstructure of the shell and with the hinge teeth and ligament. Neontologists (*i.e.*, specialists in recent, or nonfossil, forms) must also take into account organ systems, such as gills, palps, stomach, and the nature and extent of mantle fusion. Many superfamilies of the Bivalvia have persisted with remarkably little change since the Early Paleozoic (which began 570,000,000 years ago) and later ones from the Mesozoic (which began 225,000,000 years ago) and much that has been discovered about living bivalves appears applicable to their remote ancestors.

Distinguishing taxonomic features. The classification of the Bivalvia is based on the types of dentition, gill structure, and stomach form; also on the structure and degree of fusion of the mantle and on the structure of the valve.

Annotated classification.

CLASS BIVALVIA

Two calcareous valves joined by a noncalcareous ligament; mostly marine, some live in brackish water or freshwater; internally and externally symmetrical; head is rudimentary; two symmetrical gills used for respiration and feeding; locomotion usually by means of a well developed foot; some forms are sessile (*i.e.*, permanently fixed and nonmoving) being cemented or else attached to substrate by byssal threads. About 15,000 species.

Subclass Protobranchia

Gills with 2 divergent rows of flat, short filaments; shallow burrowers.

Order Nuculacea

Dentition taxodont (*i.e.*, with numerous similar teeth along hinge area); feed by means of palp proboscides.

Order Solemyidae

Hinge teeth reduced; feed by means of gills.

Subclass Lamellibranchia

Gills lamellibranch (*i.e.*, shaped like thin plates) or septibranch (reduced to horizontal muscular septa, or membranes).

Order Taxodonta

Taxodont dentition; gills filibranch (*i.e.*, w-shaped in cross section, with long filaments attached to each other by cilia); byssally attached or shallow burrowers.

Order Anisomyaria

Heteromyarian (*i.e.*, anterior adductor muscle reduced) or monomyarian (*i.e.*, without anterior adductor muscle); primitively byssally attached, some becoming free or cemented; hinge teeth reduced, secondary or absent; includes mussels, pearl oysters, scallops, saddle oysters, edible oysters.

Order Schizodonta

Hinge teeth few and divergent.

Order Heterodonta

Cardinal teeth (those near the anterior end of the hinge) and lateral teeth (comprising ridge below hinge area) well developed; eulamellibranch (*i.e.*, gills w-shaped in cross section); basically shallow burrowers, includes cockles, giant clams, tellins, and venerids.

Order Desmodonta

Teeth small or absent; mostly deep burrowers or borers; includes razor clams, carnivorous septibranchs.

Critical appraisal. The most important classification based on the characters of modern bivalves (of greatest concern here) is that of one authority, who divided the class into four orders: Protobranchia, Filibranchia, Eulamellibranchia, and Septibranchia. Later the Pseudobranchia were added to include groups that lay indeterminately between the Filibranchia and Eulamellibranchia.

It has been suggested that bivalves evolved along three branches, normal, sedentary, and burrowing or boring. This view persists despite the fact that the bivalve form must initially have evolved in relation to the exploitation of the infaunal habit with byssal attachment there persisting, where it persisted at all, only for temporary attachment in soft substrates.

Classification is satisfactory only at the level of superfamilies as these have been defined by some authorities. There is a remarkable degree of adaptive radiation within many of these. They exhibit many common features concerning the nature and extent of mantle fusion involving the mode of formation of the siphons and the manner in which the ligament is formed. There is evidence that might unite the primitive eulamellibranch Carditacea and Astartacea into one superfamily, which possibly might be considered an order; however, so far as living bivalves are concerned, there is no urgency in producing a definitive classification.

The position is different for fossil groups, however, since some classification schemes are based on both paleontological and neontological criteria. One, for example, involves arrangement of the superfamilies under 15 orders, which are grouped under five subclasses. But there is strong argument for grouping the superfamilies into no more than five orders.

BIBLIOGRAPHY. R. TUCKER ABBOTT, *American Seashells* (1954), the standard popular work on the classification of American shells; J.A. ALLAN, *Australian Shells, with Related Animals Living in the Sea, in Fresh-Water and on the Land* (1950); A.H. COOKE, *Cambridge Natural History*, vol. 3, *Molluscs* (1895), old-fashioned but still useful; E. FORBES and S. HANLEY, *A History of British Mollusca and Their Shells*, 4 vol. (1848–53); A. FRANC, "Classe des Bivalves," in P.P. GRASSE (ed.), *Traité de zoologie*, vol. 5, pp. 1845–2133 (1960), an up-to-date and very detailed scientific account of the bivalvia; P.S. GALTSOFF, "The American Oyster, *Crassostrea virginica* Gmelin," *Fishery Bull. Fish Wild. Serv. U.S.*, vol. 64 (1965), the fullest scientific account of this species; A. MYRA KEEN, *Sea Shells of Tropical West America* (1958), a standard scientific account of Eastern Pacific species; J.L. KELLOGG, "Ciliary Mechanisms of Lamellibranchs with Descriptions of Anatomy," *J. Morph.*, 26:625–701 (1915); P. KORRINGA, "Recent Advances in Oyster Biology," *Q. Rev. Biol.*, 27:266–308, 339–365 (1952); J.E. MORTON, *Molluscs*, 4th rev. ed. (1967), the best short account of the mollusks; R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology*, pt. N, 2 vol., *Mollusca 6, Bivalvia* (1969); N.D. NEWELL, "Classification of the Bivalvia," *Am. Mus. Novit.*, no. 2206 (1965), classification from standpoint of paleontology; P. PELSENEER, *Mollusca*, pt. 5 of E.R. LANKESTER (ed.), *A Treatise on Zoology* (1906), the standard work in its day, still of value; A.W.B. POWELL, *Shells of New Zealand*, 3rd ed. (1957); R.D. PURCHON, *The Biology of the Mollusca* (1968), excellent account of recent aspects of bivalve studies; N. TEBBLE, *British Bivalve Seashells* (1966); J. THIELE, *Handbuch der systematische Weichtierkunde*, pt. 3, *Classis Bivalvia* (1935), the standard scientific work on classification of Bivalvia; E.R. TRUEMAN, "Bivalve Mollusks: Fluid Dynamics of Burrowing," *Science*, 152:523–525 (1966); R.D. TURNER, *A Survey and Illustrated Catalogue of the Teredinidae (Mollusca: Bivalvia)* (1966), excellent systematic account of the shipworms; K.M. WILBUR and C.M. YONGE (eds.), *Physiology of Mollusca*, 2 vol. (1964–

1966), most recent account of functioning throughout Mollusca.

(C.M.Y.)

Bizet, Georges

The most naturally gifted and original of all French composers of the mid-19th century and universally remembered for his opera *Carmen*, Georges Bizet was also the composer of many other works hardly inferior in quality. During his short life he was handicapped by the French public's exclusively operatic tastes in music; and, when he died at the age of 36, he had only just begun to achieve the wholly individual style and the dramatic power that show fitfully in his earlier music.



Bizet.
The Bettmann Archive

Bizet was born in Paris on October 25, 1838, and baptized Georges, though his registered names were Alexandre-César-Léopold. His father was a singing teacher and his mother a gifted amateur pianist; the boy's musical talents declared themselves so early and so unmistakably that he was admitted to the Paris Conservatoire before he had completed his tenth year. There, his teachers included the accomplished composers Charles Gounod and Fromental Halévy, and he quickly won a succession of prizes, culminating in the Prix de Rome, awarded for his cantata *Clovis et Clotilde* in 1857. This prize carried with it a five-year state pension, two years of which musicians were bound to spend at the French Academy in Rome.

The Prix de Rome

Bizet had already shown a gift for composition far superior to that of a merely precocious boy. His first stage work, the one-act operetta *Le Docteur miracle*, performed in Paris in 1857, is marked simply by high spirits and an easy mastery of the operetta idiom of the day. His *Symphony in C Major*, however, written in 1855 but subsequently lost and not discovered and performed until 1935, will bear easy comparison with any of the works written at the same age of 17 by either Mozart or Felix Mendelssohn. Flowing and resourceful counterpoint, orchestral expertise, and a happy blend of the Viennese classical style with French melody give the symphony a high place in Bizet's output.

The young composer was already aware of his gifts and of the danger inherent in his facility. "I want to do nothing *chic*," he wrote from Rome, "I want to have *ideas* before beginning a piece, and that is not how I worked in Paris." In Rome he set himself to study Robert Schumann, Carl Maria von Weber, Mendelssohn, and Gounod, who was regarded as more than half a German composer by the admirers of the fashionable French composer Daniel Auber.

> Mozart's music affects me too deeply and makes me really unwell. Certain things by Rossini have the same effect; but oddly enough Beethoven and Meyerbeer never go so far as that. As for Haydn, he has sent me to sleep for some time past.

Instead of spending his statutory third year in Germany, he chose to stay on in Rome, where he collected impressions that were eventually collected to form a second C major symphony (*Roma*), first performed in 1869. An Italian-text opera, *Don Procopio*, written at this time, shows Donizetti's style, and the ode *Vasco de Gama* is largely modelled on Gounod and Meyerbeer.

When Bizet returned to Paris in the autumn of 1860, he was accompanied by his friend Ernest Guiraud, who was to be responsible for popularizing Bizet's work after his death. In spite of very decided opinions, Bizet was still very immature in his outlook on life (youthfully cynical, for instance, in his attitude toward women) and was plagued by an artistic conscience that accused him of preferring the facilely charming in music to the truly great. He was even ashamed of his admiration for the operas of his Italian contemporary Giuseppe Verdi and longed for the faith and vision of the typical Romantic artist, which he could never achieve. "I should write better music," he wrote in October 1866 to his friend and pupil Edmond Galabert, "if I believed a lot of things which are not true." In fact the skepticism and materialism of the dominant Positivist philosophy persistently troubled Bizet; it may well have been an inability to reconcile his intelligence with his emotions that caused him to embark on so many operatic projects that he never brought to a conclusion. The kind of drama demanded by the French operatic public of the day could very seldom engage his whole personality. The weaknesses in the first two operas that he completed after his return to Paris are a result not so much of the composer's excessive regard for public taste as of his flagging interest in the drama. Neither *Les Pêcheurs de perles* (*The Pearl Fishers;* first performed 1863) nor *La Jolie Fille de Perth* (1867; *The Fair Maid of Perth*) had a libretto capable of eliciting or focussing the latent musical and dramatic powers that Bizet eventually proved to possess. The chief interest of *Les Pêcheurs de perles* lies in its exotic Oriental setting and the choral writing, which is more individual than that of the lyrical music, over which Gounod still casts a long shadow. Although *La Jolie Fille de Perth* bears only a skeletal resemblance to Sir Walter Scott's novel, the characterization is stronger (the gypsy Mab and the "Danse bohémienne" anticipate *Carmen*), and even such conventional features as the night patrol, the drinking chorus, the ballroom scene, and the heroine's madness exhibit a freshness and elegance of language that raise the work unmistakably above the general level of French opera of the day.

Return to Paris

Although warmly acknowledged by Berlioz, Gounod, Saint-Saëns, and Liszt, Bizet was still obliged during these years to undertake the musical hackwork that only the most successful French composers were able to avoid. Stories of his moodiness and readiness to pick a quarrel suggest a profound inner uncertainty, and the cynicism and vulnerability of adolescence hardly yielded to a mature emotional attitude of life until his marriage, on June 3, 1869, to Geneviève Halévy, the daughter of the composer of the opera *La Juive* (1835; *The Jewess*). Between his engagement in 1867 and his marriage, Bizet was himself aware of undergoing "an extraordinary change . . . both as artist and man. I am purifying myself and becoming better." Adverse criticism of certain features of *La Jolie Fille de Perth* prompted him to break once and for all with "the school of *flonflons*, trills and falsehoods" and to concentrate his attention on the two elements that had always been the strongest features of his music—the creation of exotic atmosphere and the concern with dramatic truth. The first of these was brilliantly exemplified in the one-act *Djamileh* (1872), original enough to be accused of "exceeding even Richard Wagner in bizarrerie and strangeness"; and the second in the incidental music for Alphonse Daudet's play *L'Arlésienne* (1872), which is marked by a delicacy and tenderness quite new to his music. Besides the happiness of his marriage, which was crowned by the birth of a son in July of this same year, his letters show that he was deeply stirred by the events of the Franco-Prussian War, and, during the siege of Paris, he served in the national guard.

Effect of Bizet's marriage

First production of *Carmen*

It was in the first flush of this new emotional maturity, but with the ardour and enthusiasm of youth still unshadowed, that he wrote his masterpiece, *Carmen*, based on a story by the contemporary French author Prosper Mérimée. The realism of the work, which caused a scandal when it was first produced in 1875, was to inaugurate a new chapter in the history of opera; and the combination of brilliant local colour and directness of emotional impact with fastidious workmanship and a wealth of melody have made this opera a favourite with musicians and public alike. The philosopher Friedrich Nietzsche regarded it as the type of "Mediterranean" music that was the antidote to Wagner's Teutonic sound. The scandal caused by *Carmen* was only beginning to yield to enthusiastic admiration when, on June 3, 1875, Bizet died suddenly at his home in Bougival, on the Seine River near Paris.

MAJOR WORKS

OPERAS: Eight published operas, including *Les Pêcheurs de perles* (first performed 1863); *Carmen* (1875).

ORCHESTRAL WORKS: *Symphony in C major* (composed 1855); *Petite Suite*, arranged from *Jeux d'enfants* (1871); suite, *L'Arlésienne* (1872).

INCIDENTAL MUSIC: Incidental music for Daudet's *L'Arlésienne* (1872).

PIANO MUSIC: *Jeux d'enfants*, twelve pieces for piano duet (1871); *Variations chromatiques de concert* (1868).

SONGS: Thirty-seven published songs, including "Chanson d'avril" (?1866); "Berceuse" (1868).

BIBLIOGRAPHY. WINTON B. DEAN, *Bizet* (1948), a short, but extremely thorough study of the composer and his work, with useful background information; MINA K. CURTISS, *Bizet and His World* (1959), a fascinating re-creation of Bizet as a man and of his position in his family and among his contemporaries, based on correspondence and documents that only became available after 1950.

(M.Du P.C.)

Black Muslims

The Black Muslims (officially titled the Nation of Islam) are a quasi-religious, black nationalist organization among Afro-Americans. Based in Chicago, Illinois, the Nation of Islam was led, almost from its inception, by Elijah Muhammad, reverently called by his followers "The Honorable Elijah Muhammad, Holy Prophet and Messenger of Allah." To his followers he was divinely chosen and inspired to unite black Americans under Islam for their ultimate emancipation from white rule.

NATURE AND SIGNIFICANCE

Elijah Muhammad fulfilled a traditional desire among less advantaged black Americans, going back to slave days, for a messiah who will set them free [see also NEGRO CULTS (IN THE UNITED STATES)]. This messianic impulse, common among politically and socially oppressed groups, is one of the reasons the Nation of Islam has gained such popularity among black Americans, many of whom see their situation as similar to biblical accounts of the Egyptian bondage of the children of Israel and their eventual divine deliverance. Muslims believe all black Americans are chosen as members of the "ancient tribe of Shabazz, . . . the lost-found Nation of Islam in the wilderness of North America."

At first the leaders of the Nation of Islam placed its antecedents in what they believed to be "Arabian civilization," the highest development of which was reached in Egypt. Since the 1960s, however, and the emergence during that decade of independent black African nations, their organization gradually begun to centre around Africa.

In its ideology the Nation of Islam is political; its proclaimed objective is an "Afro-American homeland" and a forthcoming "Black Nation." In practice, however, it shuns anything political, because it holds that Allah himself (who is both a religious and political leader) will bring about the New World and the New Islam—the religion of the "Black Nation." While it awaits the advent of this New World, Elijah Muhammad, the Messenger of Allah, acted as its political and religious leader.

Elijah Muhammad made his appeal primarily to urban, lower class American blacks, who were, for the most part, poorly educated migrants from rural areas of the South. Northern-born Muslims were generally better educated, though college graduates comprised only a small percentage of the movement's followers. During the early 1960s Malcolm X succeeded in attracting more college graduates to the Nation of Islam, though its additional appeal to this group was also based on the movement's shift in emphasis from a political to a more social orientation.

Before 1935 membership in the Nation was very small; its meetings were usually held in homes or rented places. Exact data on membership is difficult to secure since the figures are kept secret. The two largest memberships are at Mosque (formerly called Temple) No. 27 in Los Angeles and Mosque No. 7 in New York City. Total membership was estimated in the early 1960s at between 10,000 and 25,000. There were then between 5,000 and 15,000 registered followers, at least 50,000 believers, and a larger number of sympathizers. With the recent rise in popularity of nationalistic ideas among black Americans, however, the Nation of Islam has become more and more popular, and its membership has probably increased far beyond these early estimates. One 1970 source records claims of a membership of from 100,000 to 750,000.

EARLY HISTORY AND BACKGROUND

Noble Drew Ali's Moorish-American Science Temple

The origins of the Nation of Islam are found in the religious tradition of Noble Drew Ali's Moorish Science Temple of America (established 1913). It also embodies the secular tradition of Marcus Garvey's Universal Negro Improvement Association (founded 1914). Elijah Muhammad was himself a member of both these organizations. The philosophy of the Moorish Science Temple was based on the assertion that American blacks are "Asiatics" and, specifically, Moors whose forebears inhabited Morocco before they were enslaved in North America. In 1929, after Noble Drew Ali's death, the leadership of the movement was assumed by W.D. Fard, who reportedly claimed that he was Noble Drew Ali reincarnated. In 1930 a split developed in the movement. The faction that called itself the Moors remained faithful to the teachings of Noble Drew Ali and still exists as a small group; the other faction followed W.D. Fard, and later Elijah Muhammad, to become the Nation of Islam.

Leadership of W.D. Fard. W.D. Fard became prominent in the black nationalist movement upon the death of Prophet Noble Drew Ali. In 1930, he founded a temple in Detroit. He "proclaimed that his mission was to secure 'freedom, justice, and equality' for his 'uncle' living in the 'wilderness of North America, surrounded and robbed completely by the caveman.' " (From Arna Bontemps and Jack Conroy, *They Seek a City;* Doubleday, Doran & Co., Inc., 1945.) W.D. Fard used various names: Walli Farrad, Professor Ford, Farrad Mohammed, F. Mohammed Ali, Wallace Fard Muhammad, and even God (Allah). He is said to have peddled silks and raincoats from door to door in Chicago. After organizing the Detroit temple he gradually receded into the background, finally disappearing and leaving no trace.

Development and expansion. Fard's mysterious disappearance led some of his followers to believe that he was, truly, the "Supreme Ruler of the Universe," or "God, Allah." Not all his followers, however, believed in his divinity, and a divisive controversy developed. One result was the founding of the Chicago branch of the Nation of Islam in 1933. The Chicago Temple, led by Elijah Muhammad, severed all connection with the parent group. The nucleus of the present Nation of Islam was Elijah Muhammad, his mother, his wife, and his six children. They constituted the early membership of his first temple, which became Mosque No. 1 in Detroit.

IDEOLOGY

Social and racial aspects. The Nation of Islam insists that knowledge of one's own identity—one's self, nation, religion, and God—is indispensable to a creative life for the individual and for the group and is the true

meaning of heaven. The members believe that consciousness of self and identity remain incomplete unless complemented by knowledge of one's "enemy." The enemy of the Nation of Islam is the Caucasian race, specifically the American white man, held responsible for the moral and material conditions of black Americans. The enemy also is the disastrous effects on blacks of the white man's claim to cultural, moral, and spiritual superiority—the myth of white supremacy.

Religious aspects. Elijah Muhammad proclaimed that his mission on earth was to deliver this message to the "so-called American Negroes," who until the coming of Master Wallace Fard Muhammad had no knowledge of themselves or of their enemy. Muhammad's concern was with organizing blacks so that they may return to their own religion and their own kind in fulfillment of the covenant that he believes Allah made with their "patriarch," Abraham of the Old Testament.

W.D. Fard Muhammad is the God of the Black Muslims. He is known to his followers as "Allah (God) in the Person of Master W.F. Muhammad, to Whom all Praise is due, the Great Mahdi or Messiah, as the Christians say." He is also "the Son of Man and the Saviour." According to Muhammad's account, "Allah came to us from the Holy City Mecca, on July 4, 1930." Prophet Fard, the Mahdi or Saviour, is also believed to be the God of the Black Nation prophesied in the Bible (Rev: 18:1).

Although W.D. Fard is considered to be a black man and God, he was also said to be a very fair-skinned Arab. Nonetheless, the Black Muslims assign to him all the attributes of God. He is referred to as the "Creator of Heaven and Earth, Most Wise, All Knowing, Most Merciful, All Powerful, Finder and Life-Giver, Master of the Day of Judgment." Muhammad's followers pray to him and implore his help in everything. One of the criticisms by orthodox Muslims of Muhammad's followers is that they worship W.D. Fard rather than Allāh (God) of the "true" Islām. They contend that even the Prophet Muḥammad, founder of the religion of Islām, was merely a prophet and never claimed to be God (Allāh).

Economic aspects. The Nation of Islam stresses the importance of economic self-sufficiency for racial advancement, continuing Garvey's belief that economic self-determination was fundamental to the realization of the political objectives of black nationalism. In an "Economic Blue Print for the Black Man," Muhammad advises his followers, thus: "(1) Recognize the necessity for unity and group operation (activities). (2) Pool your resources; physically as well as financially. (3) Stop wanton criticisms of everything that is black-owned and black-operated. (4) Keep in mind—*Jealousy Destroys From Within.* (5) Observe the operations of the White Man. He is successful. He makes no excuses for his failures. He works hard—in a collective manner. You do the same. . . ."

Political or apolitical aspects. The leader of the Nation of Islam categorically stated that it is not a political movement and that the answers to the problems of black Americans do not lie in politics and the vote. Black Muslims do not participate in local or national politics. According to Muhammad, the American government is unjust and corrupt in the eyes of Allah, and therefore it is sinful for righteous Muslims to participate in its affairs. He claimed that Allah enjoins his followers not to vote. Although Elijah preached the need for a homeland for American blacks, he had no political program for the establishment of a national home.

Members of the Nation also believe that Allah forbids them to bear arms or do violence to anyone whom he has not ordered to be killed. Muhammad himself and many of his followers were arrested and jailed for refusing to serve in the armed forces. Politically and socially they have shown a tendency to isolate themselves from the American black community and from the larger American community.

The national principle is the ideological basis upon which the Nation of Islam is organized. Elijah claimed that "Allah in the Person of Master Wallace Fard Muhammad" explained to him the history and significance of the "Black Nation." The "Black Nation" consists of the entire world population of the "black, yellow, and red races." He distinguishes the Nation of Islam from the Black Nation in that the Nation of Islam is a chosen people within the "Black Nation," elected by Allah as his special instrument for the redemption of the entire Black Nation.

INSTITUTIONS AND PRACTICES

> The organization of the Nation and the symbols of authority suggest to the members a form of "private" government where they seek freedom, justice and equality, among themselves rather than in the larger American community. Consequently, members feel that they belong, both to a nation and a government of their own. . . . The nation has its own flag and its own corps of private police who maintain discipline within the [organization]. (From E.U. Essien-Udom, *Black Nationalism: A Search for an Identity in America;* University of Chicago Press, 1962.)

The Nation is oligarchic and highly disciplined.

Organizational structure. Authority in all matters of ideology, theology, and policy resided solely in Elijah Muhammad. He was the only leader of the Nation. National officers consist of two Supreme Captains—one male, the other female—and a National Secretary. The male Supreme Captain was Elijah Muhammad's son-in-law and the female his daughter.

Temples. The basic organizational unit is the temple or mosque. Each mosque is autonomous. Mosques are theoretically equal, and each through its minister has direct access to Elijah Muhammad. He formerly appointed all his national officers and ministers. In later years, however, due to Muhammad's ill health, much of this authority was assumed by the National Officers. The most important officers of a mosque, after the Ministers, are the Captains, Secretary, Treasurer, and the Investigator. Mosque captains are assisted by officers ranked as first, second, and third lieutenants.

Organizations. Within each mosque there is a men's organization, the Fruits of Islam (FOI), and a women's organization, the Muslim Girls Training and General Civilization Class (MGT-GCC). Only members of the Fruits of Islam and the Muslim Girls Training classes are eligible to become officers of a mosque. The FOI is described as a "military" organization but members neither bear arms nor have military objectives. It does have some features characteristic of military training in its organization, discipline, and ranking officers. Members of the FOI enjoy a certain amount of prestige and regard themselves as protectors of the interest of the Nation and especially of the dignity of black womanhood. Officers include the Captain, First, Second, and Third Lieutenants.

Fruits of Islam and women's programs

According to members of the Nation, the Muslim Girls Training and General Civilization Class (MGT-GCC) is "the name given to the training of women and girls in North America how to keep house, sew, cook, and in general, how to act at home and abroad." The MGT-GCC is organized along the same lines as the FOI. Among the women, hygiene and personal cleanliness are emphasized. The women are taught reading and writing (especially those who did not receive formal education), history, and domestic science—which includes sewing, cooking, housekeeping, and child care. In general, both the FOI and MGT-GCC serve the adult education and initiation polices of the Nation. Members of these organizations are considered servants of the Nation.

Finance. The income of the Nation is kept secret. Some of its sources of income are known, however. These sources include contributions by members and by other interested individuals. Most of the funds come from voluntary contributions of members. Each member is required to pay a percentage of his income into the central treasury of the mosque. This fee is called duty. Additional funds are derived from public functions such as bazaars, dinners, and collections at public meetings. Another source of income is from sale of the Nation's weekly newspaper, *Muhammad Speaks Newspaper.*

National Convention. Each year a national convention, called the Muslim Annual Convention, is held on February 27. Members of all the mosques attend these

conventions, and some of the meetings are open to the press and the general public. During the convention business meetings are held and members and officials of the various mosques renew old acquaintances. There are no debates on the reports given at these conventions, nor any decisions made by the delegates. The business of the Nation is conducted behind the scenes. Ministers and other mosque leaders meet in conferences with the Messenger and with the Supreme Captains.

Membership. Theoretically, black people—and all red, yellow, or brown peoples—are eligible for membership in the Nation of Islam. Various methods are used to recruit new members. It is the duty of every Muslim to invite nonmembers to the mosque meeting. Offers of jobs, free meals, and dinners at Muslims' homes are also used to attract prospective members. Many new members are gained through both the Nation's newspaper *Muhammad Speaks* and weekly radio broadcasts that can be heard in every major city in the United States. Motion pictures are occasionally shown portraying the businesses and educational activities of the Muslims, and of Elijah's or other member's visits to Africa and the Middle East. Public rallies are another means of recruiting new members.

Recruitment of members

Initiation. Initiation of new members is aimed at facilitating their withdrawal from society, reorienting their values, and maintaining discipline, cohesion, loyalty, and enthusiasm in the movement. A new member is taught to believe that his nationality is "Muslim" or "Asiatic." He is made to change his "slave name" and submit himself totally to the will of Allah and Elijah Muhammad. Lessons designed to give the initiate such new perspectives and values facilitate the aims of the movement.

Discipline. Muslims disapprove of undisciplined, spontaneous impulses. The pursuit of a "righteous" life, as prescribed by the "Laws of Islam" and by Muhammad's directives, is seen as the major purpose of individual existence. These laws and directives prohibit the following: extramarital sexual relations; use of alcohol, tobacco, and narcotics; indulging in such activities as gambling, dancing, movie going, dating, sports, long vacations from work, sleeping more than is necessary to health, quarrelling between husband and wife, lying, stealing, discourtesy (especially toward women), and insubordination to civil authority, except on the ground of religious obligation. Maintaining personal habits of cleanliness and keeping fastidious homes are moral duties. The eating of pork, corn bread, collard greens, and other foods traditional among American blacks is forbidden. Violation of any of these or other rules is punished immediately by suspension from the movement for periods ranging from 30 days to 7 years. The most important sanctions that appear to regulate the behaviour of Muslims are loss of membership and fear of Allah's chastisement.

Religious ceremony and ritual. There is virtually no religious ceremony or ritual at mosque meetings, except the prayers said at the opening and closing of meetings; also there are usually verses read from the Qur'ān (Islāmic scripture) and the Bible during the minister's lecture. Members are required to study Muslim procedures for prayer, to practice them, and to say daily prayers. Elijah's followers do not worship in accordance with the prescribed traditions of Eastern Muslims. Their meetings are devoted to lectures. All persons entering a mosque for meetings, including members, officers, and visitors, are searched. Men are searched by FOI members, women by MGT members. Male and female visitors sit separately at all temple functions.

Symbols

All Muslim mosques have symbols that are used by the Ministers during the lectures. The main symbol consists of an American flag and a tree with a black man hanging from a branch, which appears on a blackboard. This symbolizes justice under the American flag. Opposite the tree is a cross, another symbol of oppression, suffering, and death. Below the cross appears the word Christianity. In the upper right hand corner the flag of the Nation of Islam is painted: the moon and stars in white, against a red background that represents the sun. The letters I.F.J. & E. are inscribed on the flag and stand for Islam (Peace), Freedom, Justice, and Equality. Directly opposite the word Christianity is inscribed Islam. Between the two flags and the names of the two religions is a large question mark with the question: "Which one Will Survive the War of Armageddon?"

Members of the Nation of Islam hold a spiritual reunion every year on February 26, the birthday of the Great Mahdi (Allah in the person of Master Wallace Fard Muhammad). Most Muslims worship at their local Mosques, but many make the "pilgrimage" to Chicago, the American Mecca. February 26 is a day of worship, contemplation, and rejoicing. Gifts and greeting cards are exchanged.

Business enterprise. Economic self-sufficiency is one of the major doctrines of the Nation. Habits of thrift and hard work are expected of members. In addition to businesses owned and operated by the various Mosques, individual members are encouraged to start businesses. Most of these businesses are small enterprises. The Nation also owns several farms, and increasingly large sums are spent in buying farmlands in the South. A major project for which the Nation has been collecting funds for several years is a multimillion dollar Islamic Center in Chicago.

Education. The Nation of Islam has always maintained a day school. The first school, called the University of Islam, was established in Detroit in 1932, the second in Chicago in 1934. Both schools have been in continuous operation since. The University of Islam in Chicago provides an elementary and secondary school curriculum and is approved by the Board of Education of Illinois. Boys and girls receive instruction in the same building but in separate classrooms, except in kindergarten and the first grade. Successful graduates have gained admission to accredited colleges and universities. Until early 1959 no tuition was charged at the University of Islam. At that time very nominal fees were introduced. Facilities at the University of Islam serve also for adult education classes that include Arabic, along with basic reading, writing, and mathematics.

Communications. Today, the Nation of Islam publishes a national weekly tabloid newspaper, *Muhammad Speaks*, which features, mainly, news of various mosques and of members. It also carries news items pertaining to the conditions of black Americans and of African and Asian countries. During the Nigerian civil war (1967–70) it was the most consistently profederalist newspaper in the United States. Each issue uses a great deal of space for the editorials of Elijah Muhammad and a documented outline entitled "What The Muslims Want" and "What the Muslims Believe." Muhammad or one of his ministers can be heard on radio stations in every major city in the United States and Canada.

NATION OF ISLAM VIS-A-VIS OTHERS

The reaction of other black Americans and leaders of civil rights groups to the Nation has been mixed. Some civil rights leaders have publicly denounced it. They generally agree with white critics that the Nation is a "black supremacy" movement. Other black Americans and leaders, however, do not appear to be as disturbed by the Nation as are "respectable Negroes." Muslims are considered merely another facet of the black community. They are generally looked upon with respect by members of the black masses, although this respect is often mingled with incredulity. The general reaction among whites who are aware of the Nation of Islam is, to say the least, unfavourable. Many have described it as a "Black Supremacy Movement," and as "anti-American," "anti-Semitic," and "anti-Christian."

The U.S. Senate Internal Security Subcommittee and the House Un-American Activities Committee have both investigated the movement. It has also been listed by the Federal Bureau of Investigation (FBI) as a movement dangerous to the security of the United States. Even though local police frequently invade mosques and harass members of the Nation, no direct conflict of any major proportions has occurred between the Muslims and federal authorities. Movements of its members are closely ob-

served by FBI agents, however. No constitutional issue has yet arisen with regard to the legitimacy of the Nation's existence.

Elijah Muhammad recognized Muslims from the Middle East, Asia, and Africa, but he believed that many Muslims living in America have forsaken the teachings of Islam. In turn, many orthodox Muslims criticized Muhammad for his racist doctrines. They state that to teach racial hatred and supremacy "is against the Qur'ān and the Prophet." In addition, Eastern Muslims claim that Muhammad's ministers teach more from the Bible than from the holy Qur'ān.

Malcolm X

The ejection of Malcolm X, one of the Nation's leading ministers, in 1964 caused a temporary setback in the popularity and prestige of the organization. Before this time the Nation's membership had grown to its highest level, due primarily to the popularity of Malcolm X. In addition, college students and middle class blacks were becoming members or sympathizers. With the assassination of Malcolm X in February 1965, the Nation of Islam lost some of its influence as a nationalist group because of the suspicion that the Muslims had some connection with his death.

In recent years, as a part of the growing nationalist tendency in black communities, the Nation did begin to reacquire some of its former popularity, but it has not regained the importance it had during the years 1960–64 under Malcolm X. The Nation once more seems to be reverting to the status of a religious cult, though with increased emphasis on economic activities.

SUMMARY AND PROSPECTS

The Nation of Islam provides for its members a community with common religious, cultural, and economic interests. Through years of persistent effort it has made a genuine contribution to both the material and spiritual rehabilitation of thousands of alienated and socially despised blacks. Significantly, it stimulated and fostered the spread of "black power" consciousness, the increasing demand among black Americans for community control, and the upsurge of group pride. Through its economic activities, the Nation may well show the way to the emergence of some form of "co-operative capitalism" among black Americans. This, rather than "black supremacy" or political action, may well become its most important contribution. Whether the Nation of Islam can survive the lifetime of its aging leader and maintain its present cohesion remain in question.

BIBLIOGRAPHY. E.D. BEYNON, "The Voodoo Cult Among Negro Migrants in Detroit," *American Journal of Sociology*, 43:894–907 (1938), a scholarly study of an early phase of the Nation of Islam; A.B. CLEAGE, *The Black Messiah* (1968), a discussion of the need for a black Christian Church with its own black messiah; E.D. CRONON, *Black Moses: The Story of Marcus Garvey and the Universal Negro Improvement Association* (1955), a good study of Garvey's public career; T. DRAPER, *The Rediscovery of Black Nationalism* (1970), a history of the roots of black nationalism that offers a critical analysis of contemporary black nationalist groups; H. EDWARDS, "Black Muslim and Negro Christian Family Relationships," *Journal of Marriage and the Family*, 30:604–611 (1968), a comparative study of families of members of the Nation of Islam with black families belonging to various black Christian Churches; E.U. ESSIEN-UDOM, *Black Nationalism: A Search for an Identity in America* (1962), a comprehensive study of the Nation of Islam; A.H. FAUSET, *Black Gods of the Metropolis* (1944), an excellent study of prophetic religious groups among urban black Americans; C.E. LINCOLN, *The Black Muslims in America* (1961), a scholarly analysis of the Nation of Islam and an assessment of the Nation's potential for social disruption; MALCOLM X (with the assistance of ALEX HALEY), *The Autobiography of Malcolm X* (1965), a personal account of the impact of the Nation of Islam on an important black nationalist leader; M. PARENTI, "The Black Muslims: From Revolution to Institution," *Social Research*, 31:175–194 (1964), an assessment of the possibilities of survival of the Nation of Islam; W.A. SHACK, "Black Muslims: A Nativistic Religious Movement Among Negro Americans," *Race*, 3:57–67 (1961), an examination of the nativistic aspects of its ideology and a comparison of the factors giving rise to the Muslim movement and those giving rise to independent religious cults in other parts of the world; L.L. TYLER, "The Protestant Ethic among the Black Muslims," *Phylon*, 27:5–14 (1966), an application of Max Weber's thesis to the Muslims.

(E.U.E.-U.)

Black Sea

The roughly oval-shaped Black Sea (Chernoye More in Russian; spelled Chernoje More in the transliteration system of the Akademiya Nauk) occupies a great basin, strategically situated at the southeastern extremity of Europe but connected to the distant waters of the Atlantic Ocean by the Bosporus (which emerges from its southwestern corner), the Sea of Marmara, the Dardanelles, the Aegean Sea, and the Mediterranean Sea. The famous peninsula of the Crimea thrusts into the sea from the north, and just to its east the narrow Kerch Strait (Kerchensky Proliv) gives onto the smaller Sea of Azov. The Black Sea coastline is otherwise fairly regular. The maximum east–west extent of the sea is 733 miles (1,180 kilometres), while the shortest distance between the tip of the Crimea and the Kerempe Burmi Cape to the south is 163 miles (263 kilometres). The water surface area is 162,280 square miles (420,300 square kilometres), almost the size of Sweden or Morocco, and the total volume of water is 131,000 cubic miles (547,000 cubic kilometres), with a maximum depth of more than 7,250 feet (2,210 metres) in the south central sector. The shores of the Black Sea lie within the territory of the Soviet Union (on the north and east), Turkey (on the south), and Bulgaria and Romania (on the west).

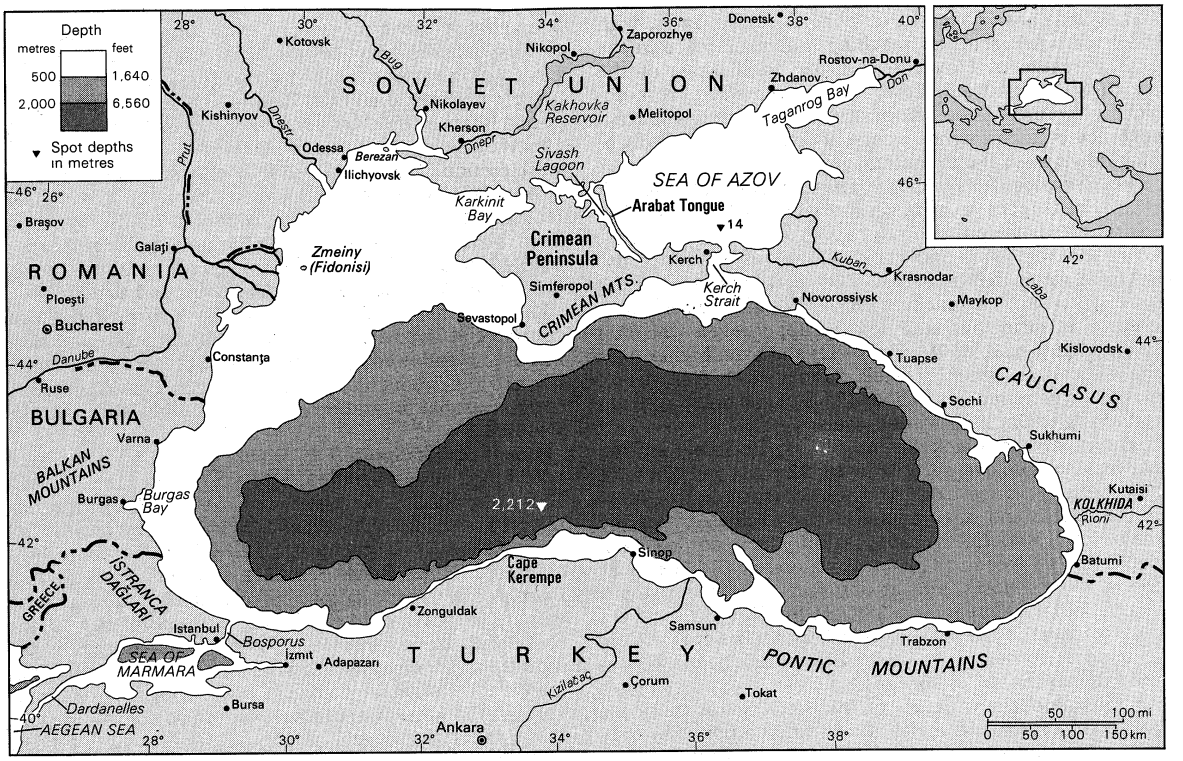
In ancient Greek myths, the sea—then on the fringe of the Mediterranean world—was named Pontus Axeinus, meaning inhospitable. Later explorations made the region more familiar, and, as colonies were established along the shores of a sea the Greeks came to know as more hospitable and friendly, its name was changed to Pontus Euxinus, the opposite of the earlier designation. It was across its waters that Jason and the Argonauts set out, according to legend, to find the Golden Fleece in the land of Colchis, a kingdom at the sea's eastern tip (now the Georgian S.S.R.), which, according to some accounts, was possibly inhabited by a black race. Yet the Turks, when they came to control the lands beyond the sea's southern shores, encountered only the sudden storms whipped up on its waters and reverted to a designation reflecting the inhospitable aspect of what they now termed the Kara Dengiz, or Black Sea.

The naming of the sea

To scientists the Black Sea is a remarkable feature because its lower levels are, to all intents and purposes, almost "dead"—not as a result of modern pollution but because of continued weak ventilation of the deep layers.

To the nations of the region, the Black Sea has been of immense strategic importance over the centuries; the advent of more settled conditions has brought its economic importance to the fore. (For further information on human geography of the Black Sea and its shores, see the articles on the nations of the region; for a related physical feature, see AZOV, SEA OF; for information on the deltas of rivers flowing into the Black Sea, see the articles DANUBE RIVER; DNESTR RIVER; DNEPR RIVER; and DON RIVER.)

Topography. *The shoreline.* In the east and south, the great mountain ranges of the Caucasus and the Pontic Mountains encircle the Black Sea in a pincer-like grip around the Kolkhida structural depression, now filled with riverborne sediments, at the sea's eastern tip. The whole region has little coastal lowland, with only the deltas of a few rivers slowly pushing out to sea. Farther west, near the Bosporus outlet, the shoreline relief, although still steep, moderates somewhat. In the southwest the smooth contours marking spurs of the Istranca (Dağları in Turkey; Strandzha in Bulgaria) ranges become northward more jagged. Farther north, in the Burgaz Bay area, low mountains emerge where the Balkan Mountains of Bulgaria extend eastward. Continuing northward, a flatter plateau region gives way to the great Danube Delta, which thrusts its mass out into the sea. Northwestern and northern shores are low and furrowed by numerous ravines and river valleys, the mouths of which are often impeded by sandy spits. The moun-



The Black Sea and associated bodies of water.

tains of the southern Crimea form the only precipitous cliff areas.

The submarine relief. The Black Sea contains but three small islands: Zmeiny (Fidonisi), Berezan, and Kefken Adasi. The submarine relief may be visualized as a series of concentric and occasionally asymmetrical rings. Beyond the shoreline a shallow shelf zone occupies about a quarter of the entire area. It is broadest in the west and at the head of Kerch Strait but elsewhere forms a rim about 6–7 miles wide, and the depth of the edge is usually less than 330–360 feet. It gives way at its edge to a slope, broken by submarine valleys and very steep in its upper parts. Between the ports of Sinop and Samsun (in the central southern sector), the coastline is paralleled by a rugged range of underwater mountains extending for nearly 100 miles. The hollow forming the basin's core covers about a third of the total area and is completely featureless flat plain, with depths increasing evenly toward the centre to a little over 7,200 feet, with the axis of maximum depth displaced toward the Turkish coast, where the maximum depth of about 7,257 feet (2,212 metres) has been recorded.

Variety of rock ages and types

Geological background. *Rock types.* Underlying rocks reflect the regional diversity of both type and age. The ancient Precambrian rocks of the southern tip of the structural block known as the Russian (or East-European) Platform, dating from at least 570,000,000 years ago, appear in the northwest. The associated Skifsky Platform has a deep cover of sedimentary rocks that were laid later. The deepwater depression, considered by some geologists as a geosyncline (or vast downwarp), has unique significance in the structure of the Earth's crust. The centre of the depression is made up of sedimentary and basaltic crustal layers respectively with a granite layer thrust between them at the periphery. Seafloor deposits generally change from coarse pebbles and gravel at the periphery to fine silts at the centre of the basin, although the progression is by no means regular.

Geological history. Much of the geological history of the Black Sea remains to be clarified, but it seems clear that it is a residual basin of the ancient Tethys Sea, dating from about 250,000,000 to 60–40,000,000 years ago. The present form of the Black Sea probably emerged at the end of the Paleocene Period, about 40,000,000 years ago, when structural upheavals in Asia Minor split off the Caspian Basin from the Mediterranean. The newly created Black Sea Basin became gradually isolated from the ocean, and its salinity was reduced; at that time the Crimea and the Caucasus were probably islands.

By the upper Miocene, about 25,000,000 years ago, the Black Sea flowed into a chain of sea lakes but gradually became separated from the Caspian region. As mountains—the Pontic, Caucasian, Crimean (Krymskye), and Carpathians—rose all around, outwashed sediments filled the basin. Further earth movements and changes in sea level associated with Ice Age glaciers then occurred and led to intermittent connections with the Mediterranean. During the last of the great glaciations the freshwater Lake Novoevksinsky was formed, and 6,000 to 8,000 years ago the present connection with the Mediterranean Sea was made. Strong earthquakes—for example, the Crimean earthquake of 1927—remain associated with the area.

Climate. On the whole, the Black Sea climate is mild, with cool summers, warm autumns, short winters, and prolonged springs, with the southern Crimea and the southeastern shores enjoying the best conditions. Conditions over the water are affected by the surrounding terrain, especially that of the dry plains of the northwest. In winter, spurs of the Siberian anticyclone (clear, dry, high-pressure air mass) create a strong current of cold air, and the northwestern Black Sea cools down considerably, with regular ice formation. The eastern portion, sheltered from cold air by the Caucasus, is milder. The winter invasion of polar continental air (which prevails for an average of 185 days annually) is accompanied by strong northeast winds, a rapid temperature drop, and frequent precipitation, with the air becoming warm and moist after passing over the milder eastern portions of the sea. Tropical air from the Mediterranean regions (87 days affected on average) is always warm and moist. Occasionally, winds that have travelled from the Atlantic across eastern Europe bring rain and sharp squalls.

The average January air temperature in the central portion of the sea is about 46° F (8° C) and decreases to 36–37° F (2–3° C) to the west. Spring air temperature everywhere approaches 61° F (16° C), rising to about 75° F (24° C) in the summer. Absolute minimums approach −22° F (−30° C) during the winter frosts in the northwest, while the Crimea may enjoy 99° F (37° C) in sum-

mer. Winds are strongest everywhere in the winter, with the cruel northeasterlies reaching hurricane force in the coastal region of Novorossiysk, just to the east of the Kerch Strait, and gale force on the sea itself. A noteworthy feature of the southern Crimean area is the occurrence of foehn winds—currents of hot, dry air dropping down from the mountains.

Periodicity in water temperatures

The hydrological system. The temperature of the Black Sea's upper layer has a marked yearly periodicity. In winter, water temperature ranges from 31° F (−0.5° C) in the northwest and about 43° F (6° C) generally, down to water depths of 230–260 feet; lower down, the water gets warmer, reaching 48.4° F (9.1° C) at the very bottom. In spring and summer, the upper levels warm up to an August peak of 79° F (26° C) and even higher near the shores; at middle levels (about 160 feet deep), a cold layer remains at 45° F (7° C), and the depths do not change from their winter levels.

Salinity of the Black Sea is almost half that of the oceans and is on average just under 22 parts per thousand (‰). Salinity varies from 17‰ on the surface to 22.3‰ in the depths and is, of course, greatly reduced near the major river mouths. A 20‰ salinity is recorded at a depth of 260 feet in the central waters, at 575 feet near the southern shores, and at 650 feet near the northern coast. Salinity increases to 38‰ at the Bosporus, where waters from the Sea of Marmara intrude. In composition, as opposed to degree, the salt of the Black Sea is almost the same as that of the oceans.

A most important feature of the Black Sea is that oxygen is dissolved (and a rich sea life made possible) only in the upper water levels. Below a depth of only 510 feet at the centre but below twice this depth near the edge, there is no oxygen, for the sea has been contaminated by hydrogen sulfide, forming a saturated, gloomy, "dead" zone frequented only by adapted bacteria.

Currents in the Black Sea are wind generated, with the main current running around counterclockwise, its branches forming gyres and sometimes large closed rotations. The current is relatively slow on the surface, except where shallow submarine relief may cause local speed-ups, but its speed is a mere inch or so a second in the depths. Flows in the Bosporus are complex, with surface Black Sea water going out and deep, saltier water coming in from the Sea of Marmara. Surface winds are an important complicating factor, especially in the shallow sill, or threshold. This situation also holds for flows to and from the Sea of Azov through the Kerch Strait.

Complex water flowage in the Bosporus

The overall water balance of the sea results from a combination of the factors of precipitation (55 cubic miles per year), inflow from the continental mass (74) and the Sea of Azov (6), surface evaporation (86), and exit through the Bosporus (50). The water level each year, therefore, varies slightly according to factors influencing any one or more of these components. Tides are virtually nonexistent, their range being exceeded by the foot or so variation induced by seiches (the changes resulting from rapid atmospheric pressure movements).

Vertical intermixing of water, except at or near the wind-whipped surface, is limited because of the compact and hence stratified nature of the sea. It has been estimated that 130 years are required to bring water in a cycle from depth to surface, although there is some limited sea-bottom turbulence caused by the warmth of the Earth's crust, and by chemical reactions in the seabed.

Plant and animal life. The Black Sea derives a Mediterranean genetic heritage from a series of invasions from that area and has Caspian elements dominant in freshwater estuaries and river mouths. In some zones, it has a rich biological productivity; and its offshoot, the Sea of Azov, is the richest in the world in this respect.

All the main groups of micro-organisms (which are predominant by a biomass, or total size, one and a half times larger than that of the groups of phytoplankton and zooplankton combined) are found in the sea. Most of them occur in a thin surface layer, with a few anaerobic bacteria in the hydrogen sulfide zone, which is otherwise lifeless. The tiny phytoplankton number some 350 species; compared to numbers in the Mediterranean, the zooplankton are even poorer, with but 80 species, including jellyfish. In coastal areas, there are eggs and larvae of invertebrates and fish. The diffusion of sea-bottom plants and animals (benthos) is four or five times poorer than in the Mediterranean, again because of the effects of the hydrogen sulfide layer. In the shallow northwest section, there is a notable extensive field of the water plant known as phillophora, whose approximately 10,000,000-ton accumulation rivals that of the Sargasso Sea. Oysters and mollusks have commercial significance in this region. The wood-boring teredo mollusk, however, is very destructive of wooden ships and structures.

There are about 180 species of fish, a fifth of them of commercial importance. The famous sturgeons are most important, followed by herring, khamsa, sprat, gray mullet, and others, including the Black Sea shark. There is some seasonal migration of fish, notably through the Bosporus.

Human utilization. The Black Sea is an important year-round transportation artery, linking the eastern European nations with world markets. Odessa, the historic Ukrainian city, together with the nearby new Soviet port of Ilichyovsk account for half the sea's freight turnover; the Soviet Black Sea fleet accounts for half that nation's sea transportation. Novorossiysk and to a lesser extent the ports of Tuapse and Batumi farther to the east specialize in petroleum. In Bulgaria, Burgas is the main export centre and Varna the main import hub. Constanţa, in Romania, connects oil-bearing regions with foreign markets. In Turkey, Istanbul is the main port, while the Danube acts as a huge trade artery for the Balkan countries.

Fish constitute the most widely used biological resource of the Black Sea, and the Soviet Union takes up fully two-thirds of the total catch. One conservationist measure has been the banning of dolphin fishing by Soviet authorities since 1966: formerly, 80,000 of these gentle creatures were being killed annually. Anti-pollution measures include restrictions on oil tankers and on the disposal of industrial wastes.

Finally, the magnificent climate and mineral springs of the regions around the Black Sea have made it the Soviet Union's major recreational and recuperative centre, with the Crimea the most important region. The Golden Sands region of Bulgaria and the Mamaia region of Romania have also attracted an increasing number of tourists, many from foreign countries, in recent decades.

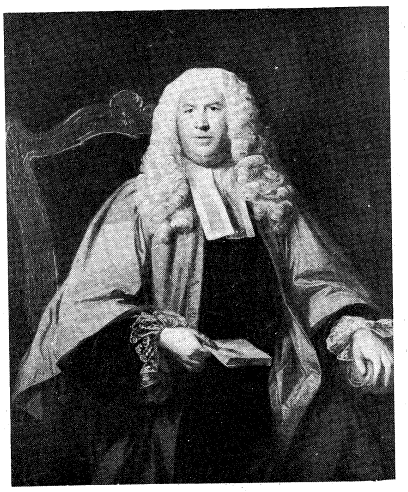
BIBLIOGRAPHY. L.A. ZENKEVICH, "The Black Sea," in *Biology of the Seas of the U.S.S.R.*, ch. 9 (1963; orig. pub. in Russian, 1963); V.P. GONCHAROV and Y.P. NEPROCHNOV, "Geomorphology of the Bottom and Tectonic Problems in the Black Sea," in the *International Dictionary of Geophysics* (1967).

(V.P.G./L.M.F.)

Blackstone, Sir William

One of the most famous of English jurists, William Blackstone was also a judge, member of Parliament, and university administrator. Blackstone's lectures at Oxford established English law, hitherto studied only in the Inns of Court, as a subject of university study alongside civil and canon law. His *Commentaries on the Laws of England*, a literary masterpiece based on his lectures, was the first complete description of English law ever written. It achieved immediate renown, shaped the future of legal education both in England and America, and is today a legal classic.

Education and early academic career. Blackstone was born in Cheapside, London, on July 10, 1723. He was the fourth and posthumous son of Charles Blackstone, a silk merchant of moderate means, and his wife Mary, daughter of Lovelace Bigg of Chilton Foliat, Wiltshire. His mother died when Blackstone was 12. He was educated by his uncle Thomas Bigg, a London surgeon, first at the Charterhouse (1730–38) and then at Pembroke College, Oxford, where he read not only the classics, but logic and mathematics. Everything that he wrote shows a wide knowledge of literature and an allusive and elegant literary style.



Blackstone, oil painting attributed to Sir Joshua Reynolds (1723–92). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

In 1741 he became a student at the Middle Temple (Inn of Court). In 1743 he was elected a member of All Souls College, Oxford; a year later he became a fellow, and by 1746 he had become a barrister. His practice was not very successful, but in college affairs he at once proved himself an active and efficient man of business, zealous for order and improvement. His greatest service was to hasten the completion and furnishing of the Codrington Library, which was finished in the summer of 1751. His persistence also brought about the building of the Wharton rooms adjoining the library. These were clearly congenial tasks: in the summer of 1743 Blackstone had written a treatise, *Elements of Architecture*, that he revised and expanded in 1746 and 1747.

In 1750 Blackstone took the degree of Doctor of Civil Law. In July 1753 he decided to retire from his practice and concentrate on the teaching of academic law and legal work in and around Oxford. He had been recorder of Wallingford since 1749 and assessor (that is, judge) of the Chancellor's Court since 1751.

Lectures on the common law

Blackstone had developed a great interest in common law, and in 1753 he began to lecture in that subject. These were the first lectures on English law ever delivered in a university. He stated his aims in a notice of his lectures dated June 23, 1753:

> It is proposed to lay down a general and comprehensive Plan of the Laws of England; to deduce their History; to enforce and illustrate their leading Rules and fundamental Principles; and to compare them with the Laws of Nature and of other Nations.

In 1756 he published *An Analysis of the Laws of England*, a synopsis of his lectures for the guidance of his pupils.

In October 1758 Blackstone was elected first holder of a chair (the Vinerian professorship) of common law. His lectures formed the basis of his *Commentaries*, which were published between 1765 and 1769. In this Oxford period of Blackstone's life he did the university another great service: appointed delegate of the university press in 1755, he rapidly achieved important reforms both in its administration and in its methods of printing. His edition (1759) of the *Magna Carta* and the *Charter of the Forest* is an example of the reformed typography. The edition led in February 1761 to his election as a fellow of the Society of Antiquaries; he read papers to the society in 1762 and 1775.

Public life. For several years Blackstone combined academic life in Oxford with increasingly active public life in London. In 1759 he resigned the office of assessor of the Chancellor's Court and in 1761 became a bencher (lecturer and supervisor) of the Middle Temple. In March 1761 he was elected member of Parliament for Hindon, in Wiltshire, on the recommendation of William Petty Fitzmaurice, 1st marquess of Lansdowne, but with no great enthusiasm on his own part. In May 1761 he married Sarah, daughter of James Clitherow of Boston House, Brentford, and bought Priory Place, Wallingford. Marriage lost him his fellowship at All Souls, but in July he was appointed principal of New Inn Hall, a medieval hall that survived until 1887, when it was annexed to Balliol College. In 1763 he became solicitor general to the queen. By this time it could not have been easy, even for a man of Blackstone's tireless industry, to combine his public life in London with his Oxford life. His decision to leave academic life came after the rejection of his proposal to convert New Inn Hall into a specialized college of common law. In 1766 Blackstone resigned both his chair and his headship of the hall.

Parliamentary career

Blackstone sat in the House of Commons for nine years as member of Parliament for Hindon (1761–68) and for Westbury (1768–70). He was not a zealous politician. He described himself as "amid the Rage of contending Parties, a Man of Moderation." But Blackstone had exhibited some previous interest in politics, for in 1751 he had taken part in a campaign to elect a country Tory, Sir Roger Newdigate, as member of Parliament for Oxford University, and in 1754 he had supported the old Tory interest in the Oxfordshire election. Blackstone was not a particularly successful politician and spoke mainly on legal and constitutional questions. The most conspicuous of these were the repeal of the Stamp Act (1766), which Blackstone opposed, and the Middlesex election of 1769, when he supported the expulsion of the political reformer John Wilkes from the House of Commons. On May 8, 1769, Blackstone gave his "unbiassed opinion" that Wilkes was disqualified from sitting in the House. This opinion was attacked as being inconsistent with the *Commentaries*. But the attack was based on genuine or pretended misunderstanding of Blackstone's position, for he did not, as he explained in 1770, maintain that expulsion from the House automatically conferred incapacity for election.

In 1770 Blackstone refused the office of solicitor general but accepted that of judge of the Common Pleas and was knighted. He was now at the height of his fame. Little is known about his extracurricular activities in the last ten years of his life. He probably had more leisure than ever before. He spent it at Wallingford, where he played a prominent part in the making of two new turnpike roads and in the repair of St. Peter's Church. The church spire, completed in 1777, was built at his expense. He urged that convicts be put to work in penitentiaries instead of being transported and helped draft the Hard Labour Bill that became law in 1779. Blackstone described this reform as "a species of Punishment in which Terror, Benevolence, and Reformation are . . . happily blended together." Toward the end of the 1770s his health failed; he died on February 14, 1780, and was buried, by his wish, in St. Peter's Church.

Assessment. Blackstone was a good judge but a better commentator. The *Commentaries* is a systematic, clear, and elegant description of the state of English law in the middle of the 18th century. It had an immediate and outstanding success. In England and America the *Commentaries* became the basis of university legal education.

Criticisms of Blackstone's views

Contemporary criticism of the *Commentaries* was directed not against the book as a whole but against particular points. His treatment of dissent was particularly criticized. Blackstone, in response, made some alterations in wording but did not change his view that dissent remained, in law, a crime. The most damaging criticism came near the end of Blackstone's life and had its greatest effect after his death. This was the philosopher Jeremy Bentham's attack on him as an "enemy of reformation."

Bentham's attack began with the anonymous publication, in 1776, of *A Fragment on Government*. It was directed against Blackstone's introduction, in which he treats government, sovereignty, and law in general. It was basically the attack of a dogmatic legal reformer upon a historical expositor of the law. Blackstone's conduct at Oxford showed that he was a defender neither of

the status quo nor of vested interests. In the House of Commons he had promoted an important reform in the law relating to the administration of assets, and in the *Commentaries* he criticized other existing laws, including the game laws, the Poor Law, and the laws of inheritance. He certainly believed that the constitution was "wisely contrived," but he was aware that it had faults. Passages in the *Commentaries* could well have been quoted in favour of parliamentary reform—the statement that there might be "a more complete representation of the people," and the disapproval of influence and of rotten boroughs, for example. But it would be absurd to expect the *Commentaries* to be primarily a plea for reform. Its purpose, like that of the lectures it is based on, is to explain and describe. Blackstone's description of the law as it existed was accurate and comprehensive and was of great use to those who wished to reform it. His description of the constitution was much more in keeping with the facts than some of his critics allowed; his statement of the sovereignty of Parliament and his recognition of the implications of sovereignty went beyond those of all predecessors.

Bentham's picture of Blackstone as an arch conservative was influential because it fitted in with the climate of early-19th-century English legal reformers. In America, where Bentham and Utilitarianism never attained much popularity, Blackstone's reputation suffered no eclipse, but in England it was, for a time, clouded. It was restored, after the middle of the 19th century, by the rise of a group of jurists who, like Blackstone, believed history to be necessary to the understanding of law.

BIBLIOGRAPHY. JAMES CLITHEROW, preface to *Reports of Cases Determined in the Several Courts of Westminster-Hall from 1746 to 1779, Taken and Compiled by Sir W. Blackstone*, 2 vol. (1781), the basic account of Blackstone's life, written by his brother-in-law; WILLIAM S. HOLDSWORTH, *A History of English Law*, vol. 12 (1938), a full account of Blackstone's career and an assessment of the *Commentaries*; *Some Makers of English Law* (1938), includes an essay on Blackstone and his influence in England and America; HAROLD G. HANBURY, *The Vinerian Chair and Legal Education* (1958), contains chapters on Blackstone, the *Commentaries*, and minor works; "Blackstone and his *Commentaries* in Retrospect," *Law Quarterly Review*, 66:318–347 (1950); LEWIS BERNSTEIN NAMIER, entry on Blackstone in *The History of Parliament: The House of Commons 1754–90*, vol. 2, ed. by NAMIER and J. BROOKE (1964); W.R. WARD, *Georgian Oxford: University Politics in the Eighteenth Century* (1958), gives an account of Blackstone's part in Oxford politics; for references to Blackstone's constitutional ideas, see J.W. GOUGH, *Fundamental Law in English Constitutional History*, 2nd ed. (1961); B. KEMP, *King and Commons, 1660–1832* (1957)); and E. BARKER, *Essays on Government*, 2nd ed. (1951).

(B.Ke.)

Blake, William

Poet, painter, engraver, and visionary, William Blake created a new, simple, and emotionally direct mode of thought and expression in the arts about 1790 and is now seen to have possessed a mind of outstanding originality and power. Yet he was ignored by the public of his day; poets and designers who did know him called him mad because he was single-minded and unworldly; he lived on the edge of poverty, and died in neglect. Only in recent times, more than 100 years after his death, has the breadth of his invention both in poetry and in painting been recognized. In addition, it is now clear that Blake had another imaginative gift: he saw with prophetic vision into the future, to the state of man in modern society.

William Blake was born in London on November 28, 1757. He was the second of five children. His father was a hosier, with a household typical of a modest shopkeeper in the late 18th century. The family lived and had their shop (which Blake's older brother later inherited) in a respectable area near Piccadilly. Although Blake's parents were Nonconformists—Protestant dissenters from the Church of England—they had the boy christened on December 11 at St. James's Church in Piccadilly. Blake did not go to school, but he learned from his mother and read widely even as a boy. Personal teaching was always to mean much to him, and he probably taught his younger brother Robert to read, and, when he married, he taught his wife.

London then was a small central city surrounded by villages and fields with market gardens. Blake was brought up in the denser part of London, and this makes his background different from that of most English poets before him, who grew up in the countryside. He walked to the villages outside London and in later life recalled the pleasure of these solitary rambles. On one of them, to Peckham Rye, before he was ten years old, he had a vision in which he saw a tree full of angels; when he told the story at home he found that his mother was kinder about it than his father.

Blake all his life had a strongly visual mind: whatever he imagined, he also saw. The accounts of those who watched him in his middle age draw imaginary personages make it plain that he had what is now called eidetic imagery—*i.e.*, the rare ability to see mental images as if they are suspended outside the head, so that they can be inspected like solid figures by shifting one's gaze from one side to the other. When Blake talked of his visions, he meant images he saw so vividly that they seemed to fill his outer rather than his inner eye. His poetry is everywhere charged and lit up with the almost physical presence of these images.



Blake, watercolour portrait by John Linnell (1792–1882). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

Training in the visual arts. At the age of ten, Blake was sent to a good drawing school, and he wanted to become a painter. His father apprenticed him to an engraver—much as today a father in a like case might urge his son to become a commercial artist. He was first taken to a fashionable engraver, William Ryland, but he did not like the man's face which, he is supposed to have said, "looks as if he will live to be hanged." Even if he did not use exactly those words, the young Blake proved to be a shrewd judge of faces, and Ryland was hanged for forgery 12 years later.

Apprenticeship

A more humdrum engraver, James Basire, was found to take Blake as an apprentice, at the age of 14, for the usual seven years. In a professional sense, the choice of Basire was not happy because the dry and formal manner of engraving that Blake learned from him was already going out of fashion and later made Blake unfashionable as an engraver. Yet to a young, bright mind like Blake's, untutored and open, working for Basire was a kind of education. At the time, Basire was making a record of the monuments in Westminster Abbey, and in the long hours Blake spent drawing in the Abbey, he acquired a sense of the presence of history and its heroes.

When Blake completed his apprenticeship in 1779, he enrolled as a student in the Royal Academy. He was an unhappy student; he took a strong dislike to Sir Joshua

Reynolds, who was president of the academy, and he felt that his talents were being wasted in the setting of the academy. He made his living by engraving for publishers and painting watercolours. By this time, after 1778, he had become friendly with a number of people who belonged to a new intellectual class then growing in England. Unorthodox in their religious views, some of them, such as John Flaxman, a fellow artist, had joined the mystical sect of Swedenborg; but most of them were Unitarians and rationalists like Joseph Priestley. Their politics were liberal and sometimes even republican. They thought that the actions of their government against the American Colonies, both before and after the Revolution, were tyrannical. They were much concerned with education, but not as it was traditionally practiced in the endowed schools (in which the chief justice, Lord Eldon, ruled as late as 1805 that modern languages were not admissible subjects). The schools which they supported were the so-called dissenting academies, in which the sons of manufacturers in the north of England were for the first time taught contemporary subjects, including science. Their outlook was, then, practical, modern, and Quakerish; and they despised the establishment as a fortress of out-of-date privilege.

Friendship with England's new intellectuals

Blake met some of these people regularly at the house of the Rev. Anthony Mathew and his learned wife, and later, when revolutionary fervour ran high, at the house of Joseph Johnson, a radical publisher. They were the only people who ever took an interest in getting Blake's poems into print. Blake had begun to write poetry when still a boy; and Mathew and his wife, with help from John Flaxman, in 1783 published a small book of these poems under the title *Poetical Sketches By W.B.* The poems are described in the preface as "the production of untutored youth, commenced in his twelfth, and occasionally resumed by the author till his twentieth year."

Blake married Catherine Sophia Boucher on August 18, 1782, and they set up house near Leicester Square. Though she was illiterate and signed the register with a cross, Blake's loving tutelage made her into a companion with whom he could read Milton's *Paradise Lost* (naked in the garden, to represent Eden) and who would share the work of printing and colouring the books of his poems that he thereafter engraved himself. In 1784 his father died, and William Blake moved next door to his boyhood home and opened a printshop with a fellow engraver, James Parker, who had been with him at Basire's. At the same time he brought his younger brother, Robert, to live with them. Blake and his wife never had children, and Robert long took the place of a child in Blake's thought. After Robert died in 1787, Blake often saw him in his dreams. The death of his brother and the failure of the printshop in the same year always hung together in Blake's memory.

Career as an engraver

Nevertheless, Blake seems to have been able to make a fair living as an engraver in those years and to win the regard of a small group of artists—chiefly John Flaxman and Henry Fuseli—who respected his direct and single-minded way of doing whatever he set his hand to. In 1784 he had begun to write poetry again, in the form of a satire called *An Island in the Moon*, which he did not finish; but snatches from it began to form some of the themes for poems that he would put together in 1789 in *Songs of Innocence*. Blake engraved the text himself on small copperplates by a method of his own, with his own decorations, and printed and coloured them by hand. There are some grounds for believing that earlier, at Mrs. Mathew's, he had sung some of his poems to tunes of his own creation.

Songs of Innocence and Songs of Experience. *Songs of Innocence* and its sequel *Songs of Experience*, which Blake added in 1794, were epoch-making works, though they had no impact at all when Blake printed them. They were virtually unread and unknown until at least 50 years after Blake died. Nonetheless, they were as formative for the culture of the 20th century in Europe and America as the Bible and *The Pilgrim's Progress* had been for an earlier age.

It is not clear why *Songs of Innocence* did not touch off the romantic storm that was to come so soon after, when Wordsworth and Coleridge published the *Lyrical Ballads* in 1798. One obvious explanation is that Blake's books were not sold or reviewed in the ordinary way. But there would seem to be more to his neglect than either the mechanics of publication or his other eccentricities. Blake's failure to fire his contemporaries on the one hand and his striking power to fire the 20th century on the other are more than matters of the form and scope of romantic poetry. It is Blake's insight into the dilemmas of civilized city life and into the hierarchies of power among warring states and industrial societies that left his contemporaries cold and seems to speak so directly to today.

The *Songs of Innocence* and *Songs of Experience* were written on the simple model of Charles Wesley's hymns and the moral verses of Isaac Watts and other respectable Nonconformists. But Blake's works reverse the roles of those didactic fables; in them it is not the poet who instructs the child, but the child who teaches the poet. In *Songs of Innocence* the child recounts the pleasures of a life in nature; in *Songs of Experience* the reader is shown children trapped and bewildered in the prisons of state and church.

Blake was 35 or more when he engraved the *Songs of Experience;* he had turned out, after all, not to be a precocious poet; and his outlook, though it had been foreshadowed in earlier work, was not fully formed until he was past 30. In part this is because his outlook was complex and coupled elements that more often than not are found on opposite sides of an argument.

Intellectual temperament

In the first place, Blake was a revolutionary. He was 18 when the Declaration of Independence inspired not only Americans but idealists everywhere in Europe. He was 22, and was present, when a rioting crowd expressed its general hatred of authority by burning Newgate Prison on June 6, 1780. Blake sympathized with the French Revolution in 1789, raged against the attacks on Tom Paine in 1798, and in 1801 still dreamed "that France & England will henceforth be as One Country."

Second, Blake belonged by temperament and sympathy to a Puritan tradition that took the Sermon on the Mount literally as its "Everlasting Gospel"—a name Blake was to use for the fragments of a long poem that he worked on around 1818. Blake was therefore opposed to private property, to any established church, to formal government and the laws of his time, to the machinery of war or any other kind of machinery which, in his phrase, keeps men working to "polish brass & iron hour after hour, laborious task, kept ignorant of its use."

Third, Blake did not distinguish between political and religious dissent. As he grew older, and particularly after a short attempt to live away from London, he increasingly used the language of religion. But what he wrote about was always human and social justice, and he identified state and church equally as prisons in which their subjects suffered physically and spiritually together. Both are identified in his books with the Old Testament Yahweh, so that Blake's form of Christianity was in fact the so-called gnostic heresy which elevated Jesus against Yahweh. Blake's combination of political radicalism with Christian Nonconformity is still an important element in the politics of the left in England.

Fourth, Blake was unusually perceptive in realizing that mechanical labour can reduce man to the status of an animal in a mill and in seeing that industry might produce the same result if it is directed only toward economic gain. Blake had worked endless hours to earn a living as an engraver in the difficult times from 1793 onward, and the experience had sharpened his vision of the threat of industrial exploitation to modern society. He was a man of remarkable character: intelligent and acute in observation, with a sweeping, imaginative grasp of the implications of his own experience, entirely upright in what he said and did, and with an unforced, natural acceptance of the remarkable gifts of language and design that he was endowed with.

From 1789 to 1794 Blake was writing for *Songs of Experience* lyrical poems of the kind he had put into *Songs*

of Innocence. Though the mood of the later is different from that of the earlier work and is bitter, the manner is the same. The disillusion of the *Songs of Experience* was a part of a general despondency among liberals in England in those years during which the increasingly violent French Revolution lost public sympathy and became a bogey.

Narrative poems. Between 1790 and 1793, Blake also wrote a book in prose, *The Marriage of Heaven and Hell*. He chose the title deliberately to contradict the Swedenborgian division between the two symbolic states, which did not seem to him to go beyond conventional theology. This remarkable book of fewer than 30 pages includes a series of "Proverbs of Hell," which spell out in simple images the themes that Blake was developing at the same time in the *Songs of Experience*. One of the most searching and troubling of these is the theme of "The Tyger": *i.e.*, the understanding that energy, enthusiasm, commitment, and a willingness to fight for a cause are all parts of the divine scheme and do not necessarily contradict the harmony of nature as it is expressed in the meekness of other creatures. In *Songs of Innocence*, the symbol of Christ had been the lamb; after *Songs of Experience*, Blake's thought returned constantly to more active symbols such as the tiger—a creature who has been designed by God to express himself fiercely. So "The Proverbs of Hell" say "The wrath of the lion is the wisdom of God," and "The tygers of wrath are wiser than the horses of instruction." In such opinions Blake tried to heal and even to turn upside down the division between the good or angelic and the bad or Satanic. In this period, the word devil was a mark of approbation in Blake and the word angel a mark of disrespect, for the climate of reaction then settled in England had made "angel" a synonym for "hypocrite" in Blake's eyes.

Composition of *The Marriage of Heaven and Hell*

During this time Blake began to write a different kind of poetry, one which is not lyrical in manner but has instead a narrative form in long, unrhymed lines with the rhythm of the King James Version of the Old Testament. Blake later used this "prophetic" manner in all his long poems, and it has come to be taken as his personal style. An early work in this style was the poem, *The French Revolution*, which was to have been printed by Joseph Johnson, who was then publishing the works of the radicals, Paine, Mary Wollstonecraft, and William Godwin. The book, however, was not issued, and the single copy that survives, dated 1791, is evidently a proof. There is no trace of six other parts that are implied by the title page. The poem is more directly political than anything else that Blake wrote, and it may be that Johnson did not issue it for the same reason that made him in 1791 turn over the second part of Paine's *Rights of Man* to another publisher (who was in fact prosecuted for it). Prosecution for sedition had become a constant threat to authors, printers, and booksellers in England and imposed an effective though silent censorship.

Narrative form of *The French Revolution*

Blake continued to engrave a number of books himself that voice his opposition to those who feared every kind of change. For a time he harked back to themes from the American Revolution. But from 1795 the world of his books becomes imaginary, and the wars that rage in them between authority and liberation are fables, written in a symbolism in which even the names of the protagonists are inventions.

The atmosphere of England at war with France was difficult for Blake: he was out of sympathy with the turn in public opinion, and in London he was surely known as an opponent of the government. Further, the inflation and economic depression of the war years hit Blake hard because he depended for his livelihood on designing and engraving, which in such times were luxury crafts with failing prospects. Blake and his wife had moved to a more modest part of London: south of the Thames. In 1796 a bookseller commissioned Blake to illustrate the *Night Thoughts* of Edward Young, and the first part was printed in 1797 but was a failure. Though the economic crisis was in part to blame, it is also clear that Blake, though he was acknowledged by his friends to be a highly original artist, was neither a man to the public taste nor one who could compromise with it. He scraped along from this time with the help of patronage from Thomas Butts, a Swedenborgian who bought drawings and small paintings from him for nearly 30 years.

In 1800 John Flaxman found work for Blake with William Hayley, a man of letters and of some means, who lived near Felpham on the Sussex coast. When the work seemed likely to grow, Blake moved to Felpham to be near Hayley. It was the only time in his life that he left London, and he stayed away for three years. Blake recognized the move as a watershed in his life: he wrote to Thomas Butts soon after he got to Felpham that "in future I am the determined advocate of Religion & Humility, the two bands of Society." Like Flaxman, Hayley was a well-wisher and a political liberal but not connected with the radical groups in London, and the move assured Blake's livelihood, gave him a friendly patron who did not think his ideas outrageous, and—quite as important—removed him from his more revolutionary friends and their persecutors in London.

Association with Hayley

The arrangement seemed idyllic, but it was one that could hardly last. Blake was not a man who could settle down in servility to Hayley, whose flat poetry and elevated sentiments became in his mind an insult to his own talent. There were small disagreements and growing friction. The country was lonely and the cottage damp, and Mrs. Blake had rheumatism. Above all, Blake could not stand what he called "the meer drudgery" of the work he was asked to do by Hayley, while his own art was neglected: several passionate letters to Thomas Butts and to his older brother are eloquent about this. Blake decided that he must move back to London, and he made plans to do so in the summer of 1803.

His stay at Felpham was, however, to end more dramatically. On the morning of August 12, 1803, a private in a troop of dragoons stationed in the village came into the cottage garden and seemed to Blake to behave offensively. Although Blake was a small man and not far short of 50, he summarily threw the soldier out of the garden. Thereupon the soldier swore a complaint, supported by a second soldier, that Blake had made seditious statements about the King, the army, and the country and had assaulted him. Blake had to post bail of £250, and, on January 11, 1804, had to stand trial. He had by that time gone back to live in London and was convinced that the soldier had been sent by informers from there. Hayley stood by Blake throughout, and some of the villagers gave evidence for him. In the end, he was found not guilty, and nothing more was heard of the matter except that Blake put the soldier into his last book, *Jerusalem*, as a peculiarly devious figure of treachery.

From this time, in 1804, Blake worked at his two longest books: *Milton*, which he finished in 1808 or a little later, and *Jerusalem*, which he finished only in 1820. The vision of an ideal of human life that they offer is the same as in the *Songs of Innocence* and *Songs of Experience:* man must free himself from conventional authority, including the hidden conventions that rule family life, the hierarchies of seniority, sex, and status, and the anxiety to conform to the usual canons of success. A good society, which Blake calls the New Jerusalem, cannot be achieved by political and social reforms alone. There must be universal grace of spirit, a sense of human dignity, and a flow of respect between person and person before life can be lived in a way that is true to human nature and the natural world.

Themes of *Milton* and *Jerusalem*

A new theme now appears in the long, prophetic poems, which, as they describe the work of building the New Jerusalem, refer more often to other forms of labour, and particularly to labour at the forge and the loom. These symbols of male and female occupations, which have many overtones in Blake's philosophy of the natural world, did not crop up in his poems much before 1800; and it must be supposed that Blake had only recently begun to hear about the new industries in the north of England. Though Blake never went north of London, many of the landscapes he describes and the place names he uses come from the north; and he was evidently an acute listener and reader.

In passages about male and female labour, Blake set forth a penetrating vision of the Industrial Revolution as it entered the life of the poor: the smoke, the noise, the stupefying drudgery, the estrangement from all that is individual in a man's life. The most powerful lines on this theme stand at the beginning of *Milton*, where Blake asks whether Jerusalem was built "among these dark Satanic Mills." The word mill in Blake's usage does not yet mean a factory but a mechanism, such as an arrangement of gear wheels or other machinery. But the spirit in which the line is sung and quoted nowadays is right. With grave foresight, Blake saw in his mind's eye that men were being chained economically to machines, and he knew prophetically what damage these abuses would work, physically and spiritually.

Last years. In 1809 Blake made a last effort to put his work before the public and held an exhibition of 16 paintings and watercolour drawings. One of the paintings shown was based on Chaucer's Canterbury pilgrims; it may have prompted Blake to mount the exhibition because he felt he had been cheated by a fellow artist, Thomas Stothard, who had been commissioned to make a rival painting of the same subject. Blake wrote a thoughtful and pugnacious *Descriptive Catalogue* for the exhibition, but only a few people attended, among them Charles Lamb.

Last exhibition

Blake found it difficult to get work now, and the engravings that can be identified as his from this period are often hack jobs. He was almost forgotten, even by those well read in poetry. In 1824 Charles Lamb did not even know that he was still alive. About 1818, however, some young painters who shared his religious seriousness (they called themselves "The Ancients") began to cultivate Blake. One of them, John Linnell, in 1821 commissioned him to make drawings for the Book of Job and later to engrave them; and in 1825 commissioned him to make watercolour drawings for Dante's *Divine Comedy*. In his 60s, Blake thus, for the first time, found a following and support for the imaginative work that he had longed to do all his life. As a result, it was in his last years that he produced his most assured and beautiful designs. It must be added, however, that when his engravings from Job were published in 1826, the book once again failed.

Toward the end of his life Blake suffered much from gallstones and wrote no more, though marginal notes he made here and there had his old ironic flash of contempt for state institutions and pious conformity. He still coloured copies of his books in bed, and that is how he died in a room off the Strand in his 70th year, on August 12, 1827. He was buried in an unmarked grave in Bunhill Fields.

Long after his death Blake was appreciated only by fellow artists and poets. Dante Gabriel Rossetti, himself a painter and a poet, bought a sketchbook in which Blake had also written drafts and versions of poems (among them "The Everlasting Gospel") from an attendant at the British Museum in 1847 for ten shillings. He lent it to Blake's first biographer, Alexander Gilchrist, in 1861 and to Swinburne for his long *Critical Essay* on Blake in 1868. William Butler Yeats helped to prepare and annotate a major edition of Blake's works in 1893. In 1920 T.S. Eliot included in *The Sacred Wood* an essay on Blake that made his name familiar to a new reading public after World War I. The centenary of his death in 1927 was marked by a number of articles and books, and thus became the turning point at which William Blake's modern reputation was established.

MAJOR WORKS

ILLUMINATED BOOKS: Relief etching on copperplates afterward tinted with watercolours—*There Is No Natural Religion*, 2 series, 19 small plates (unfinished); *All Religions Are One*, 10 small plates (*c.* 1788), prose aphorisms; *Songs of Innocence*, 31 plates (1789), lyrical poems; *The Book of Thel*, 8 plates (1789), allegorical narrative poem; *The Marriage of Heaven and Hell*, 27 plates (1793), prose manifestos, including "The Proverbs of Hell," paradoxical aphorisms, and "A Song of Liberty"; *Visions of the Daughters of Albion*, 11 plates (1793), and *America: A Prophecy*, 18 plates (1793), symbolical political poems: *Europe: A Prophecy*, 18 plates (1794), historical and political parable; *Songs of Innocence and of Experience, shewing the Two Contrary States of the Human Soul*, 54 plates (1794–*c.* 1801); *The Book of Urizen*, 28 plates (1794), in early copies *The First Book of Urizen*, first book of Blake's cosmic epic, or "Bible of Hell"; *The Book of Ahania*, 6 plates (1795), second book of the cosmic epic; *The Book of Los*, 5 plates (1795), retelling the myth of Urizen from Los's viewpoint; *The Song of Los*, 8 plates (1795), continues the myth of Urizen; *Milton: A Poem in 2 Books*, 50 plates (1804–08), epic poem, including the lyric popularly but erroneously known as "Jerusalem"; *Jerusalem*, 100 plates (*c.* 1804–*c.* 1820), epic poem.

OTHER WRITINGS (POETRY): *Poetical Sketches by W. B.* (1783); *Tiriel* (written *c.* 1789, first printed 1874), first of the so-called prophetic books; *Vala or the Four Zoas*, (1795–*c.*1804, printed 1925 unfinished), includes 9 full-page drawings and many marginal illustrations; *The Everlasting Gospel* (*c.* 1818, printed 1925 unfinished). A notebook used by Blake as a commonplace book (*c.* 1793–1810), and later called the Rossetti Manuscript (first printed in full, 1957), contains lyrical poems, epigrams, and fragments; a manuscript of *c.* 1803 (later called the Pickering Manuscript) contains 10 poems printed 1905, among them "The Smile," "The Golden Net," "The Mental Traveller," The Land of Dreams," "The Crystal Cabinet," and "Auguries of Innocence." (PROSE): *A Descriptive Catalogue* (1809), for an exhibition of Blake's watercolours and drawings—art criticism and aesthetic theory; *Public Address*, a critical essay (written *c.* 1810, printed 1925); *A Vision of the Last Judgement* (written *c.* 1810, printed 1925), essay on a lost tempera print given its present title by Dante Gabriel Rossetti; aesthetic theory and criticism.

COLOUR PRINTS OR MONOTYPES: "The Ancient of Days" (1794; Whitworth Art Gallery, University of Manchester, England); a series of 12 large prints of "Historical and Poetical Subjects" (1795; Tate Gallery, London), including "Elohim Creating Adam" (called by Blake "God Creating Adam"), "God Calling Adam," "The Lazar House," "Nebuchadnezzar," "Hecate," "Pity," "Newton," and "Good and Evil Angels Struggling for Possession of a Child"; "Glad Day" (after 1800; British Museum), from a drawing of 1780 (Victoria and Albert Museum, London).

ILLUSTRATIONS (BIBLICAL): Two series for the whole Bible, one in tempera or fresco (1799–1800; 37 paintings); the other in watercolour (*c.* 1800–05; more than 80 paintings—some lost). The tempera series includes "Bathsheba at the Bath" (Tate Gallery), "The Nativity" (Sydney Morse Collection), "The Procession from Calvary" (Tate Gallery), and "Christ the Mediator" (George Goyder Collection); the watercolour series includes "Jacob's Ladder" (British Museum), "David and Goliath" (Museum of Fine Arts, Boston), "The Stoning of Achan" (Tate Gallery), "The Soldiers Casting Lots for Christ's Raiment" (Fitzwilliam Museum, Cambridge, Cambridgeshire), "Christ in the Sepulchre" (Sydney Morse Collection), and "The Red Dragon and the Woman Clothed with the Sun" (Brooklyn Museum, Brooklyn, N.Y.). *The Book of Job*, 22 watercolour paintings (*c.* 1818–21, engraved 1823–25). (FOR MILTON'S POEMS): Two series of watercolour illustrations: for *Paradise Lost* (1808; Museum of Fine Arts, Boston), including "The Downfall of the Rebel Angels," "God Creating Eve," "Adam and Eve Asleep," "The Temptation of Eve," and "The Prophecy of the Crucifixion"; and for "L'Allegro" and "Il Penseroso" (*c.* 1818; Pierpont Morgan Library, New York). (FOR DANTE'S DIVINE COMEDY): 102 watercolours (1825–27), including "Inscription over Hell-Gate" (Tate Gallery), "Antaeus" (National Gallery, Melbourne), "The Simoniac Pope" (Tate Gallery), "The Circle of the Lustful" (City Museum and Art Gallery, Birmingham, England), "Lucia Carrying Dante in his Sleep" (Fogg Art Museum, Harvard University), "Beatrice Addressing Dante from the Car" (Tate Gallery), and "Dante and Virgil with SS. Peter, James and John" (British Museum). (OTHER ILLUSTRATIONS): Designs for engravings for a translation of G.A. Bürger's *Leonora* (1796); 12 designs for engravings for Robert Blair's poem *The Grave* (1805); watercolour (fresco) of *The Canterbury Pilgrims* (1808; Pollok House Collection, Glasgow; engraved 1810); wood engravings for R.J. Thornton's *Pastorals of Virgil*, the First Eclogue (1820–21).

BIBLIOGRAPHY. GERALD EADES BENTLEY and MARTIN K. NURMI, *A Blake Bibliography: Annotated Lists of Works, Studies and Blakeana* (1964).

Editions and selections: E.J. ELLIS and W.B. YEATS (eds.), *The Poetical Works of William Blake*, 3 vol. (1893); J. SAMPSON (ed.), *The Poetical Works of William Blake* (1913); GEOFFREY KEYNES (ed.), *The Complete Writings of William Blake* (1957) and *The Letters of William Blake*, 2nd ed. rev. (1968); JACOB BRONOWSKI (ed.), *A Selection of Poems and Letters* (1958); DAVID V. ERDMAN (ed.) and HAROLD BLOOM

(commentary), *The Poetry and Prose of William Blake* (1965).

Reproductions: GEOFFREY KEYNES (ed.), *The Note-Book of William Blake, Called the Rossetti Manuscript* (1935); ALBERT S. ROE (ed.), *Blake's Illustrations to the Divine Comedy* (1953); SAMUEL FOSTER DAMON, *Blake's Job: William Blake's Illustrations of the Book of Job* (1966); WILLIAM BLAKE TRUST, facsimile editions of various *Illuminated Books* (1951–69).

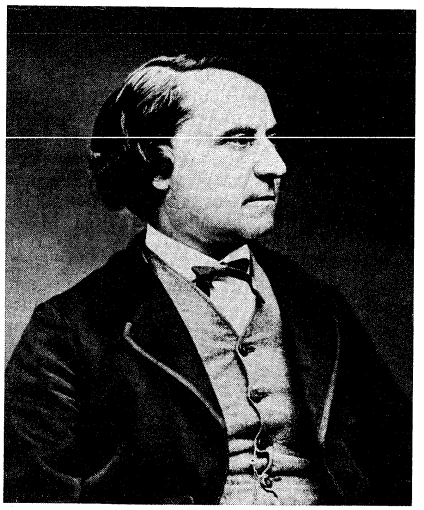
Biography and criticism: A. GILCHRIST, *Life of William Blake* (1863; rev. ed. by R. TODD, 1945); A.C. SWINBURNE, *William Blake: A Critical Essay* (1868; with illustrations from Blake's designs in facsimile, 1967); J.H. WICKSTEED, *Blake's Vision of the Book of Job,* (1910; rev. ed., 1925); SAMUEL FOSTER DAMON, *William Blake: His Philosophy and Symbols* (1924 and 1947); MONA WILSON, *The Life of William Blake* (1927; reset with additional notes and revisions, 1948); JACOB BRONOWSKI, *William Blake, 1757–1827: A Man Without a Mask* (1943; revised and enlarged as *William Blake and the Age of Revolution,* 1965); MARK SCHORER, *William Blake: The Politics of Vision* (1946); NORTHROP FRYE, *Fearful Symmetry: A Study of William Blake* (1947); DAVID V. ERDMAN, *Blake, Prophet Against Empire: A Poet's Interpretation of the History of His Own Times,* (1954; rev. ed., 1969); ARTHUR LESLIE MORTON, *The Everlasting Gospel: A Study in the Sources of William Blake* (1958); JOHN BEER, *Blake's Humanism* (1968); KATHLEEN J. RAINE, *Blake and Tradition,* 2 vol. (1969).

(J.Br.)

Blanc, Louis

Louis Blanc was a French utopian socialist of the mid-19th century whose writings had a powerful influence on radical thought. He proposed that the government provide jobs for workers and the social ideal: "from each according to his ability, to each according to his needs."

H. Roger-Viollet



Blanc.

Journalism. Jean-Joseph-Charles-Louis Blanc was born on October 29, 1811, in Madrid, where his father was inspector general of finances in the Spanish regime of Joseph Bonaparte. When that regime collapsed in 1813, the Blancs returned to France. Louis studied at schools in Rodez and Paris. He became a tutor in the family of a wealthy industrialist in northern France who employed more than 600 workers; there he came in contact with liberal political circles and found employment on a Republican newspaper. In 1837 he became a member of a committee for electoral reform directed by leaders of the opposition to King Louis-Philippe. In 1839, at the age of 26, he founded the *Revue du Progrès.* It was in this newspaper that his most important work, "L'Organisation du travail" ("The Organization of Labour"), appeared serially in 1839. The principles laid down in that essay, which first brought him to public attention, formed the basis of his subsequent career.

Blanc's theory of socialism

Blanc believed that the competitive capitalism then developing in France tended to stunt the human personality, pitting one man against another and driving the weaker to the wall. The first step toward a better society would be to guarantee work for everyone. This could be done by establishing "social workshops" financed by the state. These workshops, controlled by the workers themselves, would gradually take over most production until finally a socialist society would come into being. Blanc did not believe in human equality. But he did not agree with the followers of the socialist Saint-Simon, who held that workers should be paid according to their performance; he argued that justice would be satisfied only "when each one in accordance with the law written in some shape in his organization by God Himself, produces according to his faculties and consumes according to his wants."

Politics. In 1843 Blanc joined the committee of *La Réforme,* the journal of the extreme left-wing Republicans. In 1847 he became prominent in the so-called banquets campaign for electoral reform, holding large audiences with his oratory. The culminating banquet, arranged to take place in Paris on February 22, 1848, was banned, but a riot on the following day led to an insurrection and the fall of the monarchy. Blanc became a member of the provisional government of the Second Republic. On February 25, 1848, following a motion by Blanc, the government undertook "to guarantee the livelihood of the workers by work" and "to guarantee work for every citizen." But the government was divided. For the majority the revolution represented a political change in which a monarchy with a restricted franchise was to be replaced by a free democratic republic based upon universal suffrage; for the minority, including Blanc, it also heralded a social and economic transformation.

Although Blanc and his friends were a minority in the government, they had many supporters in the streets; and their colleagues made important concessions to their ideas by reducing working hours, proclaiming the right to work, appointing Blanc chairman of a permanent commission to investigate labour problems, and establishing national workshops to relieve the more acute unemployment. The national workshops were a travesty of those envisaged by Blanc; they were established by his opponents to discredit him and become little more than a gigantic system of outdoor relief. Meanwhile, unemployment grew from 6,100 on March 7 to 118,310 on June 15. The celebrated Luxembourg Commission, of which Blanc had been made chairman, became an arbiter in trade disputes and a centre of socialist propaganda; it was unable, however, to win acceptance of its recommendations for the reorganization of labour and industry.

Exile. Blanc was forced to flee to England after the workers unsuccessfully revolted in June 1848. He did not return to France until the fall of the Second Empire of Louis-Napoleon in 1870. He supported himself during his exile by teaching and lecturing; he wrote a history of the Revolution of 1848 and a history of the French Revolution as well and also a series of books on British political and social conditions; and he carried on polemics with other political exiles.

When he returned to France after 22 years, he was still a famous man. He was elected a deputy to the National Assembly, receiving more votes than Victor Hugo. He refused to join in the revolutionary commune that seized control of Paris in the spring of 1871, but after the commune was crushed he sought to obtain a political amnesty for the communards. He remained a man of the left, although without much following. One of his last speeches in 1881 was in support of a proposal to reduce the length of the working day. He died at Cannes on December 6, 1882.

BIBLIOGRAPHY. Louis Blanc's principal works are *Histoire de dix ans, 1830–1840,* 5 vol. (1841–44; Eng. trans., *The History of Ten Years, 1830–1840,* 2 vol., 1844–45, reprinted 1969), an attack on the government and person of King Louis-Philippe; and *Histoire de la Révolution française,* 12 vol. (1847–62), a political, economic, and social interpretation. *Historical Revelations: Inscribed to Lord Normanby* (1858, reprinted 1971), was written in English during his exile and concerns the revolution of 1848. *Lettres sur l'Angleterre,* 2 vol. (1865; Eng. trans., *Letters on England,* 2 vol., 1866), contains the author's observations on economic, poli-

tical, and social life in England. After Blanc's death, some friends collected in *Discours politiques, 1847 à 1881* (1882), the most important speeches of his long life. The earliest biography, written just after his death, was LOUIS FIAUX, *Louis Blanc* (1883), in French. The most important general study is EDOUARD RENARD, *Louis Blanc, sa vie, son oeuvre, une jeunesse pauvre* (1928). PAUL KELLER, *Louis Blanc und die Revolution von 1848* (1926); and P. VERLINDE, *L'Oeuvre économique de Louis Blanc* (1940), are narrower in scope. JEAN VIDALENC, *Louis Blanc, 1811–1822* (1948), is a short biography (in French), using the older studies and written for the centenary of the revolution of 1848.

(J.Vi.)

Blanche of Castile

The wife of Louis VIII, the mother of Louis IX (St. Louis), and twice regent of France, Blanche of Castile contributed much to the unification of French territories.

She was born in 1188 in Palencia, Spain, the daughter of Alfonso VIII of Castile and Eleanor, the daughter of Henry II of England. Her grandmother Eleanor of Aquitaine, Queen of England, travelled to Spain to take the

Marriage to Louis

11-year-old Blanche to France, where a marriage treaty was concluded with Louis, the young son of King Philip II Augustus. This politically motivated marriage had been arranged by Blanche's uncle, King John of England, and was celebrated in 1200 at Portsmouth, Hampshire. It represented only a brief truce in the struggle between England and France for control over certain French territories.

Blanche, who became French through marriage, was gradually to become French in spirit as well. Although she did not cease to be concerned for her family, among them her uncle John and his allies, her brother-in-law Ferrand of Portugal and her cousin Otto of Brunswick (later Holy Roman Emperor Otto IV), she rejoiced at the French victory over Otto and the English at Bouvines in 1214, marking the first stage of French unification, a goal for which she was constantly to strive. In the same year, she gave birth to Louis, the future king of France.

Upon John of England's death, Blanche boldly tried to seize the English throne: in 1216 Louis of France invaded England on her behalf. The English stood firm against him, and John's nine-year-old son was finally crowned Henry III.

A devout Catholic, Blanche soon became involved in what she sincerely believed to be a holy war against the heretical Cathari, a sect founded on the belief that good and evil had two separate creators, which was flourishing throughout southern France. Her husband, who became Louis VIII in 1223, took part in a crusade against the Cathari but suffered a fatal attack of dysentery upon returning to the north of France in 1226. In accordance with her husband's will, Blanche became both guardian of the 12-year-old Louis and regent of France. She zealously pressed to have Louis crowned immediately, and the coronation took place at Reims three weeks after Louis VIII's death.

Regent of France

Her most pressing problem was to deal with a rebellion of the great barons, organized by Philip Hurepel, the illegitimate son of King Philip II Augustus, and supported by King Henry III of England. In the face of such adversity, Blanche showed herself by turns a delicate diplomat, a clever negotiator, and a strong leader. Dressed in white, on a white palfrey draped in the same colour, she rode into battle at the head of her troops. After an attempted abduction of the young King, Blanche did not hesitate to replace rebel noble associates with commoners if she thought it necessary. She also created the local militias that served the kingdom so well.

Blanche was gradually able to subdue the revolt, established a new truce with England, and, in 1229, pacified the south of France by signing the Treaty of Paris with Raymond VII, Count of Toulouse. France then entered an era of peace and domestic stability, which saw the construction of many cathedrals throughout the country.

On only one occasion did Blanche fail to exhibit diplomatic conduct. In 1229 a dispute between an innkeeper and some students took place in the Latin Quarter in Paris. The police were summoned, and the students were beaten and thrown into the Seine; such intervention in the Latin Quarter, however, was contrary to the prerogatives granted to the university, and the faculty and students threatened to strike if the university's privileges were not respected. Badly advised, Blanche held firm, but the university closed its doors, and the faculty and students left Paris for the provinces and abroad. It was to take four years and the intervention of the pope before the university would return to Paris with new prerogatives, this time granted by Blanche herself.

Although Louis IX came of age on April 25, 1236, Blanche remained at his side as his most loyal and steadfast supporter. Discreet and efficient in public affairs, she lacked tact with regard to her son's private life. Although Blanche herself had chosen Margaret of Provence to be Louis's wife, she treated Margaret with great severity. In 1244, after Louis recovered from a serious illness, he and his wife, much against Blanche's wishes, made a vow to go on a crusade against the Muslims. They embarked in 1248 and once again the kingdom was entrusted to Blanche.

Informed of Louis's defeat at al-Manṣūrah, Egypt, and subsequent imprisonment, Blanche herself went to seek his ransom and that of the French Army. She petitioned her parents, her allies, and the pope himself for funds and supplies, but interest in the crusade had dwindled.

Last years

Although weakened by a heart ailment, Blanche did not neglect her obligations as a regent. Continuing to preside over council meetings, she signed laws and watched over the poor of Paris. When some of the poor were mistreated by the cathedral chapter, she herself rode, as formerly, to open the gates to their prison. On her way to the Abbey of the Lys, one of her favourite retreats, Blanche suffered an attack of the heart ailment that was to take her life. She was returned to the palace of the Louvre, dressed in a nun's habit, and laid on a bed of hay. There, after begging forgiveness of all and having received the last sacraments, she died on November 12, 1252. She was buried at Maubuisson Abbey and her heart taken to the Abbey of the Lys.

Louis IX was in Jaffa when he learned of his mother's death. The news distressed him greatly, for he was aware that he had lost not only an incomparable parent but also the strongest supporter of his kingship.

BIBLIOGRAPHY. JEAN DE JOINVILLE, *Histoire de Saint Louis* (Natalis de Wailly edition, 1874); and GUILLAUME DE NANGIS, *Chronique latine . . .*, 2 vol. (Géraud edition, 1843), two works contemporary of Blanche's reign, supply interesting opinions on her actions. The language used is sometimes archaic and difficult to comprehend. T. NISARD, *Histoire de la reine Blanche, mère de Saint Louis* (1851); and J.S. DOINEL, *Histoire de Blanche de Castille* (1870), are dated, but still remain quite interesting. ELIE BERGER, *Histoire de Blanche de Castile, reine de France* (1895), is still the most complete study; YETTE JEANDET, *Blanche de Castille, reine de l'unité française* (1965), the most recent biography. See also MARGARET WALLE LABARGE, *Life of Louis IX of France* (1968).

(Y. J.)

Blanqui, Auguste

The personification of revolutionary Communism in France, more remarkable for his revolutionary strategy than for originality of doctrine, Louis-Auguste Blanqui, the French Socialist, quickly became a legendary figure, a sort of martyr to the revolutionary cause. He spent, in fact, more than 33 of his 75 years incarcerated in nearly 30 different prisons. His disciples, the Blanquists, played an important role in the history of the workers' movement even after his death.

In relation to other Socialists, Blanqui cannot be considered either an economist or a philosopher. He was essentially a theoretician of revolution and a practitioner of insurrection. He thought that the taking of power could be the act only of a small minority. Blanqui's main idea was that there could be no Socialist transformation of society without a temporary dictatorship that would first disarm the bourgeoisie, confiscate the wealth of the church and of the large property holders, and bring the great industrial and commercial enterprises under state



Blanqui, lithograph by Menut Alophe, 1849.
By courtesy of the Bibliotheque Nationale, Paris

control. The next stage would be to establish industrial- and agricultural-production associations and develop education so as to render the people capable of organizing the country's economy to their own benefit.

Blanqui was born on February 1, 1805, in the little town of Puget-Théniers in the French Maritime Alps, where his father, Dominique Blanqui, was acting as subprefect. In 1818 he joined his elder brother, Adolphe, the future liberal economist, in Paris and studied both law and medicine until 1824. From 1827 he began taking part in the student demonstrations against the restored Bourbon monarchy, but he was disappointed by the Revolution of July 1830, which established the bourgeois monarchy of Louis-Philippe. Blanqui then began his true political career. A member of the Société des Amis du Peuple (Society of the Friends of the People), he was pursued and twice imprisoned (1831 and 1836). In these years he was much influenced by the doctrines of Filippo Buonarroti, who in 1796 had been involved in the abortive rising against the Directory government by François Noël (Gracchus) Babeuf's Société des Égaux (Society of Equals). He studied the popular insurrections of the French Revolutionary period and became increasingly convinced of the inevitability of class struggle, in which he regarded the rich as the aggressors. Blanqui was thereafter convinced that in order to establish a popular government it was absolutely necessary first to build up heavily disciplined groups of conspirators. His taste for secret societies stemmed from this conviction; he organized first the Société des Familles (Society of Families) and then the Société des Saisons (Society of the Seasons). The latter society's disastrous attempt at insurrection on May 12, 1839, was the classic prototype of the Blanquist surprise attack. Five hundred armed revolutionaries took the Hôtel de Ville (City Hall) of Paris, but, isolated from the rest of the population, they were easily defeated after two days of fighting. Blanqui escaped but was later arrested. His death sentence was commuted to life imprisonment, and he was sent to the island of Mont-Saint-Michel off the Normandy coast. After four years of solitary confinement, he was believed to be dying and was granted a formal pardon; but he was not able to leave the prison hospital at Tours until just before the Revolution of 1848.

The Blanquist insurrection of 1839

This revolution was a decisive experience for Blanqui. Returning to Paris, he founded the Société Républicaine Centrale (Central Republican Society) and urged the provisional government that had formed after Louis-Philippe's fall to pursue more Socialistic policies. Although he took an active part in the organization of workers' demonstrations, he was convinced that the people were not ready for the universal suffrage that the provisional government proposed, and he demanded the postponement of the impending elections. The election results confirmed Blanqui's apprehensions: the conservatives constituted the majority of the Constituent Assembly. Blanqui was sentenced to ten years' imprisonment for having participated, on May 15, in a popular demonstration of which he had in fact disapproved. Released in 1859, he again organized secret societies and was rearrested in 1861, remaining in prison until he escaped to Belgium in 1865. Great changes occurred in France while the man they had begun to call *l'enfermé* ("the locked-up one") was able to take no part in events. The Parisian workers were defeated on the barricades of June 1848. Louis-Napoleon executed his coup d'etat of December 2, 1851, and became, as Napoleon III, hereditary emperor of the French the following year. An unprecedented industrial growth created conditions suited for the development of a modern workers' movement. Consideration of these changes led Blanqui to study and write about political economy and Socialism; most of these works were published after his death under the title *Critique sociale*. After 1865 Blanqui often went clandestinely from Brussels to Paris, where the first Blanquist groups were being organized among students and, later, among workers. He also wrote *Instruction pour une prise d'armes* (1867–68; "Instruction for a Taking Up of Arms"), a kind of manual for urban guerrilla warfare. When the first defeats of the French Army in the Franco-German War of 1870 began to threaten Napoleon III's position, Blanqui returned to Paris. On September 4, 1870, two days after Napoleon III's surrender to the Germans, there was a bloodless revolution in Paris, as a result of which the Third Republic was proclaimed and a provisional government was formed. In this action the Blanquist groups took some part. With the German armies advancing on Paris, Blanqui showed himself a patriot as well as a revolutionary, founding both a club and a newspaper of the same extremely Jacobin name: *La Patrie en danger* ("Our Country in Danger"). He invited Parisians to unite against Germany and support the government, and he showed considerable military skill in indicating what measures should be taken for the defense of Paris. He very soon became convinced that the provisional government, fearing the populace, was failing to take adequate defense measures. Consequently, the Blanquists twice unsuccessfully attempted to overthrow the government (October 31, 1870; January 22, 1871). After the capitulation of Paris and the elections of February 8, 1871, which were won by conservatives, Blanqui retired to the country, where he was arrested on March 17 for his part in the revolt of October 31. The day after Blanqui's arrest the insurrection called the Paris Commune occurred, and the Blanquists played a very important role in it. Blanqui himself was elected president of the Commune, but the government of Adolphe Thiers refused to release him from prison. Eventually the Commune capitulated, and, in the struggle for amnesty for its adherents, Blanqui became a kind of symbol. Still in prison, he was elected deputy for Bordeaux in April 1879. His election was invalidated, but he was pardoned and set free. For two years, in spite of his advanced age, he continued as a journalist and an ardent campaign speaker in favour of Socialism. On the eve of a meeting, he was struck by apoplexy and died a few days later, on January 1, 1881. Shortly afterward, a rapproachement between the Marxists and the Blanquists resulted in the founding in 1881 of the Comité Révolutionnaire Central (Central Revolutionary Committee) and in 1898 of the Parti Socialiste Révolutionnaire (Revolutionary Socialist Party).

Revolution of September 4, 1870

BIBLIOGRAPHY. GUSTAVE GEFFROY, *L'Enfermé* (1897), an old work, but still quite useful; MAURICE DOMMANGET, *Les Idées politiques et sociales d'Auguste Blanqui* (1957), a work by a great French specialist on Blanqui, including an important bibliography; NEIL STEWART, *Blanqui* (1939); SAMUEL BERNSTEIN, *Auguste Blanqui and the Art of Insurrection* (1971), a critical biography locating Blanqui in the history of socialism and comparing him to Marx; AUGUSTE BLANQUI, *Critique sociale*, 2 vol. (1885), a collection of studies by Blanqui; *Textes choisis* (1955), with an introduction by a Soviet specialist on Blanqui, V.P. VOLGUINE.

(J.Bru.)

Bleeding and Blood Clotting

The evolution of the high-pressure blood circulation of the vertebrate animals has brought with it the risk of serious bleeding after even slight injury. Mechanisms to prevent bleeding—hemostatic mechanisms—are therefore essential to survival; and they have, in fact, evolved in parallel with the circulatory systems. In mammals, in which greater blood pressures and greater physical activity have increased the hazard, hemostatic processes are more complex and efficient than those of their amphibian ancestors. It is the mammalian, and particularly the human, system that is described in this article.

Components of hemostatic mechanism

The hemostatic mechanism has three main components, blood clotting, aggregation (collecting together) of platelets, and contraction of blood vessels, all depending on integrated and interdependent reactions. In man, defects in any of these may occur, resulting in persistent bleeding from slight injuries, or there may be overactivity, causing formation of blood clots (thrombosis) in major vessels. An understanding of hemostasis is thus of practical importance; but, though much is known, many questions are still unanswered.

When a vessel is breached, blood escapes for as long as the breach remains open and the pressure within the vessel exceeds that outside. Flow can, therefore, be stopped by closing the breach or by equalizing the pressure. The breach can be closed by contraction of the vessel wall or by the formation of a solid plug. Pressure can be equalized by a rise in external pressure due to blood trapped in the tissues (the mass of blood is called a hematoma); or by a fall in the intravascular pressure—the pressure within the vessel—caused by constriction of a supply vessel; or it may be equalized by a diversion of flow through dilated bypass vessels; or by a fall in general blood pressure, as in fainting or shock. The timing and relative importance of these events vary with the scale of the injury. Bleeding from the smallest vessels can be stopped by platelet plugs (see below); from larger ones, the cooperation of blood clotting is needed; in still larger vessels, contraction becomes of paramount importance, and the severe drop in pressure associated with shock is the last line of defense.

HEMOSTASIS

Visible events in hemostasis. Hemostasis can be studied in such vascular membranes as the hamster cheek pouch. These can be viewed with the microscope, filmed by high-speed photography, and examined by means of electron microscopy. The vessels seen are arterioles, the smallest arteries, and venules, the smallest veins, connected by capillaries, the smallest of all blood vessels. The flow through these of red and white cells and platelets is clearly visible, and these cells normally have no tendency to adhere to each other or to the lining (endothelium) of the vessels. An injury too slight to rupture a vessel may still bring about a hemostatic reaction. After the pressure of a blunt point, there may be partial vascular contraction and platelet adhesion in successive layers at the point of injury. A mass is formed that grows until it almost blocks the vessel; then it usually breaks down and re-forms, the cycle being repeated perhaps many times. Upon examination, these masses are found to consist of apparently unchanged platelets, without fibrin or other cells, and even these slight injuries are found to cause shedding of some endothelial cells and exposure of deeper tissues to which the platelets adhere.

Clot formation

If the vessel is cut so that blood escapes, the reaction is different. In muscular vessels there may be immediate contraction, but usually this is insufficient to prevent bleeding. Then a mass forms in the breach and finally stops the flow. This mass is formed of platelets in which there is a loss of granules and a closer compaction; and, between the platelets, there are bundles of fibrin fibres. These changes occur mainly near damaged collagen, the fibrous protein found in connective tissue. A day later, more fibrin can be seen, and the platelets have degenerated into an amorphous mass. Then the fibrin itself begins to be dissolved by a specific enzyme, plasmin, and is replaced by a permanent framework of new collagen and blood vessels, and healing is complete.

Platelets and their aggregation. Mammalian platelets are nonnucleated cells, roughly a third the size of red blood cells, produced by the bone marrow cells called megakaryocytes; there are about 250,000 platelets per cubic millimetre of blood. They carry enzymes and adenosine triphosphate (see below), and the vasoconstrictive substances epinephrine and 5-hydroxytryptamine. They also contain platelet factor 3, a substance active in blood clotting, and a contractile protein (thrombosthenin) that allows platelets to extend and retract long footlike projections called pseudopodia.

Platelets adhere strongly to surfaces other than that of the lining of blood vessels, such as collagen, glass, metals, and fabrics. Adherent platelets themselves become adhesive for other platelets, so that, in a flow system, a mass is built up. The propagation of this adhesiveness from one layer to the next is probably due to a mechanism causing the breakdown of internal adenosine triphosphate (ATP) to adenosine diphosphate (ADP) and the release of the latter at the platelet surface. ADP causes platelet adhesion, possibly by bonding with calcium and a plasma protein, and it also promotes the release of more ADP from the platelet interior; thus, a chain reaction captures more platelets. The process is limited by two factors. The adhesion is mechanically unable to withstand the shear forces imposed by blood flow on masses of more than a certain size, and ADP is also broken down within a few minutes by plasma enzymes. This type of aggregation is temporary and reversible.

Action of thrombin

With vascular rupture and contact of the blood with tissues outside the vessel, a new factor is introduced. This is thrombin, an enzyme produced by activation of the clotting system. Thrombin causes more permanent platelet aggregation than does ADP alone. Platelets exposed to thrombin lose their granules and release not only ADP but also platelet factor 3, which accelerates the production of more thrombin. Thrombin also attacks the precursor of fibrin—fibrinogen—trapped between platelets, to form fibrin, which consolidates the mass in a manner analogous to that of steel rods in concrete.

Blood coagulation. It is a familiar experience that shed blood soon clots. This dramatic physical change has attracted much biological interest and entire schools of investigators. Because it can be studied in the test tube, and blood can be broken up into active components that can be recombined in endless permutations, there is an enormous amount of experimental data and speculation, filling scores of books and thousands of scientific papers. Only an outline of current theory can be presented here.

Fibrin and the thrombin–fibrinogen reaction. The clotting of blood is due to the sudden appearance of fibres that entangle the blood cells. Clumps of platelets adhere to the intersections with pseudopodia extending along the fibres, and it is a normal platelet function to cause retraction of the fibrin by an active shortening of the pseudopodia. During this clot retraction, the fluid in the clot (serum) is squeezed out, leaving a solid mass of cells and fibrin. Fibrin is a tough, polymerized protein—that is, it is a complex protein made up of repeating units—and electron microscopy reveals its molecular orientation. It is formed from fibrinogen, a soluble protein of high molecular weight that is produced by the liver. Fibrinogen is converted to fibrin by thrombin, an esterase that splits the fibrinogen molecule to form one major fragment (fibrin monomer—the substance in its simple, unpolymerized form) and several smaller fragments. The monomer molecules then link together (polymerize) to form long fibres. Later, additional bonding between the units of the polymer is promoted by an enzyme known as fibrin-stabilizing factor, or factor XIII.

Thrombin does not exist in the normal circulation but is generated from a precursor, prothrombin. This is a rather stable protein formed in the liver by a process requiring vitamin K, any deficiency of this vitamin depressing production of prothrombin and of certain other clotting factors. Prothrombin is adsorbed by such inorganic gels as aluminum hydroxide, $Al(OH)_3$, or bari-

um sulfate, $BaSO_4$, a property much used in separation processes.

Activation of prothrombin. There are two physiologically relevant ways in which prothrombin can be activated. One is contact of the blood with foreign surfaces; the other is the addition of damaged tissue cells or their extracts. Contact produces clotting in four or five minutes; tissue extracts, in 12 or 15 seconds. To most workers during the first half of the twentieth century, the tissue reaction seemed the more important. A theory developed in the early 1900s postulated that the tissues release a factor ("thrombokinase" or "thromboplastin") at an injury site that activates prothrombin. Calcium is also required; and citrates and oxalates that remove ionized calcium suppress clotting until more calcium is added and are much used as anticoagulants. This theory of clotting, involving four factors and two products, is illustrated in Figure 1.

The four-factor theory was generally accepted for nearly 50 years. Then the wide use of the "prothrombin-time test" showed that other factors must be involved. In this test, tissue extract (now called thromboplastin) is added with calcium to citrated plasma (plasma—blood minus its cells—treated with citrate to prevent coagulation) and the clotting time is recorded. By the four-factor theory, because thromboplastin and calcium are in excess, and fibrinogen is seldom inadequate, a lengthening of the clotting time represents a deficiency of prothrombin. This test was useful in detecting the vitamin K deficiency of liver disease (because vitamin K is required in the manufacture of prothrombin by the liver) and in certain other situations involving vitamin K. But in some cases a lengthening of the "prothrombin" time was found not to be due to prothrombin deficiency. In one case studied, the investigator found that prothrombin-free normal plasma added to the patient's blood corrected a "prothrombin" defect, and he was able to isolate from this plasma a normal factor required for prothrombin activation that was deficient in the patient. He called this "factor V," the fifth clotting factor. By similar methods, two other factors were later discovered. As there was considerable terminological confusion at this time, an international committee ruled that each clotting factor should be identified by a Roman numeral. For chronological reasons, the factors newly recognized as concerned in the activation of prothrombin by thromboplastin became factors V, VII, and X.

Identification of clotting factors

These discoveries did not at first alter the basis of the four-factor theory; they only added accelerators to the first-stage reaction. The nature of thromboplastin itself remains obscure. It is now thought to be a mixture of substances that are active as catalysts of certain clotting reactions and another substance, known as "tissue factor," that is the true initiator of clotting by this pathway.

Contact-activated system. Preoccupation with the activity of tissue extracts distracted attention from the surface-contact reaction. Activating surfaces include glass, metals, fabrics, plastics, skin, muscle, connective tissue, and certain fatty acids. Inactive surfaces include some oils, waxes, resins, silicones, a few plastics, and endothelium, the most inert surface of all. The physicochemical properties that determine activity are not known, and neither water wettability nor surface electric charge seems to be significant. The problem is important; modern surgery requires a perfectly inactive material to make substitutes—prostheses—for heart valves and sections of blood vessel and to allow filtration of waste products from the blood outside the body.

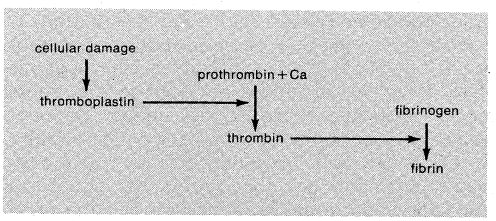


Figure 1: The classical (four-factor) theory of blood clotting.

The physiological importance of this contact system was revealed by a study of hemophilia, in which the blood clots normally with thromboplastin but abnormally in glass. The tissue-activated system in hemophilia is, therefore, apparently normal, and yet serious bleeding occurs from trivial injuries. It had long been known that a small proportion of normal plasma added to hemophilic blood corrects the clotting defect in glass, and that normal blood transfusion improves hemostasis in the patient. Investigation of this effect showed that after contact and a delay phase of three or four minutes, normal blood generates a prothrombin activator as powerful as thromboplastin. A normal factor required for this reaction was identified, and there was shown to be a deficiency of this factor in hemophilia. The element became known as factor VIII. Further investigation revealed three other necessary factors, designated IX, XI, and XII. A substance supplied by the platelets was also required. This substance is identified as phospholipid in Figure 2.

Deficiency of factor VIII in hemophilia

From R. MacFarlane, *International Review of Experimental Pathology* (1968); Academic Press

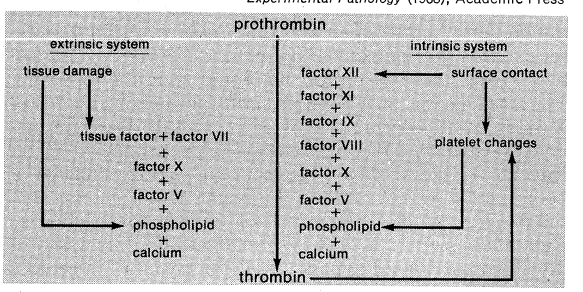


Figure 2: The clotting factors concerned in the activation of prothrombin (see text).

Thus, the factors required for prothrombin activation by the tissue-activated (extrinsic) and the contact-activated (intrinsic) systems are as shown in Figure 3. The question immediately posed is the function of this multiplicity of factors. Simultaneous interaction is improbable, and most workers agree on sequential reactions between pairs of factors. It was shown by one investigator that factor XII is the factor first activated by surface contact, becoming an enzyme that activates factor XI. Other work established the probable reaction sequence XII, XI, IX, VIII, and X, with subsequent reactions common to both extrinsic and intrinsic systems.

From R. MacFarlane, *International Review of Experimental Pathology* (1968); Academic Press

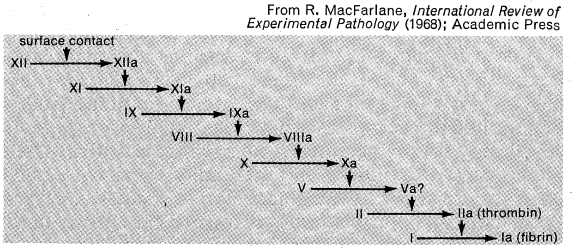


Figure 3: An enzyme cascade postulated as the basis of the intrinsic clotting system. Factor V may accelerate the action of factor Xa instead of forming an activator (Va) itself.

As regards the reactions between these factors, there is evidence that, in some factors, the activating substance is an enzyme that splits the factor X molecule to form another enzyme (Xa), which activates prothrombin (factor II) in the presence of factor V, calcium, and phospholipid. Factor X appears to be activated in this way also by both the tissue- and the contact-activated clotting systems. Thrombin was already known to be an enzyme; activated factors XII and XI have enzyme activity, and there is indirect evidence of enzyme action in other instances. On this admittedly incomplete basis, it was proposed in 1964 that the clotting mechanism might be a cascade of proenzyme enzyme transformations (Figure 3), the product of each stage activating the next. Later experiments have modified this generalization. Factor V

does not appear to be in the sequence but to act as a cofactor for the activation of prothrombin by Xa.

The extrinsic (tissue) system also activates factor X and provides an alternative pathway of more importance in nonmammalian vertebrates in which the intrinsic system with its probably greater amplification has not been developed.

From R. MacFarlane, *International Review of Experimental Pathology* (1968); Academic Press

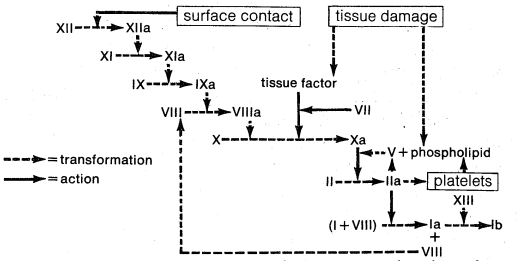


Figure 4: A scheme for the coagulation mechanism based on the cascade principle. The extrinsic and intrinsic pathways both act on factor X. Factors VII and V are shown as accelerators.

A current scheme of coagulation is shown in Figure 4. This illustrates "positive feedback" effects of thrombin, which increases the reactivity of factors VIII and V and releases platelet factor 3.

The sequential theory outlined is not accepted by all authorities. One authority considers that factors VII, IX, and X are not separate plasma entities but derivatives of the prothrombin complex formed, together with thrombin, during the process of activation. But these derivatives themselves catalyze prothrombin activation, and it is recognized that one of them ("autoprothrombin III," or factor X) is itself activated to form an enzyme ("autoprothrombin C," or Xa) that activates "prethrombin" to thrombin (Figure 5). This theory explains the otherwise

From W. Seegers, *Annual Review of Physiology* (1969); Annual Reviews, Inc. Vol. 31, p. 280

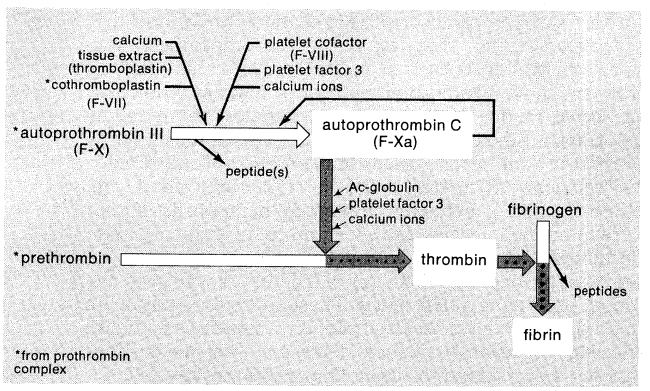


Figure 5: Basic chemical reactions: formation of autoprothrombin C, thrombin, and fibrin.

strange similarity in the properties and origin of factors VII, IX, X, and prothrombin; but it does not assign a precise function for several factors, a deficiency of which causes known clotting defects.

Inhibition of clotting. Systems involving enzyme self-activation—autocatalysis—could, if not checked, become totally activated by the most trivial stimulus. In the case of blood clotting, this could mean intravascular coagulation and immediate death; but there are inhibitory systems that destroy every activated clotting factor within a few seconds in the bloodstream so that only at sites of local production does clotting normally take place. Six different factors are described as inhibiting thrombin alone, the most important being antithrombin III, a plasma protein that combines with thrombin to form an inert complex. This action is greatly enhanced by heparin, a substance formed by certain connective tissue cells called mast cells. Activated factor X may be inactivated by the same mechanism. Thrombin itself inhibits its own production; when its concentration is sufficiently high, it destroys factor VIII.

Effect of plasmin

Another anticoagulant effect is the fibrinolytic (fibrin-splitting) action of the enzyme plasmin. Its physiological function is the removal of old fibrin at injury sites and any that may be deposited in normal vessels. The fibrinolytic and clotting mechanisms have striking similarities. Plasmin is derived from plasminogen, an inert protein precursor that can be activated by tissue factors (called lysokinases) or by contact through factor XII. An activator, urokinase, is also released by blood-vessel lining and is present in normal urine. Certain bacteria produce plasminogen activators, the streptokinase of hemolytic streptococci being the most potent. Plasmin is rapidly destroyed in the bloodstream and normally acts only when taken up by fibrin.

Vascular function. The most obvious hemostatic vascular reaction is constriction after injury. This is of great importance in large arteries, in which clotting and platelet adhesion would be unable to arrest bleeding. The survival, despite delayed surgical aid, of some persons who have lost limbs in accidents is due to constriction of their main arteries. The stimulus for contraction is not known precisely. In the capillaries, contraction has been considered to be impossible because of the absence of muscle cells from their walls, but recent studies with the electron microscope show possibly contractile elements in endothelial cells, and capillary function may have to be reassessed.

Other vascular reactions to injury have only a subsidiary hemostatic effect. Dilation of undamaged vessels in the vicinity of an injury provides an alternative pathway for the bloodstream. It is probably caused by the release of vasodilator substances, such as histamines, from damaged tissues. Another reaction is an increase in vascular permeability that allows plasma to escape through the vessel wall. This raises the tissue fluid pressure and also thickens the blood remaining in the vessel. Dilation also is caused by the action of bradykinin, a substance released from its precursor by an enzyme that is itself activated by contact through factor XII, or by tissue factors.

The relationships of these systems to hemostasis are shown in Figure 6 (below).

HEMOSTATIC DEFECTS AND ABNORMAL BLEEDING

In man, hemostatic failure may result from inherited or acquired defects in any of the clotting or platelet functions described. The usual consequence is persistent bleeding from injuries that would normally give little trouble. Some persons may bleed more easily than normal, perhaps even spontaneously, as a result of an increased fragility of the blood vessels. This fragility is not itself a hemostatic defect but is often associated with one.

Patterns of bleeding

Abnormal bleeding follows a pattern defined by the underlying defect. Platelet abnormalities are associated with spontaneous bleeding from the membranes of the nose, mouth, and gastrointestinal and urogenital tracts.

From R. MacFarlane, *International Review of Experimental Pathology* (1968); Academic Press

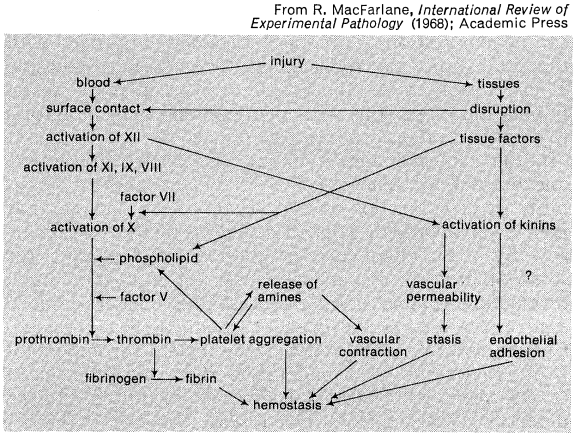


Figure 6: A comprehensive schema of hemostasis.

There is also purpura (small hemorrhages in the skin); and these hemorrhages can be produced by moderate pressure. Needle punctures in the skin also bleed longer than the normal few minutes, as in the "bleeding-time test." Massive tissue hemorrhages are rare, and the bleeding from extensive injuries, such as at operation sites, is often controllable by usual methods.

Blood clotting defects produce a different picture. Spontaneous bleeding and purpura are rare, and the effects of the bleeding-time test and of application of moderate pressure are normal. Bleeding from larger injuries, however, may be uncontrollable by any but specific replacement therapy; before this became available, many such persons died of hemorrhage following minor injury or such operations as a tooth extraction. The most characteristic feature of such defects, as typified by hemophilia, is the occurrence of massive and crippling hemorrhages into joints and muscles and bleeding into body cavities.

Vascular defects causing abnormal bleeding are rare. Groups of dilated capillaries in the skin and mucous membranes (telangiectases) may occur as an inherited or acquired condition, and these may bleed persistently if injured. There may be a capillary abnormality in von Willebrand's disease (see below). In the Ehlers-Danlos syndrome, there is an increased fragility of even major vessels, which are easily ruptured. In most cases of increased capillary fragility, there is also a platelet defect, quantitative or qualitative. This may suggest that platelets have a function in maintaining vascular integrity or that agents damaging platelets also damage vessels.

Diagnosis. The diagnosis of hemostatic defects depends on clinical and laboratory data. Investigation of the affected person's family background is important because many of these conditions are inherited. The personal history reveals the type of bleeding and the possible effects of drugs, chemicals, allergy, infection, or metabolic or dietary abnormalities. Examination reveals visible hemorrhages or vascular lesions, and the effects of internal bleeding. Clinical tests include the pressure and bleeding-time tests, already mentioned, and the collection of blood samples for laboratory testing. This includes the counting and examination of the platelets and tests for functions such as adhesiveness, response to ADP, and clot retraction. Tests of clotting function include the coagulation time in glass, the prothrombin time, and a battery of investigations of the intrinsic clotting system. These allow a deficiency of any of the clotting factors described to be identified and a quantitative assessment of activity of individual factors to be made. Most cases of abnormal bleeding can, by these means, be traced to specific defects; and precise diagnosis often allows effective treatment.

Clinical and laboratory tests

Course and outlook. The hereditary hemorrhagic disorders are life-long afflictions, and treatment, however temporarily effective, cannot alter the basic abnormality. Eventually, it may be possible to transplant into the affected person some normal organ or tissue that will provide a permanent supply of a genetically deficient factor. Until then, treatment is mainly to relieve the effects, though some prevention of the effects is possible. Even so, the expectation of life has enormously improved with the availability of replacement (transfusion) therapy. This, in itself, raises the problem of an increase in the population of persons with these genetic defects; efforts to discourage their procreation are needed.

In the acquired and secondary disorders, the course and outlook are determined by the cause and the possibility of its correction and of repair of the damage that it may have caused. Total destruction of the bone marrow by toxic chemicals, irradiation, or leukemia (or some other form of cancer) causes a failure of platelet production that is usually irreparable. Lesser damage can be repaired if the drug or chemical is eliminated. Certain cases of platelet deficiency (called thrombocytopenia) may respond permanently to removal of the spleen or treatment with adrenal hormones. Acquired clotting defects may result from liver disease, cancer, or the development of anticoagulants or excessive fibrinolysis. Again, the course and outlook are determined by the underlying condition.

Therapy. Treatment may be to relieve or prevent the manifestations and may be local or general. Local treatment consists of the application to the site of bleeding of such coagulants as thrombin or Russell's viper venom, or such vasoconstrictors as epinephrine, and the use of devices to retain dressings with firm pressure and without movement. Hemostatic dressings may be of slowly soluable materials, such as gelatine sponge. Much can be done, particularly in dental work, by these means.

General treatment consists mainly of temporary replacement therapy by transfusion. In thrombocytopenia —deficiency of platelets—transfusions of platelet-rich plasma can be given; in clotting deficiencies concentrated preparations of the missing factor can be given that will restore hemostatic function to normal for hours or days and, with repetitions, will allow injuries to heal or operations to be performed. The use of factor VIII in this way has revolutionized the treatment of hemophilia.

Hemorrhagic disorders. The hemorrhagic disorders comprise many genetic abnormalities and disease syndromes, and only a brief mention of the more important ones is possible here.

Platelet disorders. Platelets may be deficient in numbers, as the result of some identifiable disease or of injury to the platelet-producing cells of the bone marrow (secondary thrombocytopenic purpura), or may be of unknown cause (idiopathic thrombocytopenia). The platelets may be normal in number but defective in function, as in Glanzmann's and von Willebrand's diseases, both inherited as simple dominants—that is, each of these diseases becomes manifest if inherited from one parent alone. In von Willebrand's disease there is also a moderate deficiency of factor VIII in most cases, and correction of this defect may greatly improve hemostasis.

Coagulation defects. Any clotting factor (except calcium and thromboplastin) may be deficient as a hereditary or acquired defect. The commonest and most important of the hereditary disorders is hemophilia, a deficiency of factor VIII due to an abnormal gene located on the X sex chromosome. (Males have an X and a Y sex chromosome; females, two X-chromosomes.) Abnormal bleeding occurs in males carrying this gene on their single X-chromosome but not in females who carry the gene on one X-chromosome; females, though they can transmit the condition, are themselves hemostatically normal if their second X-chromosome is normal. Factor IX deficiency (Christmas disease) is a rarer condition, clinically and genetically similar to hemophilia. Other hereditary deficiencies are even more infrequent. Data on these states are given in the Table. Rarely, specific inhibitors, particularly to factor VIII, may develop in the blood, producing a deficiency. These inhibitors may arise in persons who have received blood transfusions or blood fractions, or after pregnancy. They may disappear spontaneously or respond to adrenal hormone therapy.

Excessive fibrinolysis. Excessive breakdown of fibrin (fibrinolysis) may cause bleeding, clots being rapidly digested, the blood fibrinogen becoming depleted and anticoagulant products formed. This may occur in cases of cancer, particularly of the prostate. An acute form (defibrination syndrome) may occur in certain obstetrical conditions associated with placental disease, or in injuries or operations involving lung or brain tissue, in which tissue factor may gain access to the bloodstream. Besides fibrinogen, factors V and VIII are also depleted. Infusions of fibrinogen are effective in controlling bleeding, and destruction of fibrin can be reduced by injecting inhibitors of protein breakdown.

Vascular abnormalities. In purpura, there is an unexplained and abnormal fragility of the capillaries, resulting in pinpoint hemorrhages in the skin and mucous membranes. In von Willebrand's disease capillaries may be distorted and dilated. In both these conditions, platelet defects have also been described.

Defects in capillaries

In hemorrhagic telangiectasia, visible groups of enormously dilated capillaries occur in the skin and mucous membranes of the mouth, nose, and gastrointestinal and

Recognized Coagulation Defects

deficient factor	hereditary defect		acquired defect
	probable inheritance	approximate incidence per million	
II (prothrombin)	not known	0.1	vitamin K deficiency liver disease
VII (proconvertin)	autosomal recessive	0.1	coumarin anticoagulant drugs
IX (Christmas factor; PTC)	sex-linked recessive (Christmas disease)	3–4	
X (Stuart-Prower factor)	autosomal recessive	0.1	
I (fibrinogen)	autosomal recessive	0.1	defibrination syndrome liver disease
V (proaccelerin)	autosomal recessive	0.1	defibrination syndrome
VIII (antihemophilic globulin)	sex-linked recessive (hemophilia)	30–40	defibrination syndrome circulating anticoagulant
XI (plasma thromboplastin antecedent)	autosomal dominant (PTA deficiency)	1	
XII (Hageman factor)	autosomal recessive (Hageman deficiency)	0.1	
XIII (fibrin stabilizing factor)	autosomal recessive	0.1	

respiratory tracts. The condition is inherited as a simple dominant, the lesions appearing in adult life and tending to bleed on the least provocation. Clotting and platelet functions are normal.

Thrombosis. Thrombosis, the formation of a blood clot that tends to plug in functionally normal blood vessels, is an important and increasing cause of death in Western societies. The causes of thrombosis are obscure. The commonest, and perhaps the most important cause, is a lesion that destroys the normal endothelial surface of a blood vessel. Platelets tend to adhere to such lesions, forming masses that, when reinforced by fibrin, may completely obstruct the blood flow. In arteries this obstruction may have disastrous consequences, as in coronary or cerebral vessels, or in a major artery of a limb or an organ. In veins the local results are less apparent, perhaps not detected, but, behind the original thrombus, the whole blood content of the vein may clot to form a large mass. This occurs most frequently in the leg veins and may be due to slowing of the blood flow during prolonged surgical operations or confinement to bed. The danger is that the clot may become detached and be swept into the pulmonary artery, causing fatal obstruction to the circulation. Other predisposing causes of thrombosis may be the increase in platelet numbers and adhesiveness and the increased coagulability of the blood that follows injury and blood loss.

The prevention and treatment of thrombosis aims at correcting these predisposing conditions. Specific treatment with anticoagulants includes the use of coumarin drugs, the injection of heparin, and the promotion of fibrinolysis by use of streptokinase or urokinase.

BIBLIOGRAPHY. M.P. ESNOUF and R.G. MACFARLANE, "Enzymology and the Blood Clotting Mechanism," *Advances Enzym.*, 30:255–315 (1968), a review of the biochemical basis for current theories of blood clotting, with extensive bibliography; R.G. MACFARLANE, "The Hemostatic Mechanism and its Defects," *Int. Rev. Exp. Path.*, 6:55–133 (1968), a review article on the physiology of hemostasis and the main abnormalities in man, with extensive bibliography, and (ed.), *The Haemostatic Mechanism in Man and Other Animals* (1970), an authoritative symposium covering the comparative physiology of the subject; P. MORAWITZ, "Die Chemie der Blutgerinnung," *Ergebn. Physiol.* 4:307–422 (1905), a classic contribution to the early biochemical study of blood coagulation; P.A. OWREN, "The Coagulation of Blood: Investigations on a New Clotting Factor," *Acta Med. Scand.*, suppl. 194 (1947), an important contribution marking a turning point in blood clotting research; W.H. SEEGERS, "Blood Clotting Mechanisms: Three Basic Reactions," *A. Rev. Physiol.* 31:269–294 (1969), an account of the recent theories of this authority; A.J. QUICK, "The Prothrombin in Hemophilia and in Obstructive Jaundice," *J. Biol. Chem.*, 109:LXXIII–LXXIV (1935), a brief description of a test that revolutionized blood clotting theory and practice.

(R.G.M.)

Blood, Human

Blood consists of specialized cells of several types suspended in a liquid medium, the plasma. It is well adapted for its role in the transport of many substances to and from the organs and tissues. The circulatory system provides the mechanism by which this vital function of the blood is achieved (see BLOOD CIRCULATION, HUMAN). The circulating blood continuously supplies oxygen, nutrient substances, and other materials necessary for the viability and activity of all of the cells of the body and carries away disposable cell products, including carbon dioxide and other waste materials. If blood flow ceases, death occurs within minutes because of the untoward effects of an unfavourable environment on highly susceptible cells. The relative constancy of the composition of the blood is made possible by the circulation, which conveys blood through the organs that regulate the concentration of its components. In the lungs blood acquires oxygen from the inspired air and releases carbon dioxide transported from the tissues. The kidneys remove excess water and dissolved waste products. Nutrient substances derived from food reach the bloodstream after absorption by the intestinal tract. Endocrine glands release their secretions into the blood, which transports these hormones (*e.g.*, insulin and thyroid hormone) to the tissues in which they exert their effects. Many substances are recycled through the blood; for example, iron released during the destruction of old red cells is conveyed by the plasma to sites of new red cell production where it is reutilized. Each of the numerous components of the blood is kept within appropriate concentration limits by an effective regulatory mechanism. In many instances feedback control systems are operative; thus a declining level of blood sugar leads to accelerated release of sugar into the blood so that a potentially hazardous depletion of blood sugar does not occur.

Unicellular organisms, primitive multicellular animals, and the early embryos of higher forms of life lack a circulatory system, and exchange of substances between cell and environment is accomplished by simple diffusion. In larger and more complex animals transport of adequate amounts of oxygen and other substances requires some type of blood circulation. The diffusion process then occurs between the body cells and the fluid derived from the blood, which by its constant motion maintains the constancy of the internal environment. Some simple animals, including small worms and mollusks, have blood that lacks an oxygen-binding substance analogous to hemoglobin; others are provided with pigments capable of transporting relatively large amounts of oxygen. In many invertebrates the blood pigment is dissolved in the plasma. Hemocyanin, a copper-containing protein chemically unlike hemoglobin, occurs in certain crabs and oth-

Types of blood

er lower animals. Hemocyanin is blue in colour when oxygenated and colourless when oxygen is removed. Some invertebrates have hemoglobin in solution in the plasma. In almost all vertebrates, including man, hemoglobin is contained exclusively within the corpuscles of the blood. The red cells of the lower vertebrates have a nucleus, are relatively large, and are often ovoid in shape; the cell contour is convex. In contrast, mammalian red cells lack a nucleus, are smaller, and have a biconcave shape. Red cells vary markedly in size among mammals; those of the goat are much smaller than those of man, but the goat compensates by having many more red cells per unit volume of blood. The concentration of hemoglobin inside of the red cell varies little between species (see BLOOD AND LYMPH).

Functions of the cells

The red cells permit the blood to carry sufficient oxygen to sustain life and to provide for the tremendous increase in oxygen requirement during physical exertion. Accordingly, they have an essential role in the transport function of the blood. In addition to the red cells, blood contains smaller numbers of cells of other types. The white blood cells, or leukocytes, are nucleated and larger than red cells. They are primarily concerned with defense mechanisms, preventing or coping with infection, and participating in the dissolution and removal of foreign material (phagocytosis). The smallest cells of the blood, the blood platelets, are nonnucleated. They are principally involved in maintaining the integrity of the vascular tree itself. They prevent leakage of blood from delicate capillaries, seal the walls of injured vessels, and contribute to the coagulation of the blood (see BLOOD DISEASES; BLEEDING AND BLOOD CLOTTING).

In the human embryo the primitive blood cells have their origin in the mesenchyme, the embryonic connective tissue, first appearing in the blood islands of the yolk sac. After the second month of intra-uterine life, red cell production occurs predominantly in the liver, where some white cell and platelet precursors also appear. At this time the spleen also is an active site of blood formation. In the fifth month blood formation begins in the newly developing bone marrow. At first white cell precursors predominate, but by the time of birth the bone marrow has become virtually the only source of red cells, white cells, and platelets. The lymphocytes are an exception; they arise from the thymus, spleen, and lymph nodes.

PROPERTIES AND FUNCTIONS

Properties. Blood is an opaque red fluid, freely flowing but more dense and more viscous than water. The characteristic colour is imparted by hemoglobin, a unique iron-containing protein. Hemoglobin brightens in colour when saturated with oxygen (oxyhemoglobin) and darkens when oxygen is removed (deoxyhemoglobin). For this reason, fully oxygenated blood from an artery is lighter in colour than the partially deoxygenated blood from a vein. The red cells constitute about 45 percent of the volume of the blood, and the remaining cells (the white cells and platelets) less than 1 percent. The fluid portion, the plasma, is a clear, slightly sticky, yellowish, proteinaceous liquid. After a fatty meal plasma transiently appears turbid. Within the body the blood is permanently fluid, and turbulent flow assures that cells and plasma are fairly homogeneously mixed. When blood is shed, physicochemical changes are initiated that cause the blood to coagulate (see BLEEDING AND BLOOD CLOTTING). The blood clot consists of microscopic strands of a complex protein, fibrin, forming a gel in which the blood cells are entrapped. The clot shrinks, or retracts, squeezing out an incoagulable yellowish fluid, the serum. An anticoagulant can be added to the shed blood to prevent the deposition of fibrin and to maintain the blood in a fluid state. When blood treated in this way is undisturbed, the cells gradually settle; the red cells go to the bottom, the white cells and platelets form a thin white layer (buffy coat) overlying the red cells, and the plasma appears in the upper portion of the container. The cells settle because they are denser than the plasma; specific gravity of the plasma is about 1.026, while that of red cells is about 1.093. The specific gravity of whole blood varies with the proportion of red cells that it contains; an average normal value is about 1.060. The rate at which red cells settle (sedimentation rate) is normally slow because of the small size of the red cells and their relatively large surface area. When red cells form aggregates like rolls or coins, or rouleaux, the surface–volume ratio is decreased and the sedimentation rate is accelerated. Separation of cells and plasma may be accomplished rapidly and completely by the use of a centrifuge, a device in which rapid rotation accelerates sedimentation by increasing gravitational forces. The relatively high viscosity of blood, as compared with water, is largely attributable to the suspended red cells. Viscosity increases disproportionately as the cell fraction increases. The viscous properties of blood differ from those of simple liquids in that the viscosity of blood is dependent on conditions of flow. Slowly flowing blood is more viscous than the same blood flowing through the same vessel at higher rates of flow.

Characteristics of shed blood

Separation of cells and plasma

The pH of blood is kept within narrow limits of variation at about 7.4, slightly on the alkaline side of neutrality (a pH of less than 7 indicates acidity, of more than 7 alkalinity). Venous blood is somewhat less alkaline (7.35) because of the higher carbon dioxide content. Constancy of pH is achieved by a system of efficient buffers in the blood and by the selective excretory functions of the lungs and kidneys. The rate and depth of respiration are controlled physiologically to maintain a normal tension of carbon dioxide in the blood, important in maintaining blood pH. The excretion of acid or alkaline urine by the kidneys is regulated by physiologic needs for sustaining the normal pH of the blood.

The total amount of blood varies with age, sex, weight, body build, and other factors, but a rough average figure for adults is about 7 to 8 percent of body weight. An average young man has a plasma volume of about 45 millilitres and a red cell volume of about 30 millilitres per kilogram of body weight. There is little variation in the blood volume of a healthy person over long periods, although each component of the blood is in a continuous state of flux. In particular, water moves in and out of the bloodstream with great rapidity, achieving a balance with the extravascular fluids (those outside the blood vessels) within minutes. The normal volume of blood provides such an adequate reserve that appreciable blood loss is well tolerated. Withdrawal of 500 millilitres (about a pint) of blood from normal blood donors is a harmless procedure. Blood volume is rapidly replaced after blood loss; within hours plasma volume is restored by movement of extravascular fluid into the circulation. Replacement of red cells is completed within several weeks. Tolerance to change in blood volume is made possible by the elasticity of the circulatory system. Only about 15 percent of the blood is contained in the relatively rigid arteries, while 70 percent is in the more distensible veins. The capillaries, which have a total cross-section area about 1,000 times greater than that of the aorta, contain only about 10 percent of the blood. The vast area of capillary membrane, through which water passes freely, would permit instantaneous loss of the plasma from the circulation were it not for the plasma proteins. In particular, the plasma albumin has the greatest functional significance in regulation of plasma volume. Capillary membranes are impermeable to albumin, the smallest in weight and highest in concentration of the plasma proteins. The osmotic effect of the plasma albumin retains fluid within the circulation, opposing the hydrostatic forces that tend to drive the fluid outward.

Blood volume and distribution

Functions. Broadly conceived, the function of the blood is to maintain the constancy of the internal environment, an aspect of the precise biologic regulation called homeostasis. The circulating blood makes possible man's adaptability to changing conditions of life—his endurance of wide variations of climate and atmospheric pressure; his capacity to alter his physical activity; his tolerance of changing diet and fluid intake; his resistance to physical injury, chemical poisons, and infectious

Maintaining stability of internal environment

agents. The blood has an almost unbelievably complex structure, and many components participate in its functional activities, often in an intricate and poorly understood way. Regulatory mechanisms with which the blood is involved include sensors that detect alterations in temperature, in pH, in oxygen tension, and in concentrations of the constituents of the blood. Effects of these stimuli are in some instances mediated via the nervous system or by the release of hormonal substances (chemical mediators). Some of the major functions of the blood are outlined in the paragraphs that follow.

Respiration. In terms of immediate urgency, the respiratory function of the blood is most vital. A continuous supply of oxygen is required by living cells, in particular those of the brain, and deprivation is followed in minutes by unconsciousness and death. A normal man at rest uses about 250 millilitres of oxygen per minute, a requirement increased manyfold during vigorous exertion. All of this is transported by the blood, most of it bound to the hemoglobin of the red cells. The minute blood vessels of the lungs bring the blood into close apposition with the pulmonary air spaces (alveoli), where the pressure of oxygen is relatively high.

Oxygen diffuses through the plasma and into the red cell, combining with hemoglobin, which is about 95 percent saturated with oxygen on leaving the lungs. One gram of hemoglobin can bind 1.35 millilitres of oxygen, and about 50 times as much oxygen is combined with hemoglobin as is dissolved in the plasma. In tissues where the oxygen tension is relatively low, oxygen diffuses out of the blood. Not all of the oxygen is removed, and the venous blood returning to the lungs is partially oxygenated. The added demand for oxygen during increased physical activity is met primarily by accelerated rate of blood flow, permitting more oxygen to be transported. In addition, a larger fraction of the oxygen combined with hemoglobin may be extracted by the tissues with high oxygen requirements.

Elimination of carbon dioxide

Carbon dioxide, a waste product of cellular metabolism, is transported by the blood in the opposite direction. Occurring in relatively high concentration in the tissues, it diffuses into the blood and is carried to the lungs for elimination by way of the expired air. Carbon dioxide is much more soluble than oxygen and readily diffuses into red cells. It reacts with water to form carbonic acid, a weak acid that at the alkaline pH of the blood appears principally as bicarbonate. Carbon dioxide also reacts with hemoglobin to form a carbamate compound, a substance that is significant in carbon dioxide transport within red cells. The tension of carbon dioxide in the arterial blood is regulated with extraordinary precision through a sensing mechanism in the brain that controls the respiratory movements. Carbon dioxide is an acidic substance, and increase in its concentration tends to lower the pH. This tendency is averted by the stimulus that causes increased depth and rate of breathing, a response that accelerates the loss of carbon dioxide. It is the tension of carbon dioxide and not of oxygen in the arterial blood that normally controls breathing. Inability to hold one's breath for more than a minute or so is the result of the rising tension of carbon dioxide, which produces the irresistible stimulus to breathe. Respiratory movements that ventilate the lungs sufficiently to maintain a normal tension of carbon dioxide are adequate under normal conditions to keep the blood fully oxygenated. Control of respiration is effective in regulating the uptake of oxygen and disposal of carbon dioxide, and in maintaining the constancy of blood pH.

Nutrition. Each substance required for the nutrition of every cell in the body is transported by the blood: the precursors of carbohydrates, proteins, and fats; minerals and salts; vitamins and other accessory food factors. These substances must all pass through the plasma on the way to the tissues in which they are utilized. The materials may enter the bloodstream from the gastrointestinal tract, or they may be released from stores within the body or become available from the breakdown of tissue. Turnover of substances carried by the blood may be rapid in relation to their concentration in plasma. Hundreds of grams of sugar (glucose) may be absorbed into the blood each day; yet the plasma concentration of glucose does not greatly exceed 100 milligrams (mg) per hundred millilitres (ml). All of the calcium in the skeleton and teeth has had to traverse the plasma, but the concentration of calcium in plasma is closely guarded at about 10 mg per 100 ml. The concentrations of many plasma constituents, including glucose and calcium, are carefully regulated, and deviations from the normal may have catastrophic effects. One of the regulators of the blood sugar is insulin, a hormone released into the blood from glandular cells in the pancreas. Ingestion of a carbohydrate meal is followed by increased production of insulin, which tends to keep the blood sugar from rising excessively. But an excess of insulin may severely reduce the blood sugar, causing a reaction that, if sufficiently severe, may include coma and even death. Glucose is transported in simple solution, but some substances require specific binding proteins (proteins with which the substances form temporary unions) to convey them through the plasma. Iron and copper, essential minerals, have special and necessary transport proteins. Nutrient substances may be taken up selectively by the tissues that require them. Growing bones utilize large amounts of calcium, and bone marrow removes iron from plasma for hemoglobin synthesis.

Excretion. The blood carries the waste products of cellular metabolism from their sites of release to the excretory organs. The removal of carbon dioxide in the lungs has been described. Water produced by the oxidation of foods or available from other sources in excess of needs is excreted by the kidneys as the solvent of the urine. Water derived from the blood also is lost from the body by evaporation from the skin and lungs and in small amounts from the gastrointestinal tract. The water content of the blood and of the body as a whole remains within a narrow range because of effective hormonal and other regulatory mechanisms that determine the urinary volume. The concentrations of physiologically important ions of the plasma, notably sodium, potassium, and chloride, are precisely controlled by their retention or selective removal as blood flows through the kidneys. Of special significance is the renal (kidney) control of acidity of the urine, a major factor in the maintenance of the normal pH of the blood. Urea, creatinine, and uric acid are nitrogen-containing products of metabolism that are transported by the blood and rapidly eliminated by the kidneys. The kidneys clear the blood of many other substances, including numerous drugs and chemicals that are taken into the body. In performing their excretory function, the kidneys have a major responsibility for maintaining the constancy of the composition of the blood. The liver is in part an excretory organ. Bilirubin (bile pigment) produced by the destruction of hemoglobin is conveyed by the plasma to the liver and is excreted through the biliary ducts into the intestinal tract. Other substances, including certain drugs, also are removed from the plasma by the liver.

Work of white cells, blood enzymes, and antibodies

Defense mechanisms. Cells of the blood and constituents of the plasma interact in complex ways to confer immunity to infectious agents, to resist or destroy invading organisms, to produce the inflammatory response, and to destroy and remove foreign materials and dead cells. The leukocytes (white blood cells, discussed later in this article) have a primary role in these reactions: granulocytes and monocytes phagocytize (ingest) bacteria and other organisms; they migrate to sites of infection or inflammation and to areas containing dead tissue; they participate in the enzymatic breakdown and removal of cellular debris; lymphocytes are concerned with the development of immunity and with acquired resistance to infections. Blood contains a number of substances that nonspecifically inhibit the growth of micro-organisms. One of these is lysozyme (muramidase), a bacteriocidal enzyme that occurs also in other body fluids, including the tears. Acquired resistance to specific micro-organisms is in part attributable to antibodies, plasma globulins that

are formed in response to the entry into the body of a foreign substance (antigen). Antibodies that have been induced by micro-organisms tend to prevent reinfection by the same organism and often confer a high degree of resistance. Cells and antibodies may cooperate in the destruction of invading bacteria; the antibody may attach to the organism and sensitize it so that it is more readily phagocytized. Involved in some of these reactions is complement, a group of protein components of plasma that participate in certain immunologic reactions. Micro-organisms and other cells sensitized by a specific antibody may be killed by complement.

Mechanisms to prevent blood loss

Hemostasis. The blood is contained under pressure in a vascular system that includes vast areas of thin and delicate capillary membranes. Even the bumps and knocks of everyday life are sufficient to disrupt some of these fragile vessels, and serious injury can be much more damaging. Loss of blood would be a constant threat to survival if it were not for protective mechanisms to prevent and to control bleeding. The platelets contribute to the resistance of capillaries, possibly because they actually fill chinks in vessel walls. In the absence of platelets capillaries become more fragile, permitting spontaneous loss of blood and increasing the tendency to form bruises after minor injury. Platelets immediately aggregate at the site of injury of a blood vessel, tending to seal the aperture. A blood clot, forming in the vessel around the clump of adherent platelets, further occludes the bleeding point. The coagulation mechanism involves a series of chemical reactions in which specific proteins and other constituents of the blood, including the platelets, play a part (see BLEEDING AND BLOOD CLOTTING). Plasma also is provided with a mechanism for dissolving clots after they have been formed. Plasmin is a proteolytic enzyme—a substance that causes breakdown of proteins—derived from an inert plasma precursor known as plasminogen. When clots are formed within blood vessels, activation of plasminogen to plasmin may lead to their removal.

Temperature regulation. Heat is produced in large amounts by physiological oxidative reactions, and the blood is essential both for its distribution and disposal. The circulation assures relative uniformity of temperature throughout the body and also carries the warm blood to the surface, where heat is lost to the external environment. A heat-regulating centre in the hypothalamus of the brain functions much like a thermostat. It is sensitive to changes in temperature of the blood flowing through it and, in response to the changes, gives off nerve impulses that control the calibre of the blood vessels in the skin and thus determine blood flow and skin temperature. A rise in skin temperature increases heat loss from the body surface. Heat is continuously lost by evaporation of water from the lungs and skin, but this loss can be greatly increased when more water is made available from the sweat glands. The activity of the sweat glands is controlled by the nervous system under direction of the temperature-regulating centre. Constancy of body temperature is achieved by control of the rate of heat loss by these mechanisms.

COMPONENTS

Plasma. The liquid portion of the blood, the plasma, is a solution of enormous complexity containing more than 90 percent water. Its major solute is a heterogeneous group of proteins constituting about 7 percent of the plasma by weight. Fatty substances (lipids) are present in several forms in suspension and in solution. Other constituents include salts, glucose, amino acids, vitamins, hormones, and waste products of metabolism.

Functions and distribution of water

Water. The water of the plasma is freely exchangeable with that of the body cells and extracellular fluids and is available to maintain the normal state of hydration of all tissues. Water is the largest single constituent of the body and is essential for the existence of every living cell. It forms both the solvent in which intracellular chemical reactions occur and the ambient fluid that provides the proper environment for the cells. More than half of the water is contained within the cells; the remainder makes up the extracellular fluid, of which most is outside of the blood vessels. The portion of the water that is contained in the plasma is relatively small but is the most dynamic in terms of flow and of turnover rates of its many solutes. The semipermeable membranes enclosing all tissue cells and forming the walls of the capillaries give free access to water, but limit the passage of many solutes. Water moves across these membranes to achieve osmotic equilibrium. The principal difference between plasma and the extracellular fluid of the tissues is the high protein content of plasma. Plasma protein exerts an osmotic effect by which water tends to move from the extravascular fluid into the blood. Movement of water across cell membranes is determined by the concentration of the electrolytes, especially sodium and potassium, within and outside of the cell. Since cell membranes are relatively impermeable to these ions, their osmotic effects regulate the distribution of water. Water is the source of sweat, tears, secretions of the respiratory and alimentary tracts, and the urine. In addition to its primary property as a solvent and suspending medium, water is of vital physiological significance in transfer of heat. Water has a high specific heat—*i.e.*, it can take up much heat without appreciable rise in temperature—and is ideally suited for its role in regulation of body temperature. The remarkable precision with which the water content of man is regulated is demonstrated by the day to day constancy of the weight of the body, of which about two-thirds is water. Compensatory mechanisms controlling water loss assure relative constancy of the body water, despite large variations in fluid intake. Deficiency of water gives rise to the sensation of thirst; satisfaction of this demand provides an adequate water intake. The kidneys regulate the amount of water lost from the body by varying the volume of the urine; important in this control mechanism is a hormone (antidiuretic hormone) secreted by the pituitary gland. The sensing mechanism seems to detect changes in osmotic pressure of the extracellular fluid. If the osmotic pressure rises, antidiuretic hormone is secreted in larger amount and acts on the kidney to restrict water loss in the urine.

Proteins. Proteins are complex chemical compounds occurring in all living organisms and forming essential constituents of every living cell. Chemically, they are large molecules formed of chains of amino acids, organic acids that contain both an acidic and a nitrogenous basic (amino) group. Chains (polypeptides) are formed by linkage of the acid group of one amino acid to the amino group of the next (peptide bond). The characteristics of a protein are determined by the number and types of amino acids and the sequence in which they are arranged.

When dietary protein is digested in the gastrointestinal tract, individual amino acids are released from the chains and are absorbed. The amino acids are transported through the plasma to all parts of the body, where they are taken up by cells and are assembled in specific ways to form proteins of many types. The plasma proteins are released into the blood from the cells in which they were synthesized. Much of the protein of plasma is produced in the liver. The major plasma protein is albumin, a relatively small molecule, the principal function of which is to retain water in the bloodstream by its osmotic effect. The amount of albumin in the blood is a determinant of the total volume of plasma. Depletion of albumin permits fluid to leave the circulation and to accumulate and cause swelling of soft tissues (edema). Albumin binds certain other substances that are transported in plasma and thus serves as a nonspecific carrier protein. Bilirubin, for example, is bound to albumin during its passage through the blood.

Globulins, fibrinogen, and fibrin

Albumin has physical properties that permit its separation from other plasma proteins, which as a group are called globulins. In fact, the globulins are a heterogeneous array of proteins of widely varying structure and function, only a few of which can be mentioned here. The immunoglobulins comprise the antibodies, each of which was produced in response to a specific antigen.

For example, administration of poliomyelitis vaccine is followed by the appearance in the plasma of antibodies that react with polio virus and effectively prevent that infection. Antibodies may be induced by many foreign substances in addition to micro-organisms; immunoglobulins are involved in some hypersensitive and allergic reactions.

Other plasma proteins are concerned with the coagulation of the blood. The framework of the clot is composed of fibrin, which is derived from a soluble plasma component known as fibrinogen. When a blood vessel is injured or blood is shed, fibrinogen undergoes a chemical change that causes its molecules to aggregate in a special way, forming strands and filaments. These join to produce a meshwork of fibrin in which the blood cells are trapped. The formation of the fibrin net is preceded by a complex series of poorly understood chemical reactions involving many other plasma proteins specifically concerned with the coagulation process. An enzyme, thrombin, is responsible for the conversion of fibrinogen to fibrin. Thrombin does not exist in circulating plasma but is derived from an inert plasma precursor, prothrombin, during the coagulation process. Serum, the fluid expressed from the blood clot, differs from plasma principally in that fibrinogen, prothrombin, and certain other procoagulant proteins have been altered or removed.

Thrombin, differentiation of plasma and serum

Transport of fats

Many proteins are involved in a highly specific way with the transport function of the blood. Blood lipids are incorporated into protein molecules as lipoproteins, substances of great importance in lipid transport. Iron and copper are transported in plasma by unique metal-binding proteins (transferrin and ceruloplasmin, respectively). Vitamin B_{12}, an essential nutrient substance, is bound to a specific carrier protein. Although hemoglobin is not normally released into the plasma, a hemoglobin binding protein (haptoglobin) is available to transport hemoglobin should hemolysis (breakdown) of red cells occur.

Lipids. The concentration of lipids in plasma varies, particularly in relation to meals, but ordinarily does not exceed one gram per hundred millilitres. The largest fraction consists of phospholipids, complex molecules containing phosphoric acid and a nitrogen base in addition to fatty acids and glycerol. Triglycerides, or simple fats, are molecules composed only of fatty acids and glycerol. Free fatty acids, lower in concentration than triglycerides, are responsible for a much larger transport of fat. Other lipids include cholesterol, a major fraction of the total plasma lipids. These substances exist in plasma combined with proteins of several types as lipoproteins. The largest lipid particles in the blood are known as chylomicrons and consist largely of triglycerides; after absorption from the intestine they pass through lymphatic channels and enter the bloodstream through the thoracic lymph duct. The other plasma lipids are derived from food or enter the plasma from tissue sites.

Other plasma components. Some plasma constituents occur in plasma in low concentration but have a high turnover rate and great physiological importance. Among these is glucose, the blood sugar. Glucose is absorbed from the gastrointestinal tract or may be released into the circulation from the liver. It provides a source of energy for tissue cells and is the only source for some, including the red blood cells. Glucose is conserved and utilized and not excreted. Amino acids also are so rapidly transported that the plasma level remains low, although they are required for all protein synthesis throughout the body. In contrast, urea, an end product of protein metabolism, is rapidly excreted by the kidneys. Other nitrogenous waste products—uric acid and creatinine—are similarly removed.

Glucose

Several inorganic materials are essential constituents of plasma, and each has special functional attributes. The predominant cation (positively charged ion) of the plasma is sodium, an ion that occurs within cells at a much lower concentration. Because of the effect of sodium on osmotic pressure and fluid movements, the amount of sodium in the body is an influential determinant of the total volume of extracellular fluid. Dietary sodium usually exceeds requirements, and the excess is excreted by the kidneys. The amount of sodium in plasma is controlled by the kidneys under the influence of a hormone (aldosterone) of the adrenal gland. Potassium, the principal intracellular cation, occurs in plasma at a much lower concentration than sodium. The renal excretion of potassium is influenced by aldosterone, which causes retention of sodium and loss of potassium. Calcium in plasma is in part bound to protein and in part ionized. Its concentration is under the control of two hormones: parathyroid hormone, which causes the level to rise, and calcitonin, which causes it to fall. Magnesium, like potassium, is a predominantly intracellular cation and occurs in plasma in low concentration. Variations in the concentrations of these cations may have profound effects on the nervous system, the muscles, and the heart, effects normally prevented by precise regulatory mechanisms. Iron, copper, and zinc are required in trace amounts for synthesis of essential enzymes; much more iron is needed in addition for production of hemoglobin and myoglobin, the oxygen-binding pigment of muscles. These metals occur in plasma in low concentrations. The principal anion (negatively charged ion) of plasma is chloride; sodium chloride is its major salt. Bicarbonate participates in the transport of carbon dioxide and in the regulation of pH. Phosphate also has a buffering effect on the pH of the blood and is vital for chemical reactions of cells and for the metabolism of calcium. Iodide is transported through plasma in trace amounts; it is avidly taken up by the thyroid gland, which incorporates it into thyroid hormone.

Sodium, potassium, other metals

Chloride and other anions, hormones

The hormones of all of the endocrine glands are secreted into the plasma and transported to their target organs, the organs in which they exert their effect. The plasma levels of these agents often reflect the functional activity of the glands that secrete them; in some instances measurements are possible though concentrations are extremely low. Among the many other constituents of plasma are numerous enzymes. Some of these appear simply to have escaped from tissue cells and have no functional significance in the blood. Others, like plasmin and thrombin, are derived from plasma precursors and have specific functions.

Red blood cells (erythrocytes). The red blood cells are highly specialized, well adapted for their primary function of transporting oxygen from the lungs to all of the body tissues. Rather uniform in size (about eight microns in diameter), red cells have the form of biconcave disks, a shape that provides a large surface to volume ratio. When blood is centrifuged to cause the cells to settle, the volume of packed red cells (hematocrit value) ranges between 42 and 54 percent of total volume in men and between 37 and 47 percent in women; values are somewhat lower in children. Normal red blood cells are fairly uniform in volume, so that the hematocrit value is determined largely by the number of red cells per unit of blood. The normal red cell count ranges between 4,000,000 and 6,000,000 per cubic millimetre. Hemoglobin constitutes about one-third of the weight of each red cell. The amount of hemoglobin in blood is related to the hematocrit value and to the red cell count, and in normal adults ranges between 14 and 18 grams per hundred millilitres. When fresh blood is examined with the microscope, red cells appear to be yellow-green disks with pale centres, containing no visible internal structures. They tend to adhere to each other like rolls of coins (rouleaux).

Size, shape, and volume of red cells

The red cell is enclosed in a thin membrane composed of chemically complex lipids, proteins, and carbohydrates in a highly organized structure. Extraordinary distortion of the erythrocyte occurs in its passage through minute blood vessels, many of which have a diameter less than that of the red cell. When the deforming stress is removed, the cell springs back to its original shape. The erythrocyte readily tolerates bending and folding; but if appreciable stretching of the membrane occurs, the cell is damaged or destroyed. The membrane is freely permeable to water, oxygen, carbon dioxide, glucose, urea, and certain other substances but is impermeable to hemoglobin. Within the cell the major cation is potassium; in con-

trast, in plasma and extracellular fluids the major cation is predominantly sodium. Relative impermeability of the membrane to sodium and potassium permits differences in concentration to be maintained. Because of the differential permeability of the membrane, red cells are subject to osmotic effects. When they are suspended in very dilute (hypotonic) solutions of sodium chloride, red cells take in water, which causes them to increase in volume and to become more spheroid; in concentrated salt solutions they lose water and shrink. In distilled water red cells continue to swell until they become spherical, whereupon they disrupt, releasing the dissolved hemoglobin into the surrounding fluid (hemolysis).

Breakdown of red cells

Hemolysis. When red cell membranes are damaged, hemoglobin and other dissolved contents may escape from the cells, leaving the membranous structures as "ghosts." This process, called hemolysis, is produced not only by the osmotic effects of water but by numerous other mechanisms. These include physical damage to red cells, as when blood is heated, is forced under great pressure through a small needle, or is subjected to freezing and thawing; chemical damage to red cells by agents such as bile salts, detergents, and certain snake venoms; and damage caused by immunologic reactions that may occur when antibodies attach to red cells in the presence of complement.

Blood groups. The membrane of the red cell has on its surface a group of agents that confer blood group specificity (*i.e.*, that differentiate blood cells into groups). Most blood group substances are composed of carbohydrate linked to protein, and it is the chemical structure of the carbohydrate portion that determines the specific blood type. Blood group substances are antigens capable of inducing the production of antibodies when injected into persons or animals lacking the antigen. Detection and recognition of the blood group antigens are accomplished by the use of serum containing these antibodies. Persons of blood group A do not have in their serum an antibody that reacts with the group A antigen, but they do have an antibody reacting with the red cells of persons of blood group B. By the use of serum samples, each containing one or more antibodies, a large number of red cell antigens can be identified. They include those of the important ABO and Rh systems (see BLOOD GROUPS). Normal persons have in their serum antibodies that react with the ABO antigens that they themselves lack. Antibodies reacting with other red cell antigens are not normally present in serum and appear only after the person has been sensitized, as by the administration of a transfusion of blood containing the antigen. Pregnant women may be sensitized to red cell antigens of the fetus, a phenomenon occurring most often with certain antigens of the Rh system. The blood groups are inherited characteristics and are useful in genetic studies and in the determination of paternity. The large number of different red cell antigens makes it extremely unlikely that persons other than identical twins will have the same array of blood group substances.

Hemoglobin. About 95 percent of the dry weight of the red cell consists of hemoglobin, the substance necessary for oxygen transport. Hemoglobin is a protein; a molecule contains four polypeptide chains, each chain consisting of more than 140 amino acids. To each chain is attached a chemical structure known as a heme group. Heme is composed of a ringlike organic compound known as a porphyrin to which an iron atom is attached. It is the iron atom that reversibly binds oxygen as the blood travels between the lungs and the tissues. There are four iron atoms in each molecule of hemoglobin, which, accordingly, can bind four atoms of oxygen. The complex porphyrin and protein structure may be considered to provide just the proper environment for the iron atom so that it binds and releases oxygen appropriately under physiological conditions.

Interaction of hemoglobin and oxygen

The affinity of hemoglobin for oxygen is so great that at the oxygen pressure in the lungs about 95 percent of the hemoglobin is saturated with oxygen. As the oxygen tension falls, as it does in the tissues, oxygen dissociates from hemoglobin and is available to move by diffusion through the red cell membrane and the plasma to sites where it is utilized. The proportion of hemoglobin saturated with oxygen is not directly proportional to the oxygen pressure. As the oxygen pressure declines, hemoglobin gives up its oxygen with disproportionate rapidity, so that the major fraction of the oxygen can be released with a relatively small drop in oxygen tension. The affinity of hemoglobin for oxygen is primarily determined by the structure of hemoglobin, but it is also influenced by other conditions within the red cell, in particular the pH. When the interior of the red cell becomes more acid, as it does when carbon dioxide diffuses into the cell, hemoglobin binds less oxygen. This phenomenon, known as the Bohr effect, is of physiologic importance. In the tissues, where carbon dioxide is taken up by red cells, affinity of the hemoglobin for oxygen is lessened and more is made available for the tissue cells. Other components of the red cell similarly affect oxygen binding by hemoglobin. Among these are certain organic phosphate compounds produced during the chemical breakdown of glucose, in particular 2,3-diphosphoglycerate. An increase in the concentration of this compound lessens the affinity of hemoglobin for oxygen. At high altitudes, where the atmospheric oxygen pressure is low, the concentration of 2,3-diphosphoglycerate in the red cell increases, lessening the affinity of the hemoglobin for oxygen and driving more oxygen into the tissues, where it is much needed. Acclimatization to high altitude is achieved in part by this mechanism.

Hemoglobin and carbon monoxide

Hemoglobin has a much higher affinity for carbon monoxide than for oxygen. Carbon monoxide produces its lethal effects by binding to hemoglobin and preventing oxygen transport. The oxygen-carrying function of hemoglobin can be disturbed in other ways. The iron of hemoglobin is normally in the reduced or ferrous state, both in oxyhemoglobin and deoxyhemoglobin. If the iron itself becomes oxidized to the ferric state, hemoglobin is changed to methemoglobin, a brown pigment incapable of transporting oxygen. The red cells contain enzymes capable of maintaining the iron in its normal state, but under abnormal conditions large amounts of methemoglobin may appear in the blood.

Sickle-cell anemia

Information of widely ranging importance has been derived from the study of the hemoglobin molecule. Discovery of the cause of sickle-cell anemia has led to major advances in understanding of genetics, molecular biology, and the mechanisms of disease. Persons who have sickle-cell anemia are Negroes with an inherited abnormality of the hemoglobin causing a serious and often fatal disease. Cause of the disease is the mutation of a single gene that determines the structure of one pair of the polypeptide chains of the hemoglobin molecule. Sickle hemoglobin differs from normal hemoglobin in that a single amino acid in one pair of the polypeptide chains has been replaced (glutamic acid by valine). This single intramolecular change so alters the properties of the hemoglobin molecule that anemia and other effects are produced. The entire structure of the hemoglobin molecule is known, and many other genetically determined abnormalities have been identified. Some of these also produce diseases of several types. Study of the effects of altered structure of hemoglobin on its properties has greatly broadened knowledge of the structure-function relationships of the hemoglobin molecule.

Red cell metabolism. Survival of the red cell in the circulation depends upon the continuous utilization of glucose for the production of energy. Two chemical pathways are employed, and both are essential for the normal life of the red cell. An extraordinary number of enzyme systems participate in these reactions and direct the energy evolved into appropriate uses. Red cells contain neither a nucleus nor RNA (ribonucleic acid, necessary for protein synthesis), so that cell division and production of new protein are impossible. Energy is not necessary for oxygen and carbon dioxide transport, which depends principally on the properties of hemoglobin. Energy is needed for another operation: There is a tendency for the extracellular cation, sodium, to leak into the red cell and

for potassium to leak out; energy is required to operate a pumping mechanism in the red cell membrane necessary to maintain the normal gradients (differences in concentrations) of these ions. Energy is also required to convert methemoglobin to oxyhemoglobin and to prevent the oxidation of other constituents of the red cell.

Erythropoiesis (production of red cells). Red cells are produced continuously in the marrow of certain bones. The principal sites are the marrow spaces of the skull, vertebrae, ribs, breastbone, and pelvis, and of the upper ends of the femurs and humeri. Within the bone marrow the red cell is derived from a primitive precursor or erythroblast, a nucleated cell in which there is no hemoglobin. Proliferation occurs as a result of several successive cell divisions. During maturation hemoglobin appears in the cell and the nucleus becomes progressively smaller. In a few days, the cell loses its nucleus and is then introduced into the bloodstream in the vascular channels of the marrow. Almost 1 percent of the red cells are generated each day, and the balance between red cell production and the removal of aging red cells from the circulation is precisely maintained. If blood is lost from the circulation, the erythropoietic activity of marrow increases until the normal number of circulating cells has been restored.

Nutrients needed for red cell production

In a normal adult the red cells of almost one pint of blood are produced by the bone marrow every week. A number of nutrient substances are required for this process. Some nutrients are the building blocks of which the red cells are composed. For example, amino acids are needed in abundance for the construction of the proteins of the red cell, in particular of hemoglobin. Iron is a component of hemoglobin, and without it hemoglobin cannot be synthesized. Approximately one-quarter of a gram of iron is needed for the production of a pint of blood. Other substances, required in trace amounts, are needed to catalyze the chemical reactions by which red cells are produced. Important among these are several vitamins, riboflavin, vitamin B_{12}, and folic acid, necessary for the maturation of the developing red cell; and pyridoxine (vitamin B_6), required for the synthesis of hemoglobin. The secretions of several endocrine glands influence red cell production. If there is an inadequate supply of thyroid hormone, erythropoiesis is retarded and anemia appears. The male sex hormone, testosterone, stimulates red cell production; for this reason, red cell counts of men are higher than those of women.

The capacity of the bone marrow to produce red cells is enormous. When stimulated to peak activity and when provided adequately with nutrient substances, the marrow can compensate for the loss of several pints of blood per week. Hemorrhage or accelerated destruction of red cells leads to enhanced marrow activity. The rate of erythropoiesis is sensitive to the oxygen tension of the arterial blood. When oxygen tension falls, more red cells are produced and the red cell count rises. For this reason, persons who live at high altitude have higher red cell counts than those who live at sea level. There is a small but significant difference between average red cell counts of persons living in New York City, at sea level pressure, and persons living in Denver, one mile above sea level, where the atmospheric pressure is lower. Natives of the Andes, living nearly three miles above sea level, have extremely high red cell counts.

The rate of production of erythrocytes is controlled by a hormone-like substance (erythropoietin) that is produced largely in the kidneys. When the circulating red cells decrease or when the oxygen transported by the blood diminishes, an unidentified sensor detects the change and the production of erythropoietin is increased. This substance is then transported through the plasma to the bone marrow, where it accelerates the production of red cells. The erythropoietin mechanism operates like a thermostat, increasing or decreasing the rate of red cell production in accordance with need. When a man who has lived at high altitude moves to a sea-level environment, production of erythropoietin is suppressed, the rate of red cell production declines, and the red cell count falls until the normal sea level value is achieved. When a normal person donates a pint of blood to the blood bank, the erythropoietin mechanism is activated, red cell production is enhanced, and within a few weeks the number of circulating red cells has been restored to the previous value. The precision of control is extraordinary, so that the number of new red cells produced quite accurately compensates for the number of cells lost or destroyed.

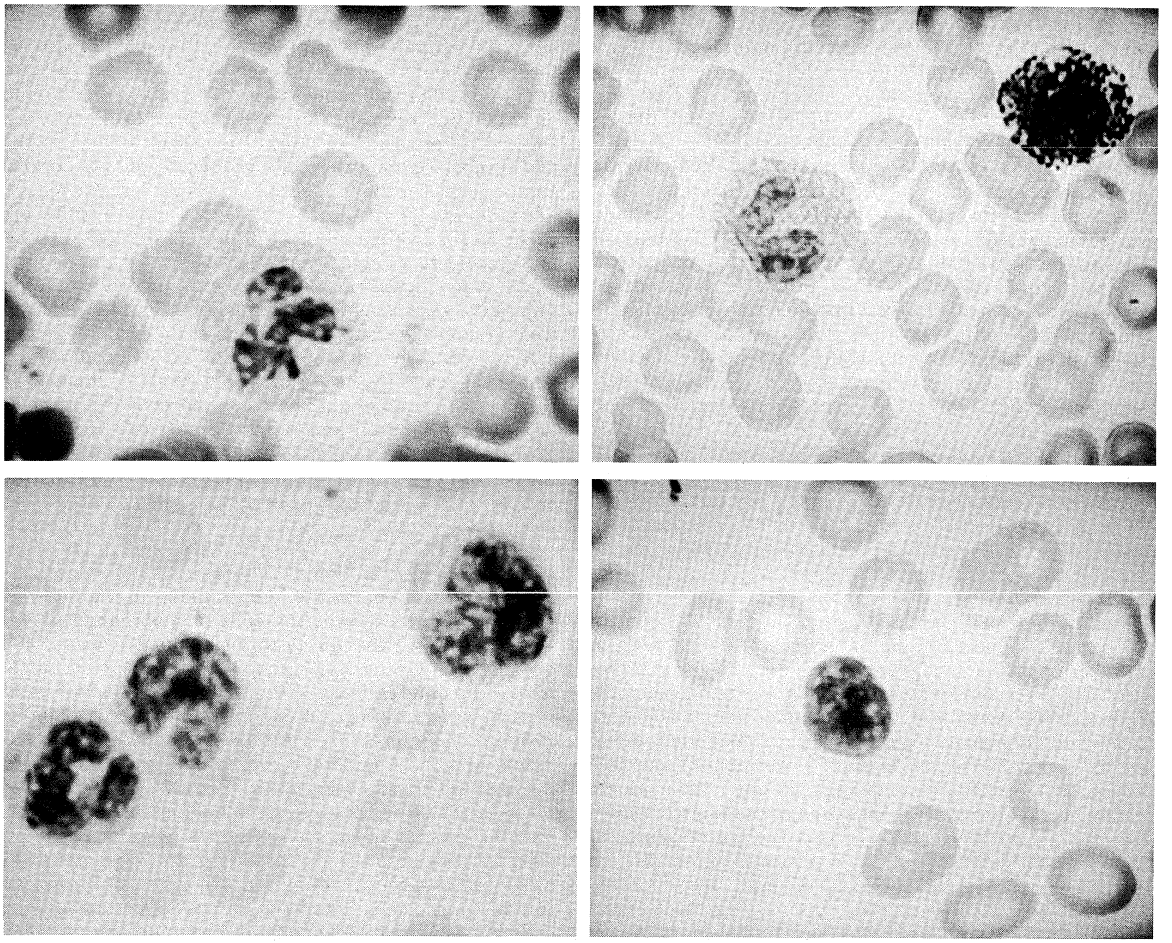
Removal of old and damaged red cells from circulation

Destruction of red cells. Red cells have a finite lifespan, on the average about 120 days. Although they utilize glucose to produce energy necessary for their survival, they are incapable of synthesizing protein; therefore reparative processes are not possible. As red cells age, wear and tear leads to loss of some of the protein, and the activity of some of the essential enzymes decreases. Chemical reactions necessary for survival of the cell are consequently impaired. The worn-out cells are removed from the circulation and rapidly destroyed. The normal process of destruction does not involve hemolysis with release of hemoglobin into the plasma. Instead, the entire red cell is engulfed by certain phagocytic cells that form a part of the lining of blood vessels, particularly in the spleen, liver, and bone marrow. These cells form a part of the reticuloendothelial system. Reticuloendothelial cells are also found in the lymph nodes, in the intestinal tract, and as free-wandering cells or macrophages. As a group they have the ability to ingest not only other cells but also many other microscopic particles including certain dyes and colloids. Within the reticuloendothelial cells erythrocytes are destroyed with remarkable rapidity. Protein, including that of the hemoglobin, is broken down, and the component amino acids are transported through the plasma to be utilized in the synthesis of new proteins. The iron removed from hemoglobin passes back into the plasma and is transported to the bone marrow, where it may be utilized promptly in the synthesis of hemoglobin in newly forming red cells. Iron not necessary for this purpose is stored within the reticuloendothelial cells but is available for release and reutilization whenever it is required. In the breakdown of red cells there is no loss to the body of either protein or iron, virtually all of which is conserved and reutilized. In contrast, the porphyrin ring structure of hemoglobin, to which iron was attached, undergoes a chemical change by which it is converted to bilirubin, a yellow pigment. Bilirubin released from reticuloendothelial cells after the destruction of erythrocytes is conveyed through the plasma to the liver, where it is extracted and secreted as bile pigment. The amount of bilirubin produced and secreted into the bile is determined by the amount of hemoglobin destroyed.

Disposal of products of red cell destruction

White blood cells (leukocytes). White cells, unlike erythrocytes, are nucleated and independently motile. Highly differentiated for their specialized functions, they do not undergo mitosis (ordinary cell division) in the bloodstream, but some retain the capability of cell division. As a group they are concerned with defense mechanisms and reparative activity. The number of leukocytes in normal blood ranges between 4,500 and 11,000 per cubic millimetre. Fluctuations occur during the day; lower values are obtained during rest and higher values during exercise. Violent physical exertion may cause the count to exceed 20,000 per cubic millimetre. Most of the leukocytes are outside of the circulation, and the few in the bloodstream are in transit from one site to another. As living cells, their survival depends on their continuous production of energy. The chemical pathways utilized are more complex than those of the red cells and are similar to those of other tissue cells. Leukocytes, containing a nucleus and able to produce RNA, can synthesize protein. They comprise three classes of cells, each unique as to structure and function, and designated granulocytes, monocytes, and lymphocytes.

Granulocytes. Most numerous of the white cells are the granulocytes, cells larger than red cells, having a multilobed nucleus and containing large numbers of cytoplasmic granules (*i.e.*, granules in the cell substance outside the nucleus). There are three types of granulocytes: neutrophils, eosinophils, and basophils. They are identi-



Types of white blood cells with red blood cells in background.
(Top left) Neutrophil; (top right) eosinophil (left) and basophil (right); (bottom left) monocytes; (bottom right) lymphocyte.
C. Lockard Conley

fied by the colour of the granules when the cells are stained with a compound dye. The granules of the neutrophil are pink, those of the eosinophil are red, and those of the basophil are blue black. About 60 percent of the white cells are neutrophils, while the eosinophils and basophils together comprise no more than 3 percent.

Neutrophils

The neutrophilic granulocytes are cells of fairly uniform size with a diameter between 12 and 15 microns. When examined in fresh blood that has not been allowed to dry, they display marked activity characterized by amoeboid motion: they extend long projections called pseudopods into which their granules flow and in this way rapidly advance along a surface. The nucleus consists of two to five lobes joined together by hairlike filaments. Within the body the neutrophils migrate to areas of infection or tissue injury. The attractive force that determines the direction of motion is known as chemotaxis and is attributed to substances liberated at sites of tissue damage. Neutrophils are actively phagocytic: they engulf bacteria and other fine particles and they may destroy living microorganisms. The granules of the neutrophil are microscopic packets of potent enzymes capable of digesting many types of cellular materials. When a bacterium is engulfed by a neutrophil, it is encased in a membranous sphere or vacuole. The granules empty their contents into the minute compartment containing the infectious organism. As this phenomenon occurs, the granules of the neutrophil are depleted (degranulation). Digestion of the invading organism is accomplished by the enzymes that surround it, and the organism is destroyed. Enzymes of the granulocyte are important not only in the elimination of micro-organisms but also in the removal of dead cells in areas in which tissue has been injured. Granulocytes, themselves, may be destroyed during the process.

Eosinophils and basophils, origins of granulocytes

Eosinophils, like neutrophils, are motile and actively phagocytic. They tend to be involved in tissue reactions in which there are antigen-antibody interactions. Less certain information is available about the specific function of the basophils.

The origin of all of the granulocytes is the bone marrow, and there is evidence that cells from which they arise, called stem cells, are the same line of cells as that which gives rise to the red cell precursors. The earliest cell recognizable as a precursor of the granulocytes is a myeloblast, extremely different in appearance from the mature cell. Granulocytic cells in various stages of development form the major component of the blood-forming cells within the bone marrow. The marrow contains a large reserve of mature granulocytes that can be released into the circulation when needed; it is thought to contain about a dozen times more mature granulocytes than are present in the blood. Within the circulation the neutrophils are not uniformly distributed in the blood but tend to concentrate along the membranes of the small blood vessels. Granulocytes needed to combat local infection may be derived from this "marginal pool" as well as from an increased output by the bone marrow. An equilibrium is normally maintained between the number of cells entering and leaving the blood. Granulocytes remain in the circulation for only a few hours, and after leaving it they do not return. A small number of these cells are found in the various secretions of the body, and many are removed by the reticuloendothelial system. Vast numbers may accumulate at sites of inflammation. From the time of their emergence from the bloodstream to the time of their loss or destruction the granulocytes are distributed through tissue spaces in numbers much larger than their number in the circulation.

Equilibrium between cells entering and leaving blood stream

The relative constancy of the granulocyte count in the blood implies that some control mechanism exists. Factors of hormonal nature, perhaps comparable to erythropoietin, probably influence the rate at which granulocytes are produced as well as the rate at which they are released from the bone marrow into the circulation. Definitive information about these control mechanisms is not yet available.

Characteristics and functions of monocytes

Monocytes. The largest cells of the blood are the monocytes, which make up on the average about 7 percent of the leukocytes. The nucleus is relatively big and tends to be indented or folded rather than multilobed. The cytoplasm contains large numbers of fine granules, which often appear to be more numerous near the cell membrane. Monocytes are actively motile and phagocytic. They are capable of ingesting infectious agents as well as red cells and other large particles, but they cannot replace the function of the neutrophils in the removal and destruction of bacteria. Closely related to the phagocytic cells of the reticuloendothelial system, they are considered to be scavengers and may enter areas of inflamed tissue later than the granulocytes. Often they are found at sites of chronic infections. Most of the available information about the origin and migrations of monocytes has been obtained from the study of animals. Few monocytes or cells identifiable as precursors of monocytes are seen in human bone marrow, but in the mouse there is evidence that monocytes have a marrow origin. The large phagocytic cells of the tissues, the macrophages, appear to have their origin from the blood monocytes.

Lymphocytes. About 30 percent of the white cells of the blood are lymphocytes. They have a single round nucleus and few if any granules in the cytoplasm. Most lymphocytes are small, only slightly larger then erythrocytes, with a nucleus that occupies most of the cell. Some are larger and have more abundant cytoplasm that stains blue. Lymphocytes are sluggishly motile, and their paths of migration outside of the bloodstream are very different from those of granulocytes and monocytes. Lymphocytes are found in large numbers in the lymph nodes, spleen, thymus, tonsils, and lymphoid tissue of the gastrointestinal tract. They enter the circulation through lymphatic channels that drain principally into the thoracic lymph duct, which has a connection with the venous system. Unlike other blood cells, some lymphocytes may leave and re-enter the circulation, surviving for about a year or more. The principal paths of recirculating lymphocytes are through the spleen or lymph nodes. Lymphocytes freely leave the blood to enter lymphoid tissue, passing barriers that prevent the passage of other blood cells. When stimulated by antigen and certain other agents, some lymphocytes "transform" to immature cells capable of cell division.

Functions and types of lymphocytes

The lymphocytes are concerned with vital defense mechanisms pertaining to acquired immunity to foreign cells and antigens. They are responsible for immunologic reactions to invading organisms and to foreign cells such as those of a transplanted organ or a cancer. In addition, they are involved in immunologic reactions to foreign proteins and other antigens not necessarily derived from living cells. Two classes of lymphocytes are recognized, indistinguishable by the usual microscopic examination. One class, represented by a small fraction of the blood lymphocytes, is capable of producing immunoglobulins, including various types of antibodies. These cells, when appropriately stimulated by a particular antigen, synthesize the corresponding antibody. The other class of lymphocytes, which includes most of those in the blood, is concerned with cell-mediated immunity. These are the lymphocytes that participate in the rejection of transplanted tissues and are involved in certain types of allergy.

In birds the differentiation of these two classes of lymphocytes is known to be dependent upon two organs. The cells capable of producing antibody have differentiated under the influence of lymphoid tissue in an outpouching of the gastrointestinal tract known as the bursa of Fabricius. Lymphocytes concerned with cell-mediated immune reactions have acquired their properties as the result of an influence of the thymus (see below). Lymphoid cells in various parts of the body have the properties of one or the other class. Whether a comparable relationship exists in man is uncertain, although there is abundant evidence of the role of the human thymus in immunologic reactions, and lymphoid tissue is plentiful in the gastrointestinal tract.

Thymus, rejection of foreign tissues

The thymus is a small organ lying behind the breastbone in the upper portion of the chest. It is relatively large at birth, begins to regress after puberty, and may be represented only by a fibrous cord in the elderly. During embryonic life, lymphocytes appear in the thymus before they are found elsewhere in the body. The effect of the thymus on the differentiation of lymphocytes begins before birth, and lymphocytes leaving the thymus populate lymphoid tissue elsewhere. Removal of the thymus from certain animals at the time of birth prevents the development of normal immunologic responses. The manner in which the thymus induces the differentiation of lymphocytes is unknown.

A primary function of the lymphocytes is to permit the body to distinguish its own tissues from material of foreign origin. If skin is removed from one part of the body and transplanted to another area of the same person, it will adhere and survive. On the other hand, if skin from one person is transplanted to another (other than an identical twin) the graft will be rejected. This is the result of an important protective mechanism that prevents the proliferation of foreign cells within the host, but it is also the mechanism that makes organ transplantation difficult. Rejection of a tissue graft is largely accomplished by cell-mediated immunity; lymphocytes migrate to the area of the graft and cause its destruction. Cell-mediated immunity is involved in other reactions including the tuberculin reaction: when persons who have had tuberculosis have an extract of the tubercle bacillus injected into their skin, a "positive" response occurs, an inflammatory reaction at the site of injection. The "positive" reaction is mediated by cells previously "sensitized" by the tuberculous infection; antibody is not involved. On the other hand, when a person has been infected or immunized with polio virus, antibodies are produced that are detectable in the serum. The ability to acquire immunity to foreign substances is called immunologic competence. Immunologic competence begins to develop during embryonic life, is incomplete at the time of birth, but is fully established soon after birth. If an antigen is introduced into the body before immunologic competence has been established, the individual will not subsequently develop immunologic reactions to it and is said to be tolerant to that antigen.

Tuberculin reaction, immunologic competence, immune tolerance

Study of immunologic competence and immune tolerance has been accelerated by interest in organ transplantation. Success of kidney transplants has been improved by increased knowledge about donor selection and techniques for suppressing the immune responses of the recipient. Comparable success has not yet been achieved with heart transplants.

Platelets (thrombocytes). The blood platelets are the smallest cells of the blood, averaging about two microns in diameter. Although much more numerous (150,000 to 300,000 per cubic millimetre) than the white cells, they occupy a much smaller fraction of the volume of the blood because of their relatively minute size. Like the red cells they lack a nucleus and are incapable of cell division, but they have a more complex metabolism and internal structure than have the red cells. When seen in fresh blood they appear spheroid, but they have a remarkable tendency to extrude hairlike filaments from their membranes. They adhere to each other, but not to red cells and white cells. In stained preparations of dried blood, platelets have a central granular area, the granulomere, and a peripheral, more clear area, the hyalomere. The granules contain substances important for the activity of platelets.

The function of the platelets is related to hemostasis, the prevention and control of bleeding. When the endo-

thelial surface (lining) of a blood vessel is injured, platelets in large numbers immediately attach to the injured surface and to each other, forming a tenaciously adherent mass of platelets. The mechanism of this reaction is not clearly understood, but platelets have an intrinsic "stickiness"; they tend to adhere to foreign surfaces and to collagen, an important component of normal connective tissue. The effect of the platelet response is to terminate bleeding and to form the site of the developing blood clot or "thrombus." If platelets are absent, this important defense reaction cannot occur, and protracted bleeding from small wounds (prolonged "bleeding time") results. The normal resistance of capillary membranes to leakage of red cells is dependent upon platelets. Severe deficiency of platelets reduces the resistance of the capillary walls, and abnormal bleeding from the capillaries occurs, either spontaneously or as the result of minor injury. Platelets also contribute substances essential for the normal coagulation of the blood, and they cause the shrinking, or "retraction," of a clot after it has been formed.

Effect of thrombin

Platelets are uniquely susceptible to the enzymatic action of thrombin, the substance responsible for converting fibrinogen to fibrin. Thrombin causes platelets to aggregate and to undergo irreversible structural and biochemical changes during which their survival as living cells is terminated. When a blood clot is formed by the action of thrombin, platelets display remarkable distortions of their membrane and extrude bulbous and thread-like pseudopods. The newly forming fibrin strands appear to become attached to the platelet filaments. The platelets subsequently draw together, causing the retraction of the clot. Platelets contain a contractile protein, thrombosthenin, which undoubtedly plays an important part in this phenomenon. On exposure to thrombin and certain other agents, the platelets release through their membranes some of their contents, including potassium and the substances that participate in the coagulation of the blood. Adenine diphosphate, one of the substances released, appears to be important in causing platelet aggregation. Platelets are capable of removing from the plasma and storing in high concentration certain chemical agents, including serotonin, a substance that causes constriction of small blood vessels. Serotonin is rapidly released when platelets aggregate and probably plays a part in control of bleeding by reducing the size of injured vessels.

Platelets are formed in the bone marrow by segmentation of the cytoplasm (the cell substance other than the nucleus) of giant cells known as megakaryocytes, the largest cells of the marrow: during differentiation in the marrow the megakaryocyte becomes large, and multiple nuclei appear; the abundant granular cytoplasm of the megakaryocyte divides into many little segments that break off and are released as platelets into the circulating blood. After about ten days in the circulation, platelets are removed and destroyed. There are no reserve stores of platelets except in the spleen, in which platelets occur in higher concentration than in the peripheral blood. Some platelets are consumed in exerting their hemostatic effects, and others, reaching the end of their life-span, are removed by reticuloendothelial cells. The rate of platelet production is controlled, but not so precisely as the control of red cell production. A hormone-like substance, thrombopoietin, as yet unidentified, is believed to be the chemical mediator that regulates the number of platelets in the blood by controlling the rate of platelet production.

EXAMINATION OF THE BLOOD IN THE LABORATORY

Physicians rely upon the clinical laboratory to obtain measurements of many constituents of the blood, information useful or necessary for the detection and recognition of disease. Some tests, such as one to detect anemia, are so regularly performed as a part of the health examination that they are considered routine. Hemoglobin is a highly coloured pigment that interferes with the passage of a beam of light. To measure hemoglobin concentration, blood is accurately diluted and the red cells broken down to yield a clear red solution. A photoelectric instrument is employed to measure the absorbance of transmitted light, from which hemoglobin concentration can be calculated. Hemoglobin is a relatively unstable pigment, and the best methods to measure it employ chemicals to form hemoglobin derivatives, in particular cyanmethemoglobin, with which more accurate measurements can be made. Changes in the hemoglobin concentration of the blood are not necessarily directly paralleled by changes in the red cell count and the hematocrit value, because the size and hemoglobin concentration of red cells may change in disease. Therefore, measurements of the red cell count and hematocrit value may provide additional useful information. Red cells can be counted visually with the use of a microscope. Highly diluted blood is introduced into a "counting chamber," a device in which a film of fluid of precisely controlled thickness is superimposed upon an accurately ruled grid, permitting the enumeration of cells in a predetermined volume of diluted blood. The method is laborious and imprecise and has been largely abandoned in favour of electronic particle counters that make possible extremely accurate red cell counts within a few seconds. The hematocrit value is obtained, after centrifuging blood in a tube of uniform diameter, by measuring the percentage of the blood column that is occupied by the packed red cells; an anticoagulant is required to prevent clotting during the test procedure. Variations in the leukocyte count beyond the normal range occur in many diseases. White cell counts can be performed with the counting chamber or with the electronic device; the blood is diluted and the red cells destroyed by hemolysis before the count is performed. Complex automated machines are available that simultaneously determine the hematocrit value, the hemoglobin concentration, and white cell and red cell counts. Only a tiny drop of blood is needed for the analyses, which are completed within a minute; results are printed out on a laboratory report that is sent to the doctor. Although expensive, the equipment increases the output of the laboratory and saves the technician valuable time.

Red cell count

Adequate examination of the blood cells requires that a thin film of blood be spread on a glass slide, stained with a special blood stain (Wright's stain), and examined under the microscope. Individual red cells, white cells, and platelets are examined, and the relative proportions of the several classes of white cells are tabulated. The information acquired may have important diagnostic implications. In iron deficiency anemia, for example, the red cells look paler than normal because they lack the normal amount of hemoglobin; in malaria the diagnosis is established by observing the malarial parasites within the red cells. In pneumonia and many infections the proportion of neutrophilic leukocytes is usually increased, while in others, such as whooping cough and measles, there is an increase in the proportion of lymphocytes.

Examination of individual cells

Chemical analyses are employed to measure many of the constituents of plasma. Often serum rather than plasma is used, since it can be obtained from clotted blood without the addition of an anticoagulant. Tests can be performed manually, with an individual procedure for each analysis. Quantitative determination of the amount of sugar in the blood is essential for the diagnosis of diabetes, a disease in which the blood sugar tends to be elevated. Nitrogenous waste products, in particular urea, tend to accumulate in patients with diseased kidneys that are unable to excrete these substances at a normal rate. An increase in the concentration of bilirubin in the serum often reflects a disorder of the liver and bile ducts, or an increased rate of destruction of hemoglobin. Measurements of these and many other serum constituents are so valuable in medical diagnosis that often multiple tests are performed. The autoanalyzer, a completely automated device, increases the number of procedures that can be performed in clinical laboratories. A dozen analyses may be made simultaneously by a single machine, employing a small amount of serum. The serum is automatically drawn from a test tube and is propelled through plastic tubing of small diameter. As the serum specimen

Chemical analyses

advances it is divided; appropriate reagents are added; chemical reactions occur with formation of a product that can be measured with a photoelectric instrument; and the result appears as a written tracing from which serum concentration of various substances can be read directly. The data acquired by the machine may be fed automatically into a computer and the numerical results printed on a form that is submitted to a physician. Many analyses are not performed routinely but are invaluable in special circumstances. In cases of suspected lead poisoning, for example, detection of an elevated level of lead in the blood may be diagnostic. Some analytical procedures have specific diagnostic usefulness. These include assays for certain hormones, including measurement of the thyroid hormone in the serum of patients suspected of having thyroid disease.

Immunologic reactions of blood

Other clinically important laboratory procedures are concerned with immunologic reactions of the blood. Careful determinations of the blood groups of the patient and of the blood donor, and cross matching of the cells of one with the serum of the other to ensure compatibility, are essential for the safe transfusion of blood. Determination of the Rh type of the pregnant woman is regularly performed and is necessary for the early detection of fetal-maternal incompatibility and for proper prevention or treatment of hemolytic disease of the newborn. The diagnosis of certain infectious diseases depends upon the demonstration of antibodies in the patient's serum. A serologic test is used routinely for the detection of syphilis.

Other procedures

Radioactive substances are used for special types of tests, including measurements of blood volume and estimation of the life-span of red cells. Many other kinds of blood examination yield useful results. Enzymes normally present in the muscle of the heart may be released into the blood when the heart is damaged by a coronary occlusion (obstruction of the coronary artery). Measurement of these enzymes in the serum is regularly performed to assist in diagnosis of this type of heart disease. Damage to the liver releases other enzymes, measurement of which aids in evaluation of the nature and severity of liver disease. Inherited abnormalities of proteins are increasingly recognized and identified by use of sophisticated methods. Accurate diagnosis of hemophilia and other bleeding disorders is made possible by investigations of the coagulation mechanism. Measurements of the concentration of certain vitamins in the blood provide the basis for diagnosis of some vitamin deficiencies. The number of potentially useful blood tests is so vast that they must be selected judiciously in the evaluation of the individual patient.

BIBLIOGRAPHY. Further information on this subject may be found in the following texts: I. DAVIDSOHN and J.B. HENRY (eds.), *Clinical Diagnosis by Laboratory Methods*, 14th ed. (1969); J.W. HARRIS, *The Red Cell*, 2nd ed. (1970); R.G. MACFARLANE and A.H.T. ROBB-SMITH (eds.), *Functions of the Blood* (1961); V.B. MOUNTCASTLE (ed.), *Medical Physiology*, 12th ed., 2 vol. (1968); A. WHITE, P. HANDLER, and E.L. SMITH, *Principles of Biochemistry*, 4th ed. (1968); W.J. WILLIAMS *et al.*, *Hematology* (1971); and M.M. WINTROBE, *Clinical Hematology*, 6th ed. (1967).

(C.L.C.)

Blood and Lymph

The circulatory system, which contains blood and lymph, is an important connecting pathway; it permits a continuous integration among various tissues and organs of animals and facilitates contact with the environment outside the body. Blood is a cell-containing fluid that transports oxygen, water, carbon dioxide, products of metabolism, and internal secretions (*e.g.*, hormones). Blood is a tissue that is constantly circulating throughout animals, a means by which the constancy of the internal environment is maintained, and it is also the route by which the defenses against injury and disease may be quickly mobilized. Lymph is a fluid that is derived from the tissues of animal bodies and conveyed to the blood stream by lymphatic vessels.

This article compares blood and lymph in a number of living systems. For more specific information about blood in man, see BLOOD, HUMAN.

GENERAL FEATURES

Definitions. Blood in mammals may be defined as the red fluid that is pumped by the heart into arteries and returns to the heart in veins, following a complex but completely closed circular path. The red colour of mammalian blood is caused by hemoglobin, an oxygen-carrying protein in red cells (erythrocytes). Mature red blood cells of mammals, which contain no nuclear component, are disk shaped and float in a nearly colourless liquid called plasma. Plasma also contains white cells (leukocytes) and tiny bodies called platelets, which seal wounds. The grainy cytoplasmic component of one type of white cell, called granulocytes, constantly changes shape, engulfing and destroying unwanted particles. A nongranular type of white cell (lymphocytes) originates in mammalian lymphoid tissue (*e.g.*, spleen, lymph nodes) and has important immunological functions. Lymph, a rather colourless, sometimes milky liquid, drains from the tissues into lymph vessels, which converge into a thoracic duct, which empties into the blood. The above definitions of blood and lymph also apply (with modifications) to vertebrates other than mammals; *i.e.*, the red cells of other vertebrates contain nuclei but are red, except in the icefish, which has colourless blood. White cells vary in shape not only among vertebrates but even among mammals. The platelets of mammals resemble tiny disks of cytoplasm, but those of other vertebrates have nuclei.

Invertebrate animals have a great variety of liquids, cells, and modes of circulation. In insects, a heart pumps fluid, called hemolymph, through an open system to the tissues. In some worm species the fluid contains pigmented cells, but the fluid, called coelomatous fluid, does not circulate; in others the fluid circulates, but the cells may contain no pigment.

Perhaps blood should be defined by its functions, rather than by its appearance—when passing through skin, gills, or lungs, blood picks up oxygen from outside the body and loses carbon dioxide that was picked up in the metabolizing tissues; when passing through tissues, blood does the reverse and also picks up metabolized food, which maintains the cells of the body, and passes poisonous molecules and excess water to the kidneys for excretion.

Origin. How did the need for blood and lymph arise? Life probably originated in the oceans from molecules that accidentally developed an ability to reproduce, to form cells, and to evolve by reproducing many of the accidental changes (mutations) that occurred in their hereditary material. Most of the early cells must have developed certain deficiencies, and many survived only if they also developed the ability to depend on other cells, either as food sources or for transporting food, oxygen, and waste products.

Eventually, aggregations of specialized cells became organized into bodies whose inner cells lost contact with the outer world and with each other. An electronic system (the nervous system) and liquid systems of blood and lymph developed to provide communication among specialized body regions. In addition, as organs became complicated, both structurally and functionally, they developed a mechanism (hormones) for transmitting, receiving, and integrating information. A hormone is formed in one organ and transported in the blood to another site, where its major effect is exerted. An increase in blood sugar, for example, causes an increase in synthesis (in the pancreas) of the hormone insulin, which causes a decrease in blood sugar; an increase in the activity of the pituitary gland, in the brain, causes an increase in the activity of the adrenal glands, near the kidneys; and adrenal secretions in turn cause a decrease in pituitary activity. Mechanisms such as those provided by hormones enable a multicellular organism to outlive the individual cells comprising it; the removal of dead cells and of other particles too large to be transported by the blood to the kidneys for excretion is carried out by white cells.

Hormones

The blood has a mechanism that prevents its being emptied through wounds. Platelets in mammals and cells called thrombocytes in other vertebrates aggregate at a wound site, and, in the plasma near them, a clot, composed of a protein called fibrin, forms. Invertebrates may rely on the motion of cells to plug their wounds.

In the brief discussion of blood and lymph that follows, it may seem as if in species most alien to man the least sophisticated fluids exist; *i.e.*, fewer and less specialized proteins occur, as well as more primitive, all-purpose cells, but this may reflect limitations of knowledge.

PLASMA

Small molecules and the regulation of water content. Water molecules, which are the main component of body fluids, probably form clusters that break up rapidly except at water-repellent surfaces (*e.g.*, as provided by fats), where the clusters are more stable and icelike. The complex mixture of fatty and nonfatty structures in cell membranes, therefore, profoundly affects the behaviour of water and the substances dissolved in it. Although water itself and certain small molecules and ions (charged particles) pass through the membranes of all active cells, certain other molecules cannot; urea, an organic compound larger than most ions, however, apparently passes into cells by disrupting water clusters.

When two solutions containing different quantities of a substance are separated by a membrane that allows only water to pass (a semipermeable membrane), water flows through the membrane to the solution containing the most substance, in an attempt to equalize the concentrations on both sides of the membrane. When a cell is placed in water, therefore, it swells, because water passes into the cell through the membrane in response to the dissolved molecules and ions inside the cell; these particles are said to exert an internal pressure, called osmotic pressure, which may cause the cell to burst, or lyse (or, in the case of blood cells, hemolyse). If a cell is placed in a solution containing a higher concentration of a substance than that found in the cell, water leaves the cell through the membrane, and the cell assumes a wrinkled, or crenated, appearance. Living cells, including the blood cells in plasma, must expend metabolic energy to maintain an osmotic equilibrium.

Cell lysis

The salt concentrations in present-day seas are higher than those within cells. Invertebrate animals capable of regulating the water content of their circulating body fluids usually lower it to about two-thirds that of the surrounding ocean; aquatic mammals also are able to perform this physiological regulation. Other animals apparently need no such regulation. The body fluids of echinoderms, for example, resemble seawater; among crustaceans, the shore crab (*Carcinus*) regulates its water content, but the spider crab (*Maia*) does not. *Rhithropanopeus* is able to swell when it must molt; its eyestalks produce a hormone that prevents water inflow at other times. Animals that cannot regulate their water content and thus tolerate only a narrow range of salt concentrations (stenohaline animals) probably are restricted to oceanic life; animals that are able to regulate their water content (euryhaline animals) are free of this restriction. Mechanisms other than those involving a circulating body fluid, however, enable certain animals to regulate their water content; *e.g.*, some coelenterates and flatworms (Platyhelminthes), which live in freshwater, utilize a diffusely branched system to remove water.

Some cyclostomes (hagfish and lampreys), among the most primitive fishes, live in the ocean, are permeable to water, and regulate their osmotic pressure. Elasmobranch fishes (sharks and rays) maintain a blood osmotic pressure above that of seawater by retaining urea, which is excreted rapidly in other vertebrates by the kidneys. The advantages to marine animals of maintaining specific osmotic pressures are not well known. Saltwater fishes lose some regulatory ability in the cold; as a result, the concentration of salts in the plasma increases, so that it acts as an "antifreeze." Fishes in freshwater also may be affected by cold, but their plasma loses ions, which means that it freezes more easily.

The ability of an animal to retain salts when moving to freshwater may have been a prerequisite to the evolution of animals capable of retaining both salts and water on land. The urine of freshwater fishes contains a low concentration of salts. Amphibians and many crabs return to water when some as yet unknown mechanism senses the need for water to dilute body fluids. A specific region in the brain of amphibians produces a hormone similar to the antidiuretic hormone of man, which regulates water retention. Among reptiles, the evolutionary success of crocodiles over dinosaurs may be attributable to the development in crocodiles of a skin more impermeable to water; certain crocodiles, however, when dehydrated, return to the water, taking in more water through the skin than by drinking.

Although a sense of thirst is necessary in all animals living on land, the composition of their internal fluids may vary. Most of the osmotic pressure in the plasma of water-dwelling crustaceans, for example, is provided by salts; half of that in land-dwelling insects is provided by organic compounds called amino acids.

Regulation of ion content in man

The regulation of the content of ions in the plasma of vertebrates, especially mammals, appears increasingly complex at higher evolutionary levels; but perhaps man seems most complex because he has been studied most. The plasma concentrations of ions such as sodium, potassium, and calcium are regulated by hormones; but human emotional stress increases the activity of the adrenal glands, whose secretions affect the transport of sodium and potassium ions across cell membranes.

Glucose has relatively little effect on the osmotic pressure of plasma, but it is an important source of energy in cells; it must be available at all times, therefore, within controlled levels. Blood-glucose levels in vertebrates are regulated by two hormones formed in the pancreas, insulin and glucagon. Insulin, synthesized when plasma-glucose concentration increases, stimulates the liver to convert excess glucose into a storage compound, glycogen. Glucagon acts in the opposite way; *i.e.*, it increases abnormally low levels of plasma glucose. Among vertebrates, either hormone may dominate in maintaining the proper levels of plasma glucose, whose concentration is profoundly affected by feeding and by stress.

Because fatlike substances do not dissolve in water and force a specific arrangement of water molecules around them, they usually circulate surrounded by protein molecules, so that their water-hating (hydrophobic) areas do not come in contact with water in the plasma.

Proteins and their various roles. There are about 20 kinds of amino acids that link together in a specific order to form protein molecules. Each amino-acid sequence is determined by heredity; in turn, the sequence of amino acids determines the shape of each protein. A mosaic-like arrangement of amino acids with positive and negative charges may comprise most of the exposed parts of a protein molecule; other amino acids, which are hydrophobic, are found inside the coil-shaped molecule and thus do not come in contact with water surrounding it. The function of a protein molecule depends on its shape.

During evolution, a chance mutation may result in the replacement of one amino acid in a specific protein molecule by another. If the newly inserted amino acid destroys the function of the protein, an organism may die; on the other hand, the loss of a protein function by mutation may be compensated for by another mutation, either in the protein molecule itself or in some other one. If an animal survives a partial loss of function, the loss may be discovered if the animal becomes ill. Many proteins function as enzymes (biological catalysts); some enzymes (*e.g.*, several blood-clotting factors) have been shown to exist only because people born without them develop specific symptoms. Mutations that do not affect function can be discovered only by analyzing proteins whose sequences of amino acids have been established. The insulin of vertebrates, whose amino-acid sequence has been established, functions only if the two functional chains comprising the molecule are not altered; a mutation in the nonfunctional section that connects the chains does not inactivate the molecule.

Species-specific enzymes

Since an enzyme catalyzes a reaction involving a specific compound, it can be identified. An enzyme whose shape must change before it becomes active or a protein involved in the structure of an organism usually can be distinguished only by crude tests. Many active enzymes function efficiently only with molecules prepared from the same species. Although others (*e.g.*, insulin, blood-clotting Factor VIII) derived from one mammalian species will function if injected into another species, they act as antigens; *i.e.*, they stimulate formation of proteins called antibodies in the injected species. These enzymes, therefore, are recognized in the injected animal as foreign in composition, if not in function. The specificity of proteins for their own hosts has limited the understanding of their functions in species remote from man. Crude separation of plasma proteins carried out in mammals other than man show that proteins such as albumin and alpha, beta, and gamma globulins occur in them.

Albumin, a rather small protein that dissolves in water, comprises more than half the total proteins found in mammalian plasma, contributes to the blood osmotic pressure, and transports certain ions and fatty acids. Yet, humans born without it appear normal. The second most abundant protein in mammalian blood is a long, large molecule called fibrinogen. When blood clots, a series of enzyme interactions converts an inactive plasma enzyme (prothrombin) into an active enzyme (thrombin) that removes two pairs of amino-acid groups from each fibrinogen molecule to convert it into a molecule called a fibrin monomer. Molecules of the fibrin monomer then link together to form strands of fibrin, the visible clot. The remaining liquid, plasma lacking fibrinogen, is called serum. Fibrinogen, prothrombin, and many other proteins are formed in the liver. In animals other than mammals, some clotting enzymes, as well as the events that trigger their interactions, may be restricted to the cytoplasmic component of specific sensitive cells (see below *Thrombocytes and platelets*).

Gamma globulins are produced by plasma cells and lymphocytes that probably occur only in vertebrates. Some other globulins carry specific metal ions, which colour them—*e.g.*, mammalian transferrin, which carries iron, is brown, and ceruloplasmin, which carries copper, is blue. The most deeply coloured globulins have a specific slot in which they carry a flat molecule (heme) containing iron. In vertebrates, four globulin-heme moieties form one hemoglobin molecule in the red cells, where it remains until the cells die. Many polychaete worms have developed a similar protein; slight difference in the structure of its heme, however, makes it green in colour. A variety of polychaete blood proteins exists. The noncirculating coelomic fluid of some polychaetes contains hemoglobin; other polychaetes have two hemoglobins—one in the plasma and another in coelomic cells. Hemoglobin predominates in young *Serpula* polychaetes, green pigment (chlorocruorin) in older ones, and another polychaete, *Arenicola*, has a very large hemoglobin molecule. Hemocyanin, a large protein molecule that contains copper and turns blue when it takes up oxygen, occurs in many mollusks, where it may be the main plasma protein. All the pigment types mentioned pick up oxygen molecules easily and release them where they are needed. A change of shape in each hemoglobin subunit accompanies oxygen uptake or release and is transmitted to the other units in a way that facilitates the process (see COLORATION, BIOLOGICAL).

Hemoglobin

FORMED ELEMENTS

Red cells (erythrocytes). The hemoglobin in the plasma of some worms allows them to carry about ten times more oxygen than can an equal volume of seawater. The localization of hemoglobin in red cells increases this capacity another four times. The evolution of cells with such a singular function is traceable by arranging species in order of the increasing oxygen-carrying efficiency of their red cells. The mature red cells of mammals lack nuclei. Not only does a nucleus take space; it also requires oxygen; mammals, therefore, have the most efficient red cells, at least so far as oxygen-carrying ability is concerned. It thus is tempting to presume that, if no further evolutionary progress occurs, even though mutations continue, a maximum efficiency has been attained. Indeed, one can say that in the following aspects, evolution of red cells apparently stopped long ago: the total volume of red cells in proportion to the total volume of blood varies relatively little among mammalian species; in addition, all normal red cells have about the same high concentration of hemoglobin, above which the molecules would soon interfere with each other or even form crystals. On the other hand, the diameter of red cells varies greatly, ranging from less than 3 to more than 9 microns (1 micron is 0.001 millimetre). Both Indian and African elephants have large red cells (9.2 microns in diameter), but no clear relationship between red-cell size and total body size exists among other species. Red cells apparently became smaller as life on land developed or evolution in general progressed, but the relationship between red-cell size and evolutionary advancement is not clear. The supposedly primitive hagfish, for example, has red cells 26.4 microns long; those of the related lamprey are only 14.3 microns. The red cells of many sharks, about as large as those of some amphibians, are larger than those of many other fishes.

Evolution and red-cell size

The disk-shaped mature erythrocytes of most mammals have thick, rounded edges and lack nuclei. When such a cell is cut in half, both halves may assume a disk shape and resemble small red cells. Certain membrane components in the smallest red cells (two to four microns in diameter), which occur in ruminants such as cows, occur in a proportion different from that in other mammalian red cells. The fact that red cells are thin in the centre may increase their efficiency in gas transport. The efficiency of gas exchange and transport may be increased also by the fact that the shape of red cells can be distorted into cones, clubs, and dumbbells as they pass through the extremely narrow blood capillaries.

Perhaps red and white blood cells originate in the same cells in bone marrow. The need for red cells is conveyed in mammals by a protein (erythropoietin). The protein is formed by the action of a substance (erythrogenin), synthesized by the kidneys when they lack oxygen, on another molecule (erythropoietinogen), which is synthesized in the liver. A severe loss of red cells and the subsequent decrease in available oxygen is normally followed by a flood of young red cells, which still have nuclear material. The mature red cells of mammals contain many enzymes important for their maintenance and partly responsible for their death when they fail; the oldest cells, therefore, should die first, and they do in some mammals (*e.g.*, mouse, dog, and man); in other species, however, red-cell death may occur at any age in an apparently random manner. Although it has been suggested that red cells must pump out sodium ions and retain potassium ions to survive, dogs and some other mammals have about the same proportions of sodium and potassium ions in both red cells and surrounding plasma; yet their cells live no longer than do those of man. On the other hand, the less heat a vertebrate species produces, the longer its red cells live: turtle erythrocytes, for example, live longer than those of man; red cells of birds and mice have a shorter existence.

Extreme differences between sodium- and potassium-ion concentrations in erythrocytes exist in various species (see Table) and even within breeds in one species.

White cells. All poorly pigmented cells normally found floating in blood, hemolymph, or coelomic fluid may be called leukocytes, although those in the hemolymph and coelomic fluid of invertebrates are often called hemocytes and coelomocytes, respectively. White cells in hemolymph and coelomic fluid include cells that eat (phagocytize) foreign particles, as do the granulocytes of mammals; cells that recognize and chemically attack only specific foreign matter, as do the lymphocytes and plasma cells of mammals; and cells that clump together when disturbed, as do the thrombocytes. The distinctions among functions and blood-cell types of animals other than man, however, have not yet been clearly established.

Concentration of Potassium and Sodium Ions in Plasma and Red Cells of Certain Vertebrates

	potassium ions*		sodium ions*		percentage of red cells replaced per day
	red cells	plasma	red cells	plasma	
Man	95	4	19	138	1
Dog	8	4	97	143	1
Chicken	119	6	18	154	3

*Milli-equivalents per 1,000 millilitres.

A superficial resemblance exists among some mammalian blood cells and those of other animals. Some white cells of sea cucumbers (Holothuroidea), for example, resemble mammalian white blood cells called basophils; gastropod mollusks have cells that resemble lymphocytes and granular amoebocytes that resemble certain granulocytes of man (eosinophils); crustaceans have coelomocytes that clump around foreign cells and eat them; and insects have a multitude of hemocytes, some of which resemble human white cells while others do not. The evolution of insects and other arthropods diverged from that of man and other chordates so long ago that the blood cells of these two animal groups do not provide any evidence of a common ancestry. The most active white cells in insects, called plasmatocytes, can spread either over surfaces of parasites, forming membranous capsules, or over foreign particles; often they become granular. Hemocytes, most abundant in the last immature stage (larval instar) of insects, decrease in numbers during the nonfeeding stage called pupation, after which, in certain insects, new cells are formed. When a pupa is injured, hemocytes appear to seal the wound; actual clotting in insects, however, is initiated by cells called cystocytes (coagulocytes). When hemolymph flows from a wound, cystocytes enlarge, break down, and either turn the plasma around them cloudy or form threads around which the plasma jells. Other insect hemocytes show no identifiable activity.

Plasmatocytes

Echinoderms and worms also contain blood cells very different from those of man. Echinoderm groups differ in their types of blood cells; all, however, have phagocytes to eat foreign particles, and other hemocytes to repair damage and to seal wounds. Echinoderm blood cells probably form from stem cells (so-called lymphocytes); active movement of cells within the heartless hemal system may transport food throughout the body. Hemoglobin-containing cells occur in some burrowing echinoderms. The variety of cells in hemal and coelomic fluids among echinoderms and even within various areas of one individual possibly reflects a range of requirements totally different from those of man, so that there are no blood kinships between man and echinoderms. Among vertebrates, relationships are clearer; *e.g.*, fishes and amphibians have granulocytes, lymphocytes, and, probably, a type of white cell present in human blood and referred to as a monocyte.

Thrombocytes and platelets. The thrombocytes of fishes, amphibians, reptiles, and birds resemble the cells that aggregate at wounds in invertebrates; in vertebrates, however, the plasma apparently has more responsibility in the clotting process. Since no thrombin (an enzyme involved in clotting) forms when thrombocytes are removed from the plasma of fishes, these cells, which break down on contact with a wound, must liberate substances that convert plasma prothrombin (an enzyme precursor) to thrombin, which then converts plasma fibrinogen into fibrin strands (a clot). In contrast to other vertebrates, the plasma of mammals contains an inactive protein, Factor XII, also called the Hageman factor. Mammalian Factor XII, upon contact with some hard foreign surface, becomes an active enzyme that initiates clotting. The contact-sensitive thrombocytes of other vertebrates are analogous to small disks of cytoplasm (platelets) in mammals. Platelets are derived from large bone-marrow cells (megakaryocytes). Although platelets aggregate in plasma under a variety of conditions, which also cause the aggregation of thrombocytes, and their clumping, or sticking, is accompanied by the liberation of several important substances, they contribute to the clotting of plasma only by making available a chemical component of their membranes (phospholipids). Platelets, however, are essential for sealing wounds in mammals.

Factor XII

Lymphocytes and lymph. Animals must recognize their own cells, so that they do not eat them. The system by which vertebrates protect themselves from foreign cells (*i.e.*, development of specific immunities), however, has not yet been found in the invertebrates. Hagfish appear to be the most primitive animals that have lymphocytes, form antibodies, and reject tissue grafts. From the higher fishes to man, lymphocytes and plasma cells become progressively more abundant, antibodies more specific, and tissue grafts less successful.

As blood flows through the capillaries, dissolved substances pass between the blood and the surrounding tissues. Most of these substances pass readily through the capillary walls, but a small amount of protein, once forced from the plasma, cannot return because the protein molecules are too large. This protein and waste fluids pass from the cells into the lymph vessels, which have much more permeable walls and within which the pressure is much lower than that in the capillaries. The fluids are pushed through the vessels by the motion of organs and muscles, their flow directed by valves along the way; in fishes, amphibians, reptiles, and some birds, a heartlike organ (the lymph heart) helps the circulation. As it flows, the lymph passes through many glandlike masses of tissue, called lymph nodes, and through the thymus, a glandular body present in all vertebrates. Birds have an extra lymphatic organ, the bursa of Fabricius. In all vertebrates, the lymph vessels combine, finally reaching a duct (*e.g.*, the thoracic duct in mammals) from which the lymph enters the venous bloodstream.

In very young animals, lymphocytes are formed in the thymus; later they form from cells in the bone marrow and are transported to, and deposited at, certain nodes or organs. The fate of each lymphocyte depends on the information or stimulation it receives at its point of deposition. The lymphocytes in gut-associated lymph tissues or those in the bursa of birds, for example, are apparently prepared to form antibodies if they are needed. In mammals, a lymphocyte at a lymph gland may be stimulated by a molecule (antigen) to form a specific antibody. Certain lymphocytes first form a globulin known as immunoglobulin M (IgM); then the lymphocytes divide, and each daughter cell receives some IgM. If stimulated again by more antigen molecules, the lymphocytes again divide, liberating their IgM and transforming into plasma cells that form a smaller globulin (IgG). Some antigen molecules probably remain with certain cells in the lymph node and eventually stimulate new lymphocytes passing the lymph node. Speed, intensity, and complexity of these reactions apparently increase from cyclostomes through elasmobranchs to bony fishes. Most of the globulins of fishes probably resemble human IgM. Amphibians, however, may be more advanced; bullfrogs form IgM, then IgG, or IgG first, depending on the type of antigen molecule injected. The antibodies formed by a bullfrog injected with serum of one mammalian species react with the serums of other mammals; the antibodies of mammals, however, are more specific. The serum of a rabbit immunized to serum of one fish species, for example, reacts less with serums of other fishes; in fact, the extent of interaction of the serums serves as a measure of evolutionary relationships among species.

Fate of lymphocytes

In vertebrates the thymus plays a role in determining which lymphocytes survive, apparently stimulating their division and then destroying undesired ones. Lymphocytes may spend only a few hours at a time in the blood and many months elsewhere. The variety of lymphatic systems among vertebrate species probably has some relationship to the immunological functions required for each. The ability of man's immunological system to distinguish differences in blood-group substances means that there are many blood incompatibilities; *i.e.*, few humans, unless they are twins, have completely identical sets of blood groups. Perhaps the thymus destroys cells that

know too much; *i.e.*, cells that would form antibodies against substances or tissues that are useful (see IMMUNITY).

HOMEOSTASIS

A large number of lymphocytes constantly are mobilized to fight random invasions in animals; the relatively low number of them that briefly reside in the blood at any time, however, remains remarkably constant. The complicated feedback systems that maintain a constant internal environment are called homeostasis. Homeostatic mechanisms mute the changes between rest and action or starving and eating, so that they provoke only small changes in the amounts of salts in cells and in the blood. The things that remain constant only with great effort are most precious to survival. These efforts result in variations in blood among different species; homeostasis, however, creates the uniformity within a species that allows distinctions to be made.

BIBLIOGRAPHY. E.C. ALBRITTON (ed.), *Standard Values in Blood* (1952), with comprehensive tables; F.M. BURNET, *Cellular Immunology*, 2 vol. (1969), an essential work; O.F. KAMPMEIER, *Evolution and Comparative Morphology of the Lymphatic System* (1969), a classic, but almost exclusively morphological; R.G. MACFARLANE and A.H.T. ROBB-SMITH (eds.), *Functions of the Blood* (1961), a unique, somewhat outdated collection of comparative physiological studies; L. VROMAN, *Blood* (1967), a readable account.

(L.V.)

Blood Circulation, Human

The circulatory system, in which the blood is propelled throughout the body, consists of four types of hollow containers: (1) the heart, or the central pump; (2) the arteries, or efferent vessels, delivering the blood to the tissues; (3) the capillaries, representing a fine network of very small vessels contained within the body tissues; and (4) the veins, larger vessels, returning the blood back to the heart. The quantity of blood within this closed system—the circulating blood volume—is, in the average adult, 5–6 litres, of which 60 percent is the liquid medium, plasma, and 40 percent the solid part, the blood cells. The entire amount of the circulating blood passes through each ventricle (lower chamber) of the heart every minute. The circulation consists of two complete cycles: the greater circulation, termed the systemic circulation, with vessels serving almost all the tissues of the body; and the lesser circulation, the pulmonary circulation, which distributes blood to the respiratory membranes of the lungs. The heart is thus a double pump, consisting of its right side, ejecting blood into the pulmonary artery and supporting the pulmonary circulation, and its left side, ejecting the blood into the aorta, the main artery that supplies the systemic circulation. The blood returning from the systemic circulation is drained into the right side of the heart and then pumped into the lungs; that returning from the pulmonary circulation enters the left side of the heart from which it is ejected into the systemic circulation. Thus the two cycles are hooked together in series. The large vessels emerging from the two cardiac ventricles are called the great vessels or the arterial trunks. Each one divides into smaller and smaller branches, eventually ending in the microscopic capillaries. The pulmonary artery divides directly and immediately into its branches, but the aorta takes a long course, with many side branches, before its final division into two arteries, one for each leg. Many of the side branches lead to semi-autonomous subdivisions of the systemic circulation.

Transport functions

There are many transport functions of the circulatory system; its design, however, as well as the composition of the blood contained in it, facilitates the most important function of the system—that of mediating the process of respiration. Every living cell of the body "breathes"; that is, consumes oxygen and emits carbon dioxide. Oxygen is carried by the blood from the heart to all tissues of the body, where, in the capillaries, an exchange takes place: oxygen is taken up according to tissue needs and, in return, carbon dioxide enters the blood. The blood then, poorer in oxygen but with a higher content of carbon dioxide, is carried to the right side of the heart and thence to the lungs, where another exchange takes place: oxygen is replenished and carbon dioxide given up, to be exhaled by the process of breathing. The relative amounts of oxygen and carbon dioxide carried in the blood form the basis for the division of blood into fully oxygenated, bright red, arterial blood contained in the arteries, and deoxygenated, dark red, venous blood, with a higher carbon dioxide content, contained in the veins. It should be understood that this description of arterial and venous blood applies only to the systemic circulation, because within the smaller pulmonary circuit the situation is reversed: the pulmonary artery contains the deoxygenated blood and the pulmonary veins contain the fully oxygenated blood.

The role of circulation in acting as a vehicle for the exchange of the two respiratory gases is the *sine qua non* of life. Body tissues cannot survive long without oxygen, and sensitive tissues, such as the brain cells, die when deprived of oxygen for more than four minutes. The circulation, however, plays a part in many other vital processes. As the principal vehicle of communication between parts of the body, blood transports chemical substances from one place to another; for example, nutrient substances (the breakdown products of carbohydrates, fats, and proteins) from their points of absorption to metabolic pools or storage depots, and from there to the users of this fuel. Hormones secreted into the blood by glands are distributed where needed. Waste products (other than carbon dioxide) are transported to the sites of their excretion or destruction, such as the kidneys, liver, and lungs. Blood also is a principal factor in the maintenance of body temperature. Finally, there is one other important function of the circulation; namely, to maintain the proper environment for metabolic processes by guarding the composition of the tissue fluid, which is separated from the blood plasma by an extremely thin membrane that permits a free-flowing exchange of water and soluble substances.

The circulation works with a remarkable economy: a system of priorities is established by the body, delivering blood to places where it is most needed at any given moment. The many built-in safety mechanisms and reserve capacities of the circulation provide for its proper function, even under adverse conditions, with a minimum expenditure of energy.

THE CENTRAL PUMP

The work of the heart

The workload of moving the blood within the circulatory system rests largely with the two central pumps—the right and the left sides of the heart, working synchronously but independently of one another. The work of the heart is considerable: in an average adult male each ventricle ejects six litres of blood per minute (about five in the female), or 720 litres per hour, 8,640 litres in 24 hours. Since the work performed by a pressure pump is roughly the product of its output and the pressure it builds up, the left side of the heart, supporting the high-pressure systemic circulation, performs a workload about five times greater than the right side, which empties into a low-pressure system. Consequently, the number of muscle cells within the left ventricle is proportionately larger than that in the right ventricle, and the muscle itself is thicker. The pumping action of the heart depends upon the rhythmic contraction of the two ventricles, followed by their relaxation. Contraction, or systole, of the muscles of the ventricles reduces the size of their cavities, ejecting the blood in them into two arterial trunks (into the aorta from the left ventricle and into the pulmonary artery from the right ventricle); relaxation, or diastole, of these muscles enlarges the cavities, drawing blood into the ventricles from the atria (upper heart chambers). The two atria also contract rhythmically, but their pumping action is only an auxiliary factor in the filling process of the ventricles.

Rhythmicity of the heart. The heart is composed largely of muscle cells that have contractile properties; *i.e.*, have the capability, by response to a stimulus, of

shortening their length, thus exerting considerable force. In addition to the large number of these cardiac muscle cells, the heart also contains a special type of muscle cell

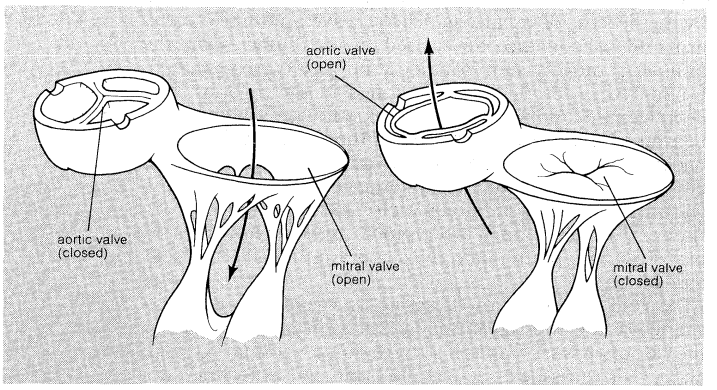


Figure 1: Cardiac valves: the mitral and aortic valves are shown in the closed and open positions.

The conducting system

that self-generates rhythmic impulses. These cells are concentrated in groups, forming identifiable structures within the inner surface of the heart. The sum total of all these structures is termed the conducting system, and it consists of: (1) the sinoatrial node (also known as the sinus node), (2) discrete pathways carrying the impulses along the atrial wall, (3) the atrioventricular node (A–V node), (4) the bundle of His, (5) the bundle branches and their secondary distributions, and (6) the Purkinje network. The sinoatrial node is a small structure, 10 × 3 mm, located in the right atrium near the mouth of the superior vena cava. The conducting pathways within the atrium are microscopic structures and have only recently been identified. The atrioventricular node is a flask-shaped structure, 8 × 6 × 4 mm, located at the junction of the interatrial septum (the partition between the two atria) and the back wall of the right atrium, near the orifice of the coronary sinus. From this node emanates the bundle of His, 2–4 mm in diameter and about 2 cm long, which runs along the interatrial septum and eventually enters the ventricular septum, where it divides equally into two bundle branches. The left bundle branch runs along the septal wall to the left ventricle for a short distance and then divides into two branches, each of which in turn subdivides into smaller branches. The right bundle branch runs along the septal wall of the right ventricle without subdividing until it reaches the apex of the heart, where many secondary branches occur. The terminal subdivisions of the two bundle branches connect with a fine network of fibres termed the Purkinje network, which covers the inner lining of both ventricles. This network transmits the conducted stimulus to the muscle, producing its contraction.

The specialized muscle cells of the conducting system differ in physiological properties from the remainder of the heart muscle cells by two features: (1) they are capable of self-excitation—*i.e.*, they generate rhythmic impulses; and (2) they are capable of conducting these impulses at rapid speeds. The sinoatrial node has the fastest rate of impulse formation. It is the primary pacemaker, where, under normal conditions, the heart beat originates. This node is richly supplied by terminals of the autonomic nervous system, which can transmit signals accelerating or slowing the basic rate—about 70 beats per minute—generated by the nodal cells. Under the inhibiting influences of the parasympathetic branch of the autonomic system, the rate may slow to about 40 per minute; sympathetic influences can accelerate the pacemaker to a rate of 200 per minute. The dominant influence is that of the sympathetic nervous system (accelerator nerves), which speed up the heart during exercise, excitement, fever, and in response to many other stimuli.

The atrioventricular node also has its inherent rhythm. Its impulse formation occurs at the rate of 50 beats per minute, but this impulse-generating function is not utilized under normal conditions. It is a standby, reserve pacemaker, ready to take over the rhythmicity of the heart only if the sinoatrial node fails to fire or slows down excessively. Its principal role lies in conducting the impulses coming from above and in delaying these impulses to permit good coordination between the contractions of the atria and ventricles. It also filters out (in a process sometimes called the triage effect) some undesirable or excessive impulses that might otherwise reach the ventricles.

The role of the atrioventricular node

The bundle of His, its bundle branches, and the Purkinje network have their own automaticity, which, however, is very slow, about 30–40 beats per minute. They represent the last line of defense, initiating slow contractions of the ventricles when both the upper nodes fail to fire or when the ventricular network becomes disconnected from the atria (heart block).

The normal impulse, originating in the sinoatrial node, activates the atria, producing their contraction about 0.02 second later. It reaches the atrioventricular node about 0.04 second after its origin, and then is conducted slowly through this node, emerging 0.1 second later. From this point on, the impulse is conducted rapidly throughout the bundles and the Purkinje network, producing contraction of the ventricles about 0.16 second after the contraction of the atria. The activation of the ventricle proceeds in a specific order producing ventricular contraction and facilitating a most efficient ejection of the blood into the arterial trunks.

The cardiac cycle. The cardiac cycle represents the complete sequence of contraction and relaxation of the heart. It is customary to divide the cycle according to the state of the ventricles: ventricular contraction, or systole, and ventricular relaxation, or diastole. At ordinary rates (70 per minute), systole occupies 40 percent, diastole 60 percent of the cycle. At faster or slower rates this ratio changes: above 100 beats per minute, systole lengthens to up to 50 percent of the cycle, and at slower rates, such as under 50 beats per minute, it shortens to 25 percent of the cycle.

The valves of the heart

The movement of blood during contraction and relaxation of the ventricles is made possible by the action of two sets of cardiac valves. The atrioventricular valves separate the ventricles from their respective atria. The left atrioventricular valve is called the mitral valve, and the right, the tricuspid valve. These valves are fairly thick structures attached to the orifice separating the left ventricle from the left atrium (mitral orifice) and to that separating the right ventricle from the right atrium (tricuspid orifice). The free edges of the three leaflets of the tricuspid valve and of the two leaflets of the mitral valve have attached to them a large number of thin chords (chordae tendineae) that extend down to muscular pyramids (papillary muscles) on the inside of each ventricle. During ventricular contraction higher pressures close the atrioventricular valves, but, in addition, contraction of the papillary muscles, participating in the general contraction of the ventricles, pulls the chordae tendineae taut, causing even firmer closure of the atrioventricular valves, thus enabling them to withstand the high pressure inside the ventricle during systole.

Two semilunar valves separate the ventricles from their arterial trunks: the aortic valve separates the left ventricle from the aorta and the pulmonary valve separates the right ventricle from the pulmonary artery. These valves are much thinner than the atrioventricular valves. Each consists of three flaps, or leaflets: thin, pocket-like structures attached to the orifices which separate each ventricle from its arterial trunk. When the valves are open, the valve leaflets float free in the stream of blood near the wall of the beginning parts of the aorta and pulmonary artery. Closure of the valve is produced at the start of ventricular relaxation when the column of blood in the aorta flows backward toward the ventricle and fills the pockets of the leaflets, pressing them tightly closed.

The two sets of valves generate acoustic vibrations (heart sounds) that can be heard by means of the stethoscope. The first heart sound occurs at the onset of systole, coinciding with closure of the atrioventricular valves. Although there is some controversy concerning the exact

mechanism of the first heart sound, it is generally accepted that its major component is produced by the sudden tensing of the atrioventricular valve leaflets, the chordae tendineae, the papillary muscles, and by the valve closure. The second sound is produced by closure of the semilunar valves, thus signifying the onset of diastole.

The onset of systole occurs when the atrioventricular valves are wide open and the atria and ventricles functionally act as a single chamber on each side of the heart. The pressure in this large chamber is 3 mm in the right side and 6 mm in the left side. The onset of systole raises the pressure in the ventricles, promptly shutting the atrioventricular valves and thus separating the atria from the ventricles. During the initial part of systole, tension in the ventricle rises and elevates the pressure, but the muscular fibres do not yet begin to shorten because both sets of valves are closed and no movement of blood can take place. When the pressure inside the left and right ventricles, measured in millimetres of mercury (Hg), exceeds that in the aorta (80 mm) and in the pulmonary artery (10 mm Hg) respectively, the semilunar valves are forced open, and the blood can be ejected. The initial period of systole, or rising tension without flow of blood, is referred to as the isometric (isovolumetric) contraction period and lasts about 0.03 second. The next period, the ejection of blood into the arterial trunks, takes place with the semilunar valves wide open. Now the ventricles and their arterial trunks have identical pressures, acting functionally as a single chamber. The ejection of blood, however, does not take place at an even rate. The initial ejection period transfers about half of the blood in the first quarter of the total ejection period and is termed the rapid ejection phase, lasting about 0.07 second. The remainder of systole involves a slower rate of ejection, with the other half of the blood entering the arterial trunks in the following 0.22 second. This is known as the slow ejection phase.

Ventricular relaxation (diastole) begins with the closure of the semilunar valves. The relaxing ventricular muscles decrease their tension, reducing intraventricular pressure, and thus cause the semilunar valves to close. No movement of blood is possible until the atrioventricular valves open. The period of fall in pressure without movement of blood is called the isometric (isovolumetric) relaxation phase and is analogous to the isometric phase of systole. This period lasts about 0.04 second. When the pressure falls below that in the atria, it forces the two atrioventricular valves open, and the blood is drawn into the ventricles from the atria. As in systole, the flow of blood into the ventricles does not occur at an even rate. The first part of the ventricular filling period (the rapid filling period) draws some 60 percent of the blood in about 0.10 second. The next period of the diastolic phase is the rest period, or diastasis, in which the rate of flow is diminished—only 10–15 percent of the blood enters the ventricle within about 0.18 second. The final period of diastole is the atrial contraction phase: the atria, having just received their excitation signals from the sinoatrial node, contract and eject the final portion of blood, 25–30 percent of the total, into the ventricles.

None of the cardiac chambers empties completely during the cardiac cycle. The ventricles eject one-half to two-thirds of their contents during systole; the atria are connected with the large venous reservoir from which they continuously receive blood, so that their blood content shows relatively little variation.

Cardiac output. The cardiac output is defined as that quantity of blood ejected by a cardiac ventricle into its arterial trunk in one minute. The quantity of blood expelled with each ventricular contraction is termed stroke volume or stroke output. The normal, average stroke output is 75 millilitres of blood; therefore, with a heart rate of 70 to 80 beats per minute, the cardiac output is between five and six litres per minute. The output into the two arterial trunks by the ventricles is identical under normal conditions; thus each ventricle ejects five to six litres of blood per minute. These figures apply to the basal state of the body; that is, to cover the minimal metabolic needs of an individual resting comfortably and motionlessly in bed. Since the principal role of the circulatory

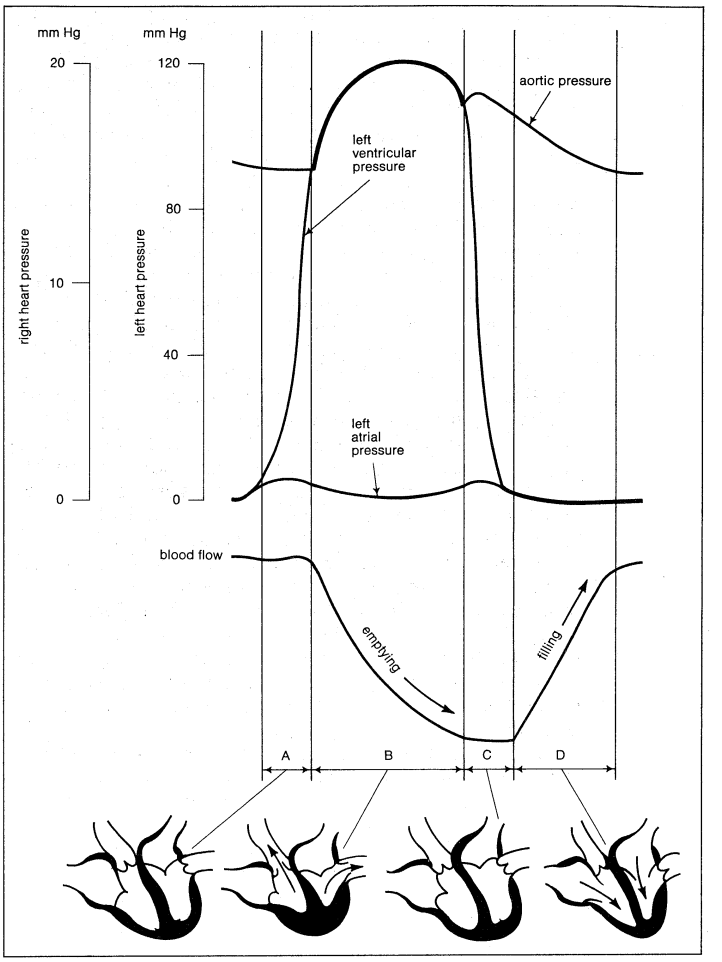


Figure 2: *Idealized curves of pressure and blood flow during the cardiac cycle, shown along with a presentation of the position of the cardiac valves.*
(A) Isometric contraction phase with all valves closed. (B) Ejection period of systole with semilunar valves open, atrioventricular valves closed. (C) Isometric relaxation phase, all valves closed. (D) Diastolic filling phase with semilunar valves closed, atrioventricular valves open. Two pressure scales are presented, to account for the systemic and pulmonary circulations.

system is the transport function of the blood, regulation of cardiac output represents one of the most intricate processes in the body.

Oxygen requirements of the body

The primary function of the circulation is the delivery of oxygen. It is customary to express the metabolic demands of the body in terms of oxygen requirement. The basal resting state, as described above for an average adult, requires the delivery of 250 millilitres of oxygen per minute. The quantity of blood delivered to the systemic circulation is six litres per minute; thus each litre of blood gives up about 40 ml of oxygen to the tissues. This figure represents the arteriovenous oxygen difference, or the difference in oxygen content between the blood entering the tissues and the blood leaving the tissues. The arteriovenous difference represents the index of oxygen utilization by the tissues and may be applied to the body as a whole or to any portion of the circulatory system. The normal arteriovenous oxygen difference of 40 ml per litre of blood represents only a small fraction of the available oxygen in the blood, which is about 180 ml per litre. At rest, therefore, tissues use less than 25 percent of the available oxygen in the blood perfusing them. This relationship can be expressed in another way: the arterial blood ejected into the systemic circulation is about 95 percent saturated with oxygen, and the mixed venous blood, returning to the right side of the heart after having perfused all tissues, still is 75 percent saturated with oxygen.

The importance of the regulation of cardiac output in response to changing metabolic demands of the body can best be appreciated when the high oxygen demands of the

body are recognized. If an individual, resting under basal conditions, begins to move about in bed or gets excited, oxygen demands may double—from 250 ml to 500 ml per minute. Mild exercise, such as walking at a slow pace across the room, may triple the oxygen demands. Strenuous exercise may increase the oxygen needs to two or three litres per minute—8–12 times the basal level. Some athletic feats demand four to six litres of oxygen per minute. How can the circulatory system solve the problem of delivering 20 times more oxygen than the basal value? First of all, exercise constitutes an emergency, during which oxygen reserves in the blood are more fully utilized. By extracting not just 25 percent but all the available oxygen from the blood, one litre of oxygen per minute, instead of 250 ml, becomes available. More oxygen than that may be needed and can only be delivered by increasing cardiac output. The heart is capable of increasing its output up to about five times its resting level, and is thus able to satisfy the demands imposed by strenuous exercise. These two factors, better utilization of available oxygen and increased cardiac output, work together for lesser degrees of metabolic needs. For example, if the oxygen demands double, utilization is likely to increase by 50 percent and cardiac output by 50 percent. Proportional contribution of the two factors satisfies the body demands in all situations. One of the principal reasons for the low level of maximum possible effort in a sedentary individual, and the very high level in the athlete, is the difference in efficiency with which the two circulatory systems can adapt themselves to the delivery of the increased amounts of oxygen required by the tissues.

It is obvious that large increases in cardiac output can take place only if more blood returns to the heart, inasmuch as the reservoir capacity of the heart itself is very limited. The quantity of blood returning to the right side of the heart from the systemic circulation is termed venous return and is the principal determinant of cardiac output. The venous return, in general, must equal the cardiac output, otherwise the circular movement of the blood could not take place. Momentary inequalities, however, frequently occur between venous return and cardiac output. The ability of the heart to handle the excess of blood returning to it rests with the adaptive mechanisms expressed by the "law of the heart" of Ernest Starling. This law postulates that when cardiac muscle is stretched, the resulting contraction becomes stronger in proportion to the degree of stretching. Thus the stretching, or increase in muscle fibre length, resulting from more blood entering the ventricle leads to a more forceful ventricular contraction, which, in turn, will expel a larger amount of blood. Conversely, a smaller venous return, by reducing the diastolic fibre length, will lower the stroke output. This mechanism can be compared to a bow and arrow: the more tension exerted upon the bow, the stronger the force generated, sending the arrow farther. A measurable expression of the degree of filling and of the fibre length during diastole is the end-diastolic filling pressure, or the pressure in a ventricle at the beginning of systole. A higher filling pressure will increase stroke output; lower pressure will decrease it. A recognized way of testing the function of the heart is to obtain the curve showing the relationship between the filling pressure and the cardiac output (*e.g.*, during distension of the heart by a blood transfusion). Such a cardiac function curve can characterize the performance of any heart, permitting the distinction between normal and abnormal performance, and also allows a study of the effect of drugs and other agents upon cardiac function.

Starling's "law of the heart"

An auxiliary mechanism aiding the regulation of cardiac output is the capability of the heart to empty itself more completely. Under basal conditions the volume in each ventricle is about 150 ml of blood, and at times only half of it is ejected with each beat. Under conditions of increased demand, the heart can reduce the residual volume (that remaining in a ventricle at the end of systole) to 10–20 ml. This stronger contraction—beyond that required to fulfill Starling's law of the heart—has been termed homeometric autoregulation of the heart, and is brought about mainly by the action of regulatory cardiac hormones (catecholamines).

Starling's law of the heart holds whenever there is an increase in both the heart rate and the stroke output. Produced by autonomic nervous reflexes, an increase in heart rate from the resting level of 70 beats per minute to the maximum rate during exercise of about 200 can increase cardiac output threefold. Further increase must involve augmentation of the stroke output. It is noteworthy that the resting heart rate is one of the determining factors of physical fitness: trained athletes often have resting rates as low as 40 to 50 beats per minute and can increase their cardiac output five times merely by accelerating the rate to 200. This is in addition to any further increase brought on by augmentation of the stroke output.

The process of adaptation of cardiac output to metabolic needs of the body has been discussed in terms of strenuous exercise, which imposes the heaviest demands upon the circulation. Obviously, between maximal demands and the basal state, there is a wide spectrum, and lesser degrees of exercise will invoke correspondingly smaller changes in cardiac output. Other factors influencing cardiac output include many situations: standing, which reduces venous return by the pooling of blood in the lower part of the body, lowers cardiac output by about 20 percent; fever increases metabolic demands and, consequently, cardiac output; high environmental temperatures may step up cardiac output. Excitement, postprandial digestive processes, high altitudes (lowered oxygen content in the air)—all increase cardiac output by invoking some of the previously described adaptive mechanisms.

THE SYSTEMIC CIRCULATION

The systemic circulation is initiated by the ejection of blood by the left ventricle into the aorta. The blood flows from there to every organ and tissue of the body (including even the lungs, which derive their nutritional blood supply from the systemic circulation; the working blood of the lungs, that to be supplied with oxygen, represents the pulmonary circulation). The systemic circulation at any given time contains about three-quarters of the total blood volume; the remaining quarter is contained in the heart and the pulmonary circulation. The systemic circulation consists of three divisions: the arterial system, the capillary system, and the venous system. The volume of blood contained in the systemic circulation is divided as follows: 20 percent is in the arterial system, 5 percent in the capillary system, and 75 percent in the venous system, which represents the great reservoir of blood.

Some hemodynamics

The arterial system. The arterial system delivers blood to all parts of the body under high pressure. Such pressure is essential for the proper perfusion of all organs and tissues. High pressure permits the proper distribution of the blood to the various tissues according to their needs. The arteries represent a system of elastic tubes (capable of increasing or decreasing their calibre), which become more numerous but of smaller calibre as they subdivide along their course. The flow of liquid through tubes is characterized by Poiseuille's law: (1) $V = \frac{dP \cdot r^2}{8\mu l}$ in which dP is the difference in pressure between two points of the circulation (pressure gradient), V is velocity of blood flow, r is the radius of the tube, μ is the viscosity of the fluid, and l is the length of the tube. In adapting the equation to conditions prevailing in the circulation of blood, the following further equations can be derived: (2) $R = \frac{8\mu l}{\pi r^4}$ (R = resistance). This indicates that resistance to flow in a vessel is directly proportional to the length of the vessel and the viscosity of blood and inversely proportional to the fourth power of the radius of the vessel. The final derived equation depicts the relationship between the flow (*e.g.*, cardiac output), pressure, and resistance: (3) $Q = \frac{dP}{R}$, or simply (4) $P = QR$, in which P is pressure, Q is output or volume of flow in a region, and R is resistance. This equation is analogous to Ohm's law, which states the relationship between electri-

cal currents, electromotive force, and the resistance of a conductor.

Resistance to blood flow represents the loss of energy through friction resulting from blood "squeezing" through a small vessel. Equation 4 indicates that pressure is proportional to resistance; *i.e.*, high pressure in the systemic circulation correlates with high resistance within the system. The resistance to flow by the aorta and the larger, and even the smaller, arteries is negligible. These vessels represent a high pressure system—a compression chamber (the German term *Windkessel* is often used in this connotation), an elastic bag, as it were, in which the high pressure is maintained by the resistance offered by the small vessels guarding the exit from the system. These vessels, the arterioles, are the smallest arteries, with an abrupt reduction in diameter from the larger ones. Arterioles have thick, muscular walls capable of constricting and relaxing. The arterioles are ordinarily in an intermediate state of partial constriction (tone of the vessel), ready to constrict further to increase their resistance, or dilate to reduce their resistance, as the need arises. Thus the energy, built up by the contraction of the left ventricle and stored in the arterial compression chamber as high pressure, is dissipated in overcoming the friction of the high-resistance arterioles. The normal systolic pressure of 120 mm Hg in the left ventricle, identical with that in the aorta, remains almost unchanged throughout even the small arteries. In the arterioles, however, the pressure drops promptly; entering the arterioles it lowers slightly to about 100 mm Hg; emerging from them it drops to about 30 mm Hg.

Equation 4, indicating that pressure is the product of flow (output) and resistance, applies to the entire systemic circulation. By resistance, then, is meant the sum total of the resistance in all arterioles in the body. This is expressed by the equation: (5) $R = R_1 + R_2 + R_3 + \ldots R_n$. It follows that the flow through each region of the body is also determined by the local resistance, so that a relaxation of a group of arterioles in any organ would automatically increase the blood flow through that organ. This principle is of fundamental importance in the process of the regulation of blood pressure as a whole, as well as the distribution of blood to areas where it is most needed, as will be discussed later.

The arteries—the elastic compression chamber—play an important role in the efficiency of the circulatory system. If blood were ejected from the heart into a rigid system of tubes, the intermittent pumping effect of the heart would be retained in the circulation. The blood would move forward during systole, but the flow would stop during diastole. Furthermore, the pressure during diastole would fall to a very low static-pressure level, which would depend upon the relationship of the volume of the blood to the capacity of the arterial system. The elasticity of the aorta and the other large arteries, however, makes the walls of these vessels distend during systole. The energy stored in the distended arteries is returned to the circulation during diastole: the pumping action then temporarily ceases, the arterial walls return to their former size, and the energy is used to propel the blood forward, making the flow continuous.

The pulse

The ejection of blood into the elastic compression chamber thus influences the flow of blood; it also determines the pressure sequence in the system. A curve, recording pressure in the arterial system, is called the arterial pulse curve. The first portion of the pulse curve, its systolic part, is identical with the corresponding part of the left ventricular pressure curve. The end of systole, the closure of the aortic valve, is signalled in the pulse curve by a sharp notch—the dicrotic notch—after which a gentle further fall in pressure takes place, related to the blood leaving the arterial system through the arterioles, until the lowest level is reached, which marks the diastolic pressure. A pressure curve can be recorded by placing a needle or tube in the aorta or an artery and connecting it to a recording manometer (pressure-measuring device). This is a method of direct measurement of the systemic blood pressure. A somewhat less accurate, but acceptable, method of measuring arterial pressure is the indirect method, using a sphygmomanometer, which is a pneumatic cuff that can be inflated and tightened around the upper arm, and which is connected with a mercury manometer or a calibrated air manometer. The pressure in the cuff is raised to a level well above the expected systolic pressure and then the air in the cuff is gradually released. High pressure completely stops the flow of blood in the temporarily collapsed artery of the arm, and as the cuff deflates the external pressure falls, the point is reached at which the blood begins to flow through the artery again. This point represents the systolic pressure level and is signified by the appearance of a sound that can be heard by means of a stethoscope placed in the bend of the elbow. Further deflation of the cuff leads to a point at which the sound becomes muffled or disappears. This represents the diastolic pressure.

The sudden distention of the first portion of the aorta by the blood ejected into it sets up a wave of elasticity that is conducted along the arterial system as the pulse wave. Since the distention creating the wave is proportional to changes in pressure, the pulse wave is identical with the pressure pulse recorded from the inside of the artery. Thus the pulse can be palpated over any accessible artery or can be recorded by an apparatus called the sphygmograph. The shape of the pulse wave is related to the quantity of blood ejected into the aorta, the speed of ejection, the elasticity of the arterial system, and the peripheral resistance. It may reflect certain abnormalities produced by disease—palpation of the pulse is one of the oldest techniques of physical examination. The pulse wave travels at high speeds—much higher than the actual movement of the blood. Its transmission occurs fastest in the more rigid, smaller arteries (about 30 m per sec), and more slowly in the large, more elastic arteries (3 to 10 m per sec). The movement of blood occurs much more slowly. For example, it takes 0.05 sec or less for the pulse wave to reach the carotid artery in the neck, but it takes 1–2 sec for the ejected blood to reach that point.

Regulation of pressure and flow

It has already been stated that the autonomic nervous system exerts an important influence upon the various functions of the circulation. It plays a paramount role in two essential regulatory functions within the systemic circulation: (1) the maintenance of arterial blood pressure, and (2) the distribution of the blood according to functions and needs of the various organs and parts of the body.

It can be surmised from Equation 4 that the systemic arteriolar resistance cannot be of a stable magnitude, for if it were, each time the cardiac output increased the blood pressure would rise in proportion; *i.e.*, a fivefold rise in cardiac output during exercise would increase the systolic pressure from 120 to 600 mm Hg. To forestall such a disaster, the arterial system possesses a sensitive self-regulatory mechanism. The arterioles are richly supplied by terminals of the autonomic nervous system—the sympathetic and parasympathetic nerves. The sympathetic nervous system has dual control over the cross-sectional size of arterioles, having both vasoconstricting (decreasing) and vasodilating (expanding) effects. The parasympathetic system has only vasodilating effects. The regulation of the size of the arterioles is a reflex phenomenon, unrelated to the conscious will of the individual. Signals from the autonomic nervous system reach the nerve terminal, where a chemical substance (transmitter) is released, changing the tone of the arteriole. Acetylcholine, a vasodilator, decreases the tone, and norepinephrine, a vasoconstrictor, increases it. The great arteries have in their walls sensing terminals, the most important of which are located in the arch of the aorta and in the major artery in the neck, the carotid artery (carotid sinus). These sensors respond to stretch: when more blood is ejected into the aorta the sensors recognize the stretch of the wall of the vessel and send the appropriate signal to the proper autonomic centre in the brain. This vasomotor centre then stimulates the vasodilating fibres in the arterioles, causing them to relax just the right amount to let out the excess blood and thus maintain the pressure at

its original level. This mechanism permits the arteriolar resistance to fall each time the output rises, and vice versa, maintaining a stable blood pressure in the arterial system.

The pressure-regulating mechanism applies to the total systemic arteriolar resistance, which, as stated, is the sum total of many subdivisions of the systemic circulation. The distribution of blood into these subdivisions represents the second essential regulating mechanism. The systemic circulation is subdivided not only topographically but also in accordance with the functional unity of certain widely spread tissues of the body. These subdivisions are connected in parallel with the central arteries; thus, the blood has the option of entering any subdivision in larger or smaller amounts—or not at all—without disrupting the systemic flow. The regulation of regional flow is related to the respective role and function of each region, which differ widely even under basal conditions, and the need for oxygen varies greatly from region to region. The main subdivisions of the arterial system are as follows: (1) The splanchnic circulation (gastrointestinal tract and abdominal organs) constitutes 24 percent of the systemic circulation; its basal oxygen utilization is roughly equivalent to the body average (expressed as an arteriovenous oxygen difference of 40 ml per litre of blood). (2) The renal circulation (kidneys) makes up 19 percent of systemic blood flow. This circuit has a low oxygen-utilization rate—the blood returning from the kidneys is almost as fully saturated with oxygen as when entering them. This overabundancy of flow above the obvious needs is caused by the fact that the blood is filtered by the kidneys of its wastes—whence the urine is formed. The kidneys are supplied with working blood over and above their nutrient blood. (3) The cerebral circulation (brain) is 13 percent of systemic flow. The brain demands and extracts more oxygen than the body average. (4) The coronary circulation (heart) comprises 4 percent of systemic flow. This circuit is the highest user of oxygen in the body, with the blood returning from the heart tissues very low in oxygen, about 30 percent (equivalent to an arteriovenous oxygen difference of 110 ml per litre of blood). The coronary circulation thus has a low reserve and it is extremely vulnerable, so that its adjustment to augmented needs is of crucial importance. (5) Circulation to the skeletal muscles contains 21 percent of systemic flow. At rest the oxygen utilization is higher than average (a work-saving device for this largest part of the body). (6) Circulation to the skin includes 9 percent of systemic flow. The skin, as is the case with the kidneys, has more abundant blood flow than it needs. This excess is related to one of its functions: blood perfuses the skin not only to nourish it but to help in regulating the body temperature. (7) Circulation to other organs in the body is 10 percent of systemic flow, and it utilizes about the average amount of oxygen.

Tissue blood supply

The above figures show the wide difference between the various regions in their respective blood supplies: some have an overabundance, others a scanty supply. The flow of blood into each region is, of course, determined by the tone of the arterioles—those more widely open and offering less resistance have higher flows; those with a higher tone have lower flows. The tone of the arterioles is capable of change in each region, independently from other regions, causing moment-to-moment shifts in blood perfusion of regions and organs. While all arterioles are under the influence of the autonomic nervous system and respond to signals from the vasomotor centre in participating in the general pressure-regulating mechanism, the differential nature of the flow is aided by the differences in the number of vasoconstrictor and vasodilator nerve fibres. In addition there are other mechanisms that produce local arteriolar dilation without the participation of the centres in the brain. Among these are local nervous reflexes (axon reflexes) that may dilate arterioles in a region. Furthermore, chemical substances play an important part in altering local resistance: these include many hormones such as vasopressin (secreted by the posterior pituitary gland), angiotensin (produced by the kidneys), histamine or the related H-substance (found in the skin), bradykinin (found in the salivary gland). In addition, other hormones referred to as unidentified vasodilator substances play a part in the regional variation of blood flow. This process is referred to as the metabolic autoregulation of blood flow and includes the dilatation of blood vessels by metabolic products of tissues in their close vicinity. Finally the essential blood gases—oxygen and carbon dioxide—play some regulatory part in vascular resistance: lack of oxygen or overabundance of carbon dioxide may influence the size of the blood vessels, generally by altering the blood pressure, or locally by redirecting flows.

The most drastic temporary adjustment of the regional blood flow occurs during exercise. It has already been pointed out that during exercise the cardiac output and oxygen utilization are greatly increased. In order to meet the extremely high demands of the working muscles for oxygen, the body also shifts all the blood it can spare to the working muscle. Thus the vital organs—the heart and brain—remain well supplied, but the splanchnic organs, the nonworking muscles, the kidneys, and the skin receive only a fraction of their usual blood supply for the duration of the exercise "emergency." A combination of chemical stimuli and nervous reflexes accomplishes this remarkable feat of body economy.

The capillary system. The capillaries are microscopically thin tubes present in every tissue of the body, forming such a dense network that virtually every cell in the body comes in close contact with the blood. The diameter of most capillaries is between 0.007 and 0.015 mm, but the total network of the capillaries has been estimated to be some 60,000 miles in length. The capillaries are the working vessels around which all the metabolic processes take place. The capillary is a thin tube, in most instances a single layer of cells surrounding a space for blood. Outside the capillaries is tissue fluid—interstitial fluid—in which these blood vessels are bathed.

The blood enters the capillaries from the arterioles, usually through intermediate-sized vessels between the two, called meta-arterioles. Blood flow through the capillary system differs from that through the arterial system in that all arteries, even the smaller meta-arterioles, always contain blood, but capillaries most of the time are empty and closed. At the junction between a capillary and a meta-arteriole there is a ringlike band of smooth muscle, the precapillary sphincter, which regulates the flow, opening some capillaries and closing others. The capillary blood flow occurs in spurts, the blood taking alternate pathways through the various capillaries in a given part of the tissue. Blood flow through the meta-arterioles and the capillaries is almost entirely regulated locally, with the most powerful regulating factor being the concentration of oxygen in the given tissue. In addition to the "true" capillaries, which represent nutritional channels supplying the tissues with oxygen and other needed substances, some organs have nonfunctioning capillaries that are merely a thoroughfare between the arterial and venous systems. These are called arteriovenous capillaries. They are particularly numerous in the skin, where they play an important role in the body's heat regulation. Even though the flow through the capillaries is intermittent, the number of capillaries is so numerous that a free metabolic exchange takes place almost continuously.

In the capillaries the circulation of blood reaches its destination and is able to fulfill its principal function—to supply all the cells with the material they need and to remove the waste products from the cells. This function of the blood is possible because the capillaries are in close proximity with every cell in the body, separated from them only by the thin capillary wall and the fluid in which the body cells float. The exchange occurs on the basis of two processes: (1) a chemical reaction that facilitates the exchange of oxygen (and to some extent carbon dioxide) and that depends on the special properties of hemoglobin (in the red blood cells), and (2) the physical process of diffusion. Both the chemical and physical processes lead to equilibration. An exchange occurs be-

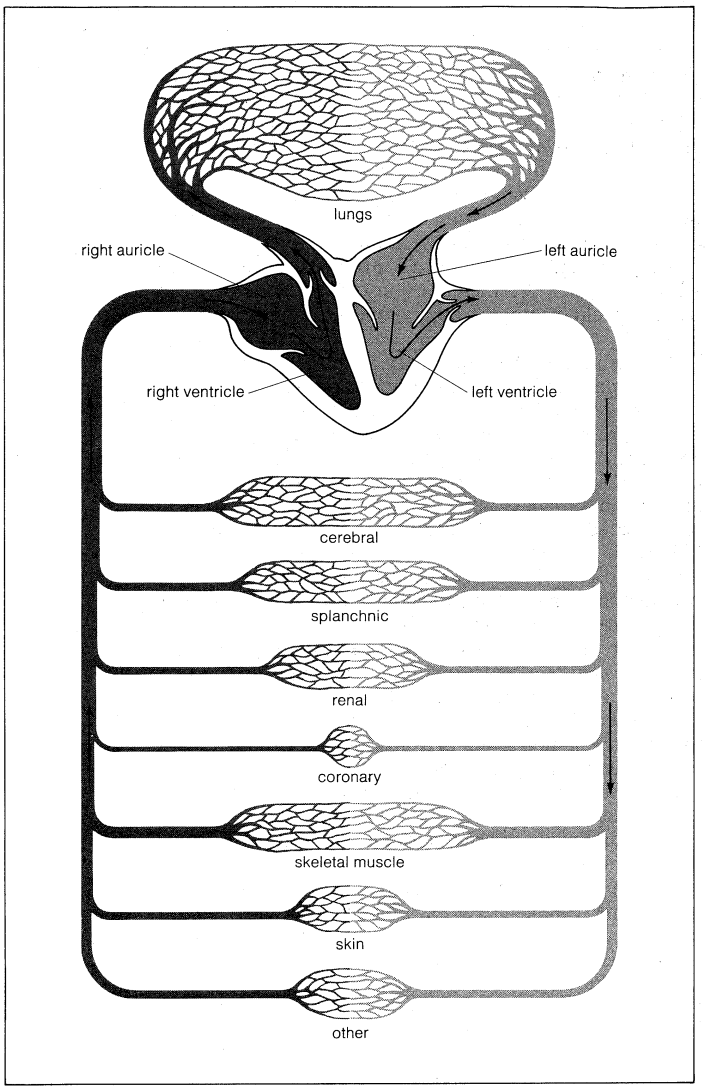


Figure 3: The systemic and pulmonary circulations, showing the principal subdivisions of the systemic circulation. Arterial blood is shown at the right, venous blood at the left.

tween the two media if the concentration of one substance is higher in one medium than in the other, and it continues until both concentrations are equal. The porous walls of the capillaries act as a permeable membrane, with selective permeability for different substances. The highest permeability (other than that for the respiratory gases) is that for water. Next come simple chemical substances of low molecular weight such as electrolytes (*e.g.*, sodium, potassium, and chloride ions) and urea. Carbohydrates and related substances of higher molecular weight have a lower index of permeability. Proteins and other giant molecular agents have a still lower index. Blood cells themselves are not exchanged in the ordinary capillaries, but there are special areas (manufacturing or hemopoietic units) that add new cells to the blood and "graveyards" that remove and destroy old cells. Thus, substances with high permeability have time to move between the two media until complete equilibrium is reached, while those of low permeability can accomplish partial exchange, so that the composition of the tissue fluid and the blood plasma never becomes identical.

Water exchange in the blood and tissues

Active exchange of water between plasma and tissue fluid occurs on a basis other than diffusion. It is driven into and out of the blood by pressure. There are four forces accountable for movement of water between the two media: (1) capillary pressure, which drives water out of the plasma toward the tissues; (2) tissue-fluid pressure, which drives water from tissue fluid to the inside of the capillary, thus opposing (1); (3) osmotic pressure of the plasma colloids, which tends to retain water in the plasma; and (4) osmotic pressure of the tissue fluid, a force similar to (3), which keeps water in the tissue spaces. Under normal conditions capillary pressure is higher in the arteriolar end of capillaries than in the venous end. As stated earlier, high pressure in the arterial system is dissipated by the resistance in arterioles, so that blood enters the capillaries under a driving pressure of about 30 mm Hg. Capillaries, as narrow vessels, offer resistance to flow and further reduce pressure to about 10 mm Hg at their venous ends. It is difficult to measure capillary pressure directly, but best estimates are that the mean, net capillary pressure is about 17 mm Hg. Capillary driving pressure (30–17 mm Hg) overcomes the opposing forces in the arterial half of the capillaries so that water is driven out of the plasma into the tissue fluid. In the venous half of the capillaries, lower pressure (17–10 mm Hg), combined with other forces, is relatively negative, and the water is returned to the plasma. Thus there is a constant exchange of water between blood plasma and extracellular fluid, with the water leaving the capillary at its arteriolar end and returning to the capillary at its venous end.

The magnitude of this capillary exchange of water can be appreciated by noting that 3 litres of blood plasma, coming in contact with 15 litres of extracellular fluid, exchange the blood volume some 45 times every minute. This large exchange makes it clear that the amount of water leaving the vascular system and that returning to it must be in nearly perfect balance—even the smallest imbalance would promptly drive plasma out of the blood or bring excess water into the circulation. This exchange of water is an important mechanism in the body's fluid retention. Should the plasma become too concentrated, as after excessive sweating, the capillary water balance is shifted to replace immediately the exact volume of water lost from the plasma, but no more. On the other hand, should excessive water imbibition dilute the plasma, the surplus of water will be removed by the capillary-exchange process until the kidneys are able to dispose of the excess.

The venous system. It has been stated earlier that the systemic venous system contains about 75 percent of the volume of blood within the systemic circulation. This inequality is related to the very slow speed of the circulation in this system, since the velocity of blood flow in any part of the circulatory system is directly proportional to the volume of blood in this part and inversely to cardiac output. The veins collect the blood from the capillary system. The smallest vessels continuous with the distal ends of the capillaries are the venules. Venules converge into small veins, which join into progressively larger ones until all the blood is draining into the caval system of two large veins returning to the heart. These two veins are the superior vena cava, which returns blood from the part of the body above the heart, and the inferior vena cava, which returns the blood from most parts of the body below the level of the heart. Both venae cavae enter the right atrium at its posterior upper and posterior lower areas, respectively.

The driving pressure returning the venous blood to the heart is very low because the arterial pressure, as it has been explained, is almost entirely dissipated by transport through the arterioles and capillaries. Thus blood enters the venous system with only a few millimetres of pressure above that in the right atrium. This low driving pressure may or may not suffice to move the blood upward toward the heart in an individual lying down, but is totally inadequate to do so in the erect human unless aided by other mechanisms, because then the blood must oppose the force of gravity. All veins located below the level of the right atrium thus have the weight of a column of blood added to their intrinsic intravascular pressure; those above the level of the heart have gravity aiding their drainage. Hydrostatic pressure in an average adult at the level of the toes is very high, about 90 mm Hg. The problem of moving the blood uphill to the heart is solved

by several pumping mechanisms. The flow toward the heart is facilitated, first of all, by venous valves. These are pocket-like structures attached to the inner walls of the veins at short distances from each other; they open to permit blood to flow toward the heart but, when the column of blood stops moving, close tightly, preventing backflow. Furthermore, the large veins in the legs are surrounded by muscle tissue. Each contraction of the muscles with the slightest motion, even the tone of the muscles when standing, exerts a milking action, moving the blood upward. This mechanism is called the muscle pump. When the veins reach the abdominal cavity and enter the inferior vena cava, valves are no longer present but another mechanism becomes operative. The abdominal cavity contains many organs whose specific gravity approximates that of blood. Pressure in the abdominal cavity is thus dissipated, as if the great veins were suspended in a liquid medium. The total pressure in the abdomen is influenced by the motion of the diaphragm, the principal respiratory muscle, which maintains the pressure at about zero, or slightly below the atmospheric pressure. To complete this mechanism, the negative pressure inside the chest (thoracic) cavity causes the blood to be sucked from the low-pressure abdominal cavity, up the vena cava, and into the right atrium. The intrathoracic negative pressure increases (*i.e.*, becomes more negative) during inspiration, accelerating the return of blood to the heart, and decreases during expiration, decreasing the return of blood. This combined action of the thorax and abdomen in enhancing the venous return of blood is called the abdominothoracic pump. The return of blood from the part of the body located above the level of the right atrium, as stated, is aided by forces of gravity. Most of the veins that are located above the level of the heart are partly closed. These veins open only enough to let the blood flow through. The principal reservoir function of the venous system is located at a level below the heart.

Venous return

The venous system, as the largest reservoir in the body, has an additional function that only recently has been appreciated. The veins have a tone—a state of partial constriction—that permits them to contract and dilate, thereby regulating the venous return to the heart, and, by the previously described mechanism, the cardiac output. This mechanism plays an important role in conditions in which cardiac output has to increase rapidly; for example, during the very initiation of exercise before the working muscles pump out the additional blood, resetting the circulation at a higher output level.

The venous system has been described as comparable to a river, collecting tributaries that arise from a merging of several smaller streams. There is, however, one exception to this system: the portal vein system, which involves all the veins draining the gastrointestinal tract and the spleen. The various veins from these sources drain, in the usual manner, into the single, large portal vein. This vein, however, takes a short course inside the abdominal cavity, enters the lower surface of the liver, divides again into smaller veins, which, inside the liver, divide into venules and capillaries—broader than average capillaries —termed sinusoids. All the blood from the gastrointestinal tract and the spleen is thus brought to the liver, where it gets detoxified and where certain fuel reserves are stored. After this process of filtration, the blood from the portal circulation mixes with the arterial blood entering the liver (through the hepatic artery) for the purpose of nutrition. The capillaries then converge into a system of venules, small and large veins, and eventually reach the large hepatic vein, draining into the inferior vena cava. This double circulation of the liver facilitates an important function of this organ—guarding the body against absorption of harmful substances from the gastrointestinal tract.

THE PULMONARY CIRCULATION

The pulmonary circulation involves the right side of the heart, the pulmonary artery and its branches, the pulmonary capillary system, and the pulmonary veins. Venous blood with reduced oxygen content, having returned to the right atrium from the systemic veins, is pumped into the pulmonary artery by the right ventricle. The volume of blood contained in the pulmonary blood vessels is about 12 percent of the total blood volume. The "lesser circulation," as applied to the pulmonary blood flow, refers not only to the smaller volume but also to the smaller circuit. Despite the differences in size and volume between the two circulations, the quantity of blood ejected into the pulmonary artery equals that ejected into the greater circulation; for the two circulations are in series with each other. An inequality in output of the two ventricles would promptly empty one circulation into the other.

There is still another important difference between the two circulations: the pulmonary arterial circuit is a low-pressure circulation. There are reasons why a high perfusion pressure is not needed in the lungs. First, the compact size of the pulmonary circulation is easily perfused under lower pressure; second, the fact that the circulation supplies only one paired organ obviates the complicated regulatory mechanisms of the systemic circulation.

It should be pointed out that the pulmonary circulation supplies the lungs only with working blood destined for the metabolic exchange of blood gases. The nutrient blood for the lungs comes from the systemic circulation as the bronchial circulation. Thus the lungs, as the liver, have a dual circulation. Unlike the liver, however, where the nutrient blood and the working blood jointly leave through the same system of veins, the lungs have two venous systems, the pulmonary and the bronchial, making two independent circulations except for minute communications between them.

The pulmonary arterial system. The right ventricle works under low pressure, and its filling pressure and ejection pressure are lower than in the left side of the heart. The filling pressure, however, is only slightly lower (3 mm Hg as opposed to 6–8 mm Hg on the left side), while the ejection pressure is one-sixth of the systemic pressure (20 mm Hg relative to 120 mm Hg on the left).

The main pulmonary artery (also called the pulmonary trunk) is a short vessel, about 4–5 cm long in the adult, which divides into two major branches, each supplying one lung. These are the right and the left branches, also called the right pulmonary artery and the left pulmonary artery. These branches, in turn, subdivide into principal branches for each of the five lobes of the lungs and then divide further and further until they reach the pulmonary arterioles, vessels comparable in size to the systemic arterioles. Being subjected to lower pressure, the pulmonary arteries and their branches are much thinner than correspondingly sized vessels in the systemic circulation. This difference is most marked in the arterioles, which have very little muscular tissue. The pulmonary arterioles are thus wide open, offering very little resistance to blood flow; hence the low pressure in this system. Unlike the systemic arterioles, the pulmonary small vessels have a very scanty supply of autonomic nerve endings. It was long believed that these vessels had no tone and were incapable of constricting or dilating under stimuli from the autonomic nervous system. It has now been demonstrated that such reactivity of the pulmonary arterioles does exist. Normally the important stimulus for this arteriolar constriction is a lowered content of oxygen in the inspired air, such as at high altitudes, where the constriction of the pulmonary arterioles can elevate pressure in the pulmonary arterial system to twice its normal value. In some diseases even more severe constriction may occur, leading to a condition that is known as pulmonary hypertension.

Although the neurogenic (*i.e.*, produced by the nervous system) responses of the pulmonary arterioles play a much lesser role than comparable responses on the systemic side, the pulmonary vessels are effectively regulated by the contents (*i.e.*, oxygen pressure) of the alveoli (thin, globular membranes enveloping the minute air bubbles).

Any reduction of oxygen in the alveolar air causes constriction of the pulmonary arterioles in adjacent areas and the blood is deflected to those parts of the lung where oxygen is more plentiful. This is called autoregulation of pulmonary blood flow. The importance of this adjustment of flow is that even in health certain portions of the lungs are better ventilated than others. As will be explained later, flow of blood through poorly ventilated alveoli could reduce oxygen saturation of the arterial blood and would be undesirable for body mechanics. The upper portions of the lungs are less well ventilated than the lower, hence more blood flow is directed toward the lower portions.

The pulmonary capillary circulation. The pulmonary arterioles lead to the pulmonary capillaries. Intermediate vessels and constricting sphincters, present in the systemic circulation, have not been identified in the pulmonary circulation. The number of capillaries located in the walls of the alveoli is greater than the basal needs—only one-fourth to one-fifth of them are utilized under conditions of rest. During exercise the blood flow through the lungs increases—as it must because the output of both circulations rises up to five times normal—and more capillaries open up to accommodate the increase. A wide-open capillary tree actually offers less resistance to flow than the more limited perfusion at rest, hence the pulmonary arterial pressure does not rise. Thus, although regulation of pressure in the systemic circulation requires a complicated mechanism for reflex dilatation of arterioles, this problem is solved in the pulmonary circulation by the expedient mechanism of opening a greater number of capillary channels.

The gas exchange in the lungs occurs efficiently despite the speed with which the blood flows through the pulmonary capillary. The process here is the reverse of that in the systemic capillaries; *i.e.*, oxygen enters the blood and carbon dioxide leaves it. Again, this is a chemical process and not one of physical diffusion. The blood in the pulmonary capillaries is separated from the air in the alveoli by a single layer of cells and a thin membrane through which these gases are easily transported.

The mechanism of exchange between water and extracellular fluid in the capillaries was explained earlier. In the lungs, the same forces are at work. Pressure in the capillaries, however, ranges from 10 to 5 mm Hg instead of from 30 to 10 mm Hg as in the systemic capillaries. Thus the same colloid pressure with a lower capillary pressure would tend to absorb water from tissue fluid in the lungs. Absorption is prevented by negative pressure in the air system, which counteracts the intravascular forces and maintains a water balance similar to that occurring in other organs.

The pulmonary venous system. From the pulmonary capillaries, blood, now fully saturated with oxygen, enters the pulmonary venules, which converge into small veins, then into successively larger veins in the manner of the venous system in general. The largest vessels, the pulmonary veins, enter the posterior portion of the left atrium of the heart. In contrast to the systemic circulation, which enters the heart via two great veins, there are four pulmonary veins, two from each lung. The blood enters the heart under a pressure of 5–6 mm Hg. Gravity, so important in the systemic circulation, plays no part in the pulmonary venous drainage. Blood is moved by various means: by pressure transmitted from the pulmonary artery, by the negative pressure in the thorax, and by diastolic suction exerted by the left ventricle drawing blood from its atrium.

The pulmonary circulation as a whole is influenced by the pressure in the thorax. The entire circuit is contained in a medium where there is negative pressure, periodically increasing and decreasing. During inspiration, when the negative pressure is accentuated, blood flow in the pulmonary circulation increases and accelerates, while at the same time the pulmonary pressure rises slightly. During expiration the negative pressure in the thorax decreases and the blood flow in the pulmonary circulation lessens and slows down.

THE LYMPHATIC SYSTEM

The transport function of blood, other than that for gas exchange, is aided by an auxiliary system of vessels—the lymphatic system (*q.v.*). This system is somewhat similar to the venous system. It starts with lymphatic capillaries, vessels as ubiquitous as blood capillaries, which converge into larger vessels, draining finally into the lymphatic duct, its largest vessel. The lymphatic duct runs upward along the inner surface of the spine and eventually empties into the venous blood system, entering a large tributary vein of the superior vena cava in the uppermost part of the thorax. Blood flow in the larger lymphatic vessels is aided by the same factors facilitating venous return of blood.

The principal role of the lymphatic system is the transport of substances consisting of large molecules, particularly proteins and fatty compounds. Lymphatic capillaries have large pores, and giant-molecular substances migrate easily into and out of such capillaries. The lymph—fluid in the lymphatic system—replenishes the blood plasma with proteins. Fatty substances are absorbed by the lymphatics from the gastrointestinal tract and then carried to the blood for distribution to the appropriate organs.

The lymphatic system contains a set of filters—the lymph nodes—where noxious substances are removed from the circulation.

BIBLIOGRAPHY. ARTHUR SELZER, *The Heart: Its Function in Health and Disease* (1966), a recent book written in technical language but aimed at the lay reader, dealing with the function and structure of the heart and circulation and changes induced by diseases of the heart; E.H. STARLING and LOVATT EVANS, *Principles of Human Physiology*, 14th edition by H. DAVSON and M.G. EGGLETON (1968); a classical textbook of human physiology; A.C. GUYTON, *Textbook of Medical Physiology*, 3rd ed. (1966), a modern textbook of physiology, *Circulatory Physiology: Cardiac Output and its Regulation* (1963), a comprehensive monograph summarizing most work dealing with the function of the heart as a pump; R.F. RUSHMER, *Cardiovascular Dynamics*, 2nd ed. (1961), a book for the physician, explaining the physiological basis for the various signs and symptoms of heart disease; O.L. WADE and J.M. BISHOP, *Cardiac Output and Regional Blood Flow* (1962), a monograph summarizing studies on the output of the heart; P. HARRIS and D. HEATH, *The Human Pulmonary Circulation* (1962), a comprehensive monograph dealing with the pulmonary circulation and its response in various forms of heart disease.

(Ar.S.)

Blood Diseases

The blood diseases are abnormal conditions that involve the corpuscular elements of the blood—the red blood corpuscles, the leukocytes, or white blood cells, and the platelets—and the tissues in which they are formed—the bone marrow, the lymph nodes, the spleen; *i.e.*, the "hematopoietic system." This definition, however, needs expansion and modification: In certain types of illness (*e.g.*, pneumonia, appendicitis) the number of leukocytes is increased, and qualitative changes in the leukocytes take place as well; these increases in the leukocytes, called leukocytosis, do not represent "blood disease." In other diseases, as will be mentioned below, a reduction in the number of red corpuscles in the blood (anemia) occurs; such anemia is not usually thought of as representing a "blood disease," but it does represent the response of the hematopoietic system to the underlying disease. Strictly speaking, the term blood disease refers only to those types of anemia and those disorders of leukocytes, of platelets, of coagulation, and of the bone marrow and lymph nodes and spleen in which the blood-forming organs are the primary tissues that are involved.

In the following discussion, all varieties of alterations in the three corpuscular elements of the blood will be discussed. For convenience, disorders chiefly affecting the red corpuscles, the leukocytes, the platelets, and the process of blood coagulation will be considered in turn, but, as will become apparent, alterations in disease do not necessarily occur in only one of these groups of corpus-

cles; in its reactions in disease the hematopoietic system often reacts as a unit.

Since the blood circulates throughout the body and carries nutritive substances as well as waste products, it is not surprising that by its examination the presence of disease can be detected. Examination of the blood may be considered in two categories; namely, the analysis of the fluid portion, the plasma, and the study of the corpuscles. Examination of the plasma includes measurement of plasma proteins, blood sugar, salts (which, being in solution and in the ionic state, are referred to as electrolytes), and various hormones. Such measurements are useful in the identification of diseases that are not classified as blood diseases; *e.g.*, diabetes, kidney disease, and thyroid disease.

Development of understanding of the blood and blood diseases. The blood has fascinated mankind as long as human thought has been recorded and human memory recalls. Long before the nature and composition of blood were known, a variety of complaints was attributed to disordered blood. The red blood corpuscles were not recognized until the 17th century, and it was another 100 years before one of the forms of the white corpuscles, the lymphocyte, and the clotting of blood were described. In the 19th century other forms of leukocytes were discovered and a number of diseases of the blood and blood-forming organs were distinguished.

In the 19th century and also in the first quarter of the 20th century much attention was given to descriptions of the morphologic changes—the changes in form and structure—that take place in the blood in disease and to the clinical manifestations of the various blood diseases. In the years that followed, a more physiologic approach began to develop, concerned with the mechanisms underlying the development of disease and with the ways in which abnormalities might be corrected. Following World War II, progress was greatly accelerated by the strong financial support that medical science received, and, in what will perhaps be looked upon as the "golden age of medical research," knowledge was gained at an ever-accelerating pace. Changing priorities in the 1970s may result in a serious slackening of this pace.

Methods of study. The study of a particular instance of disease involves inquiry into the circumstances of its development, the symptoms, and the course of the illness (the history). A thorough physical examination of the affected person and specific laboratory tests are essential for an intelligent approach to treatment.

Features of laboratory and physical examinations

In the case of the blood, certain features of the physical examination are especially important in diagnosis. These include noting the presence or absence of pallor or, the opposite, an excess of colour; jaundice, red tongue, enlargement of the heart and liver; the presence or absence of small purple spots or larger bruises in the skin; enlargement of lymph glands (nodes); enlargement of the spleen; and tenderness of the bones.

Laboratory studies particularly valuable in diagnosis include (1) determination of the existence of anemia or polycythemia (see below); (2) study of the white corpuscles, their number, and their proportions as to type; (3) enumeration of the blood platelets and a study of the blood-clotting process; and (4), in many instances, a study of the bone marrow. Further, it is sometimes necessary to remove a lymph node for microscopic examination, and X-ray examinations may be necessary for the detection of organ or lymph node enlargement or bone abnormalities. The more unusual cases may require further examinations; *e.g.*, special biochemical procedures. Measurement of total blood volume is rarely useful, partly because readjustments in plasma volume take place rapidly, thereby restoring total blood volume toward normal levels, and partly because methods of red cell and plasma volume measurement are not exact.

Disorders affecting red cells

The quantity of red blood cells (RBC) in normal human beings varies with age and sex as well as with external conditions, primarily atmospheric pressure. At sea level an average normal man has 5,400,000 red blood cells per cubic millimetre of blood. These carry an average of 16 grams of hemoglobin per 100 millilitres of blood. If such blood is centrifuged so that the red blood cells are packed in a special tube known as the hematocrit, they are found, on the average, to occupy 47 percent of the volume of the blood. In the average woman, the normal figures are lower than this (red cell count 4,800,000; hemoglobin 14 grams, volume of packed red cells 42 percent). In the newborn infant they are higher but decrease in the course of several weeks to levels below those of the normal woman; thereafter, they rise gradually. The differences in male and female blood begin to appear at about the time of puberty. It should be emphasized that these figures represent average values; those found in normal persons range approximately 15 percent on either side of this mean.

Normal values for red blood cells

From the physiologic standpoint, it is the quantity of hemoglobin in the blood that is important, because this iron-containing protein is required for the transport of oxygen from the lungs to the tissues. In disease, as well as in certain situations in which physiologic adjustments take place, the quantity of hemoglobin may be reduced below normal levels, anemia, or may be increased above normal, polycythemia.

THE ANEMIAS

Anemia varies in severity. The reduced amount of oxygen in the blood of anemic persons may be a cause of their symptoms and of the consequent poor functioning of organs such as the heart, the skeletal muscle, and the brain, but the tolerance of different persons for anemia varies greatly and depends in part upon the rate at which anemia has developed. When anemia has developed gradually, affected persons may endure severe grades of anemia with few or no complaints, whereas rapidly developing anemia will cause severe symptoms; if sufficiently severe and rapid in development, anemia can be fatal.

Anemia is always a sign, either predominant or incidental, of some underlying congenital condition or acquired disease. There are many varieties of anemia. Their clinical manifestations are, in the main, similar, and yet they must be differentiated because their causes differ and consequently their treatment is not the same. Differentiation is based on the history and physical examination, which may reveal an underlying cause, and on examination of the blood. The latter includes measurement of the degree of anemia and microscopic study of the blood corpuscles. If the number of red corpuscles, the hemoglobin concentration of the blood, and the volume of packed red cells are known, one can calculate the mean volume and hemoglobin content. The mean corpuscular volume (MCV) normally is 82 to 92 cubic microns, and about a third of this is hemoglobin (mean corpuscular hemoglobin concentration, MCHC, normally is 32 to 36 percent). These, if determined accurately, are useful indices of the nature of an anemia.

Under the microscope the red cells of man appear as round biconcave discs of uniform size with an average diameter of about 7.5 microns, or 0.0075 millimetre (a common pin is about one millimetre thick). Microscopic inspection of films of blood dried on glass slides and stained with aniline dyes allows observation of variations in the size and colour and other abnormalities of individual red cells and also permits examination of the white blood corpuscles and platelets.

Derangement of function in anemia. Red cells are formed within the marrow cavities of the central bones of the adult human skeleton (skull, spine, ribs, breastbone, pelvic bones). Under the microscope the marrow is seen to contain hosts of nucleated red cells (erythroblasts), as well as white blood cells of all stages of development and megakaryocytes, the source of blood platelets. The erythroblasts are present in various stages of development toward the mature, adult, nonnucleated, hemoglobin-containing red corpuscles that will be released into the circulating blood.

Normal production of red blood cells

The newly arrived red corpuscles in the circulation remain as reticulocytes, young red cells with a characteristic threadlike network, for two or three days. Each day's output of new red corpuscles survives an average of 120 days before succumbing to old age. The red corpuscles are able to withstand the vicissitudes of the circulation, thanks to energy supplied by glucose absorbed from the plasma; the glucose is metabolized by a variety of enzymes. Ultimately the red cells are broken down by specialized reticuloendothelial cells that are found throughout the body, and especially in the spleen and liver. The hemoglobin is digested into its components: iron, a red pigment with a ring-shaped structural formula (porphyrin), and a protein (globin). The iron and the protein remain within the body to be used over and over again in the formation of new hemoglobin, but the porphyrin ring opens and is changed chemically to become the yellow pigment of the blood plasma, bilirubin. This is then excreted by the liver and gives the bile its characteristic colour.

In a healthy person, red cell production (erythropoiesis) is so well adjusted to red cell destruction that the levels of red cells and hemoglobin remain constant. The circulation is a closed system from which there normally is no loss of blood, except that which occurs physiologically in menstruation. Anemia results when (1) the production of red cells and hemoglobin lags behind the normal rate of their destruction; (2) excessive destruction exceeds production; or (3) blood loss occurs. The bone marrow normally is capable of increasing production as much as sixfold to eightfold through an increased rate of development from the primitive cells. Anemia ensues when the normal fine balance between production, destruction, and physiologic loss is upset and erythropoiesis has not been accelerated to a degree sufficient to reestablish normal blood values.

The rate of production of red cells by the bone marrow normally is controlled by a physiologic feedback mechanism analogous to the thermostatic control of temperature in a room. The mechanism is triggered by a reduction of oxygen in the tissues (hypoxia) and operates through the action of a hormone, erythropoietin, in the formation of which the kidney plays an important role.

Failure of production may be caused by deficiency of certain essential materials, such as iron or folic acid or vitamin B_{12}, or may be due to other causes, such as the presence of certain types of disease—*e.g.*, infection; damage of the bone marrow by ionizing radiation or by drugs or other chemical agents; or anatomical alterations in the bone marrow, as by leukemia or tumour metastases (migration of tumour cells to the marrow from distant sites of origin). Accelerated destruction may occur for any one of a large variety of causes that will be discussed below (see *Hemolytic anemias*). Finally, blood loss may result from trauma or may be associated with a variety of diseases.

Classification of anemias by size of red cells

Persons whose anemia is due to increased destruction of red cells have excessive amounts of bilirubin in the plasma and appear to be slightly jaundiced, and the excess pigments darken the excreta. Certain laboratory tests serve to measure the degree of excessive pigment production. When the bone marrow is producing more red cells, the number of reticulocytes in the blood increases. These cells, in addition to their unique staining characteristics, are larger than fully mature red corpuscles. If their number is increased sufficiently, the MCV of the cells in the circulation is increased. The anemia is then characterized as macrocytic.

Macrocytic anemia also is produced when anemia results from impaired production of red cells; *e.g.*, when vitamin B_{12} is lacking. In other circumstances, as for example when there is a deficiency of iron, the circulating red corpuscles are smaller than normal and poorly filled with hemoglobin—this is termed hypochromic microcytic anemia. In still other forms of anemia there is no significant alteration in the size, shape, or coloration of the red cells—normocytic anemias.

Anemias may be classified according to the underlying abnormality in the basic physiologic mechanism (decreased production, increased destruction, blood loss) or on morphologic grounds (macrocytic, normocytic, or microcytic hypochromic) or according to their cause (*e.g.*, vitamin B_{12} deficiency). In practice it is by a combination of clinical, morphologic, and physiologic studies that the cause is determined. Accurate diagnosis is essential before treatment is attempted because, just as their causes differ widely, so the treatment of anemia differs from one patient to another. Indiscriminate treatment by the use of hematinics (drugs that stimulate production of red cells or hemoglobin) is wasteful and can be dangerous.

Other classifications of anemias

Megaloblastic anemias. Lack of vitamin B_{12} or folic acid leads to the production in the bone marrow of abnormal nucleated red cells known as megaloblasts. Such cells can be identified by their characteristic appearance. When such a vitamin deficiency occurs, bone marrow activity is seriously impaired and erythropoiesis becomes largely ineffective. Anemia develops, the number of reticulocytes is reduced, and even the numbers of granulocytes (leukocytes that contain granules in the cell substance outside the nucleus) and platelets are decreased. The adult red cells that are formed from megaloblasts are larger than normal, resulting in a macrocytic anemia. The impaired and ineffective erythropoiesis is associated with accelerated destruction of the red corpuscles, thereby providing the features of a hemolytic anemia.

Deficiency of vitamin B_{12} is manifested in other tissues as well as in the hematopoietic system. The tongue becomes sore and appears abnormally smooth. There is defective function of the intestine, resulting in "indigestion" and sometimes diarrhea. Most serious is degeneration of certain motor and sensory tracts of the spinal cord, because if vitamin B_{12} deficiency has been present for some time, treatment with vitamin B_{12} may not correct it. Initial complaints of numbness and tingling of fingers and toes may, without treatment, eventually progress to great instability of gait or virtual paralysis.

Sources of vitamin B_{12}; pernicious anemia

Vitamin B_{12} is a red, cobalt-containing vitamin that is found in animal but not in vegetable foods. Unlike other vitamins, it is not formed by higher plants but only by certain bacteria and molds and in the rumina of sheep and cattle, provided that traces of cobalt are present in their fodder. In other species, including man, vitamin B_{12} must be obtained passively, by eating food of animal source. Furthermore, this vitamin is not absorbed efficiently from the human intestinal tract unless a certain secretion of the stomach, the so-called intrinsic factor, is available to concentrate the vitamin on the intestinal wall.

The most common form of vitamin B_{12} deficiency is pernicious anemia, a condition first described by the English physician Thomas Addison in 1855 and usually affecting patients past middle life. In this disorder all stomach secretion fails, perhaps as the result of an allergic process consisting of the production of antibodies directed against the stomach lining. The tendency to form such antibodies may be hereditary.

The discovery of vitamin B_{12} came about because of the investigations of the American physician George H. Whipple, which were concerned with the study of the value of various foods in promoting the formation of hemoglobin in dogs made anemic by bleeding, and of American physicians George R. Minot and William P. Murphy, who tested the value of liver, the food found by Whipple to be most effective in the treatment of pernicious anemia. It is an interesting commentary on the ways of the growth of knowledge that the activity of the foods tested by Whipple was due mainly to their iron content rather than to the presence of vitamin B_{12}. Fortunately, from the standpoint of the victim of pernicious anemia, the once certain death was transformed to life and good health, at first if he consumed a half pound of liver per day, then if he received regular amounts of liver extract orally or by injection, and now if he is given, by injection, the equivalent of a millionth of a gram of vitamin B_{12} per day. In practice, the necessary amount of this vitamin can be given once a month or even once in three months. Oral treatment with vitamin B_{12} is possible but inefficient.

Investigations by William B. Castle, American physician and medical scientist, of the nature of the abnormality of the stomach in pernicious anemia, following the dramatic although rather fortuitous discovery of the value of liver therapy in pernicious anemia, and studies by other scientists profoundly influenced the study of hematology and converted it from a largely descriptive to a dynamic field based on a physiologic approach.

Other forms of vitamin B_{12} deficiency are rare. They are seen in complete vegetarians; in persons whose stomachs have been completely removed; in individuals who are heavily infested with the fish tapeworm *Diphyllobothrium latum* or have intestinal cul-de-sacs or partial obstructions, when competition by the tapeworms or by bacteria for vitamin B_{12} deprives the host; and in persons with primary intestinal diseases that affect the absorptive capacity of the small intestine (ileum). In these conditions, additional nutritional deficiencies, as of folic acid and iron, are likely to develop.

Effects of folic-acid deficiency

Blood changes similar to those occurring in vitamin B_{12} deficiency result from deficiency of folic acid. Folic acid is a vitamin found in leafy vegetables, but it is also synthesized by certain intestinal bacteria. In man, deficiency usually is the result of a highly defective diet, of chronic intestinal malabsorption as mentioned above, or of cirrhosis of the liver. In rare instances, additional dietary deficiency of vitamin C may produce clinical signs of scurvy and interfere with the conversion of dietary folic acid in the liver to its internally active form. Pregnancy greatly increases the need for this vitamin, and megaloblastic anemia may result. In some cases of long-continued accelerated production of red blood cells, as in some persons with thalassemia (see below), folic-acid deficiency may develop. This type of deficiency also has been observed in some patients receiving certain drugs, especially anticonvulsants.

Unless folic-acid deficiency is complicated by the presence of intestinal or liver disease, its treatment rarely requires more than the institution of a normal diet. In any event, the oral administration of folic acid, or its synthetic equivalent, pteroylglutamic acid, relieves the megaloblastic anemia. This effect can be demonstrated even in pernicious anemia, but in that condition the nervous system is not protected against the effects of vitamin B_{12} deficiency and serious damage to the nervous system may occur unless vitamin B_{12} is given.

In the above conditions, megaloblastic anemia develops as the result of dietary deficiency, faulty absorption of, or increased demands for, vitamin B_{12} or folic acid or both. In addition to these circumstances, selective vitamin B_{12} malabsorption may be the consequence of a hereditary defect. Deranged metabolism may play a role in some instances of megaloblastic anemia of pregnancy. Metabolic antagonism is thought to be the mechanism underlying the megaloblastic anemia associated with the therapeutic use of certain anticonvulsant drugs and some drugs employed in the treatment of leukemia and other forms of cancer.

Normocytic normochromic anemias. The term normocytic normochromic anemia is applied to those forms of anemia in which the mean size and hemoglobin content of the red corpuscles are within normal limits. Usually microscopic examination of the red cells shows them to be much like normal cells, but in many instances there is some variation from the normal. In other cases, there may be marked variations in size and shape, but these are such as to equalize one another, thus resulting in normal mean values. The normocytic anemias are a miscellaneous group, by no means as homogeneous as the megaloblastic anemias.

Anemias associated with loss of blood and chronic infections

Anemia caused by the sudden loss of blood is necessarily normocytic at first, since the cells that remain in the circulation are normal. The blood loss stimulates increased production, and the young cells that enter the blood in response are larger than their fellows. If they are present in sufficient number, the anemia temporarily becomes macrocytic (but not megaloblastic).

The causes of acute blood loss, too numerous to list, include trauma, peptic ulceration of the bowel, and ulcerative lesions of other types. Treatment includes replacement, by transfusion, of the blood lost and appropriate steps directed against the cause.

A common form of anemia is that occurring in association with various chronic infections and in a variety of chronic systemic diseases. As a rule the anemia is not severe, although the anemia associated with chronic renal insufficiency (defective functioning of the kidneys) may be extremely so. Most normocytic anemias appear to be the result of impaired production of red cells and somewhat shortened survival of red cells. The anemia of chronic disorders is characterized by abnormally low levels of iron in the plasma in the face of excessive quantities in the reticuloendothelial cells (cells whose function is ingestion and destruction of other cells and of foreign particles) of the bone marrow. Successful treatment depends on the possibility of eliminating or relieving the underlying disorder.

The mild anemias associated with deficiencies of the anterior pituitary, thyroid, adrenocortical, and testicular hormones usually are normocytic. As in the case of anemia associated with chronic infections or various systemic diseases, the symptoms usually are those of the underlying condition, although sometimes anemia may be the most prominent sign. Unless complicated by deficiencies of vitamin B_{12} or iron, these anemias are abolished by appropriate treatment with the lacking hormone.

Invasion of bone marrow by cancer cells carried by the blood stream, if sufficiently great, is accompanied by anemia, usually normocytic in type. It is thought that such anemia is due to impaired production of red cells through mechanical interference. Whether this be true or not, a characteristic sign in the peripheral blood is the appearance of many irregularities in the size and shape of the red cells and of nucleated red cells; the latter normally never leave the bone marrow.

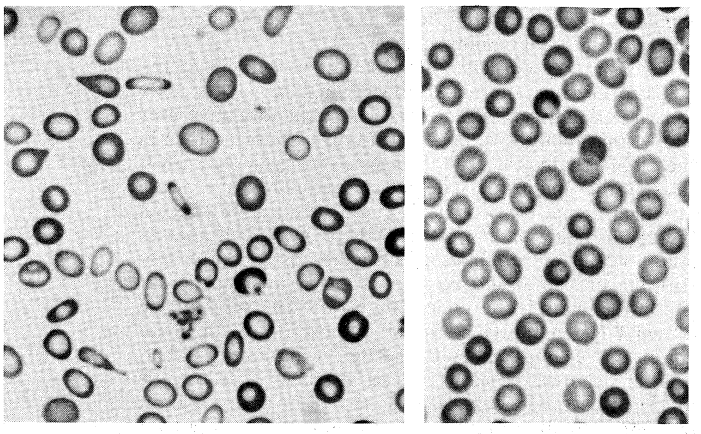
In aplastic anemia the normally red marrow becomes fatty and yellow and fails to form enough of its three cellular products—red cells, white cells, and platelets. Anemia with few or no reticulocytes, reduced levels of the types of leukocytes formed in the bone marrow (granulocytes), and reduced platelets in the blood are characteristic. The results are weakness, increased susceptibility to infections, especially of the mouth and gums, and bleeding manifestations. The last are harmless in the skin but quickly fatal in the brain. In a number of cases the onset of aplastic anemia has been found to have been preceded by exposure to such organic chemicals as benzol, insecticides, or a variety of drugs, including especially the antibiotic chloramphenicol. It is well established that sufficient exposure to ionizing radiation or to benzol or recognized myelosuppressive (bone-marrow–suppressive) agents, such as those used in the treatment of leukemia, will produce aplastic anemia, but other potential causes of such anemia seem to be harmless to hundreds or even thousands of persons. Their ill effects, therefore, would seem to depend on a peculiar sensitivity.

Aplastic anemia

Treatment consists of cessation of further exposure to any potential causative agent, administration of appropriate antibacterial drugs if infection occurs, and repeated blood transfusions in the hope of sustaining the afflicted person until spontaneous recovery of bone marrow function occurs. Sometimes the administration of the male sex hormone, with or without corticosteroid hormone, may be of some value.

Hypochromic microcytic anemias. Hypochromic microcytic anemias, characterized by the presence in the circulating blood of red cells that are smaller than normal and poorly filled with hemoglobin, fall into two main categories.

Deficiency of iron is probably the most common cause of anemia throughout the world. Iron is required for hemoglobin formation; if the supply is insufficient to produce normal quantities of hemoglobin, the bone marrow ultimately is forced to produce cells that are smaller than normal and poorly filled with hemoglobin. Iron is derived from the diet and absorbed in the intestinal tract.



Photomicrograph of a blood smear from a case of hypochromic microcytic anemia due to iron deficiency. Red corpuscles from normal blood are shown, for comparison, at right. (Wright's stain, magnified about 726 ×.)
Maxwell M. Wintrobe, *Clinical Hematology*, 6th ed.

Once in the body, it is retained and used over and over again, only minimal amounts being lost through shedding of cells from the skin and the exposed membranes and through normal menstruation. Deficiency results if the dietary supplies of iron are insufficient to meet the needs; if absorption is faulty, as in malabsorption disorders; or if blood loss is occurring. Excessive menstrual loss is a common cause of iron deficiency in women, bleeding peptic ulcer in men. Iron deficiency is common in infancy and childhood, because demands are great for the ever-expanding pool of circulating hemoglobin in the growing body, and in pregnancy when the fetus must be supplied with iron. Hookworm infestation is a common cause of iron deficiency where conditions for the worm are favourable, because the intestinal blood loss caused by the myriads of worms is great.

Sources of iron; prevalence of iron-deficiency anemia

The body of the human adult normally contains two to six grams (0.07 to 0.18 ounce) iron, of which about half is in hemoglobin. In the adult male, unless bleeding and consequent iron loss take place, there is virtually no more need for iron. Milk is a poor source of iron; meat and green leafy vegetables are good sources. Victims of iron-deficiency anemia are pale but not jaundiced. The deficiency of iron-containing enzymes in the tissues, if sufficiently great, results in a smooth tongue; brittle, flattened fingernails; and lustreless hair. Under the name of chlorosis, this type of anemia was mentioned in popular literature and depicted in paintings, especially those of the Dutch masters, until the present century. It is not necessarily less common now, but no doubt is less severe in Europe and in North America than it once was. The only treatment required is oral administration of iron salts in some palatable form, such as ferrous sulfate.

Small red cells poorly filled with hemoglobin are characteristic of a hereditary disorder of hemoglobin formation, thalassemia, that is common among people of Mediterranean stock and will be discussed below. With the exception of iron deficiency and thalassemia, hypochromic microcytic anemia is rare. It is seen in anemia responsive to the vitamin pyridoxine, where the anemia probably results from a metabolic fault in the synthesis of the heme portion of hemoglobin. Sideroblastic anemia, characterized by the presence in the bone marrow of nucleated red cells surrounded by a ring of iron granules ("ringed sideroblasts") and by a proportion of small, pale red cells in the blood, is of unknown cause and difficult to treat.

Diagnosis and causes of hemolytic anemia

Hemolytic anemias. Destruction of red corpuscles at a rate substantially greater than occurs normally, if not compensated for by accelerated production, causes a type of anemia called hemolytic. Increased red cell destruction is recognized by demonstrating increased quantities of the pigmentary products of their destruction, such as bilirubin and urobilinogen, in the blood plasma, urine, and stools, as mentioned earlier, and also by evidence of accelerated erythropoiesis, such as an increase in the number of young corpuscles (reticulocytes) in the blood. When blood destruction is extremely rapid or occurs in the blood vessels, free hemoglobin will be found in the urine (hemoglobinuria). Treatment varies with the cause of the hemolytic anemia.

There are two principal causes of hemolytic anemia: (1) inherently defective red cells and (2) an environment hostile to red cells.

Abnormalities within the red cell are usually congenital and hereditary. They include hereditary spherocytosis, sickle-cell anemia, and thalassemia, as well as a number of less common conditions. Sickle-cell anemia, thalassemia, and other diseases involving abnormality of the hemoglobin are forms of hemolytic anemia but will be discussed in a special section below.

Hereditary spherocytosis is characterized by the presence in the blood of red cells that appear small, stain densely for hemoglobin, and look nearly spherical. Such cells are mechanically fragile and readily swell up and burst in dilute salt solution. In the body they break up when deprived of free access to plasma glucose. The abnormality is aggravated by a tendency for the cells to remain longer than usual in the spleen, because of their spheroidal shape. The corpuscular defect may appear if it is inherited from either parent (*i.e.*, it is inherited as a dominant gene). The anemia varies in severity. It may be so mild as to pass for years unnoticed, but it may suddenly become severe; *e.g.*, when an incidental respiratory infection briefly suppresses the accelerated production of red cells necessary to meet the constantly increased rate of their destruction. Removal of the spleen, which always is enlarged, cures the anemia by eliminating the site of sequestration and destruction of the red cells but, of course, does not prevent future hereditary transmission of the disease.

Hereditary enzyme deficiencies of the red cell cause hemolytic anemia. The anaerobic (oxygenless) metabolism of the red cell depends on the normal function of a large number of enzymes. Hemolytic anemias resulting from "inborn errors" in the main process of glucose breakdown in the red cell usually must be inherited from both parents to become manifest. The degree of anemia varies among the different enzyme lesions and even among persons with the same enzyme deficiency. In some instances of severe hemolytic anemia, removal of the spleen has ameliorated the disease.

Abnormalities also have been discovered in the alternative process of glucose metabolism, known as the phosphogluconate oxidative or "shunt" glycolytic pathway. They include deficiencies in the enzymes glutathione reductase, glutathione, and erythrocytic glutathione peroxidase. The deficiencies of the first two enzymes are transmitted as an autosomal dominant trait (that is, it may become manifest if inherited from one parent) and an autosomal recessive trait (it needs inheritance from both parents to become manifest), respectively. Deficiency of another enzyme, glucose-6-phosphate dehydrogenase (G6PD), is rather common. The frequency of this disorder varies from 1 to 36 percent among different Caucasian population groups. It is especially frequent in Mediterranean peoples. Chronic hemolysis and anemia may be present, and in some individuals ingestion of fava beans may induce severe destruction of red cells. A different variety of G6PD deficiency has been demonstrated in 10 to 14 percent of American Negroes and is of special interest in that the defect is harmless unless the individual is exposed to certain drugs, such as certain antimalarial compounds (*e.g.*, primaquine) and sulfonamides. The full effect of the deficiency is rarely observed in females.

Still other forms of hemolytic anemias due to intrinsic red cell abnormalities are known, some of which are hereditary, and others acquired. An example of the latter is pernicious anemia (see above) resulting from vitamin B_{12} deficiency. In certain types of hemolytic anemia, the intrinsic red cell abnormality is harmful only in special circumstances; *e.g.*, the G6PD deficiency mentioned above.

Hemolytic anemia can also result as the consequence of

an environment hostile to the red cell. This type represents most of the acquired forms of hemolytic anemia as distinguished from the hereditary varieties already discussed. Certain chemical agents destroy red cells whenever sufficient amounts are given (*e.g.*, phenylhydrazine); others are harmful only to certain sensitive individuals. A number of the toxic drugs are oxidants or are transformed into oxidizing substances in the body. Injury may be accidental, as with moth ball (naphthalene) ingestion in children, or it may be the undesirable effect of a drug used therapeutically. Individual sensitivity is of several kinds. There is the susceptibility of certain Negro men to antimalarial compounds mentioned above. This is attributable to a sex-linked, inherited deficiency of the enzyme G6PD. In other instances, sensitivity is on an immunologic basis; *e.g.*, hemolytic anemia caused by penicillin, quinidine, or alpha-methyldopa (a drug used in the treatment of elevated blood pressure). The anemia develops rapidly over a few days and without transfusions may be fatal.

A long recognized type of hemolytic anemia is that associated with the transfusion of incompatible red cells. The substances alpha- and beta-isoagglutinin, which occur naturally in the blood, destroy the red cells when incompatible blood is given by transfusion. Besides the best known blood groups—A, B, and O—there are other groups, incompatibility with which may account for transfusion reactions. The Rhesus (Rh) groups and the Kell groups are examples. In hemolytic disease of the newborn (erythroblastosis fetalis) the destruction of the blood of the fetus by that of the mother may be due to Rh or to ABO incompatibility. The events that take place are, first, the passage of incompatible red cells from the fetus into the circulation of the mother through a break in the placental blood vessels; then development of antibodies in the mother; and, finally, passage of these antibodies into the fetus, with consequent anemia and jaundice.

A form of hemolytic anemia that is relatively common depends on the formation of antibodies within the patient's body against his own red cells (autoimmune hemolytic anemia). This may occur in association with the presence of certain diseases, or there may be no apparent cause. Such anemias may be severe but often can be controlled by the administration of adrenocorticosteroids and treatment of the underlying disease, if one is present. In a number of instances, splenectomy—removal of the spleen—is necessary and is usually partially or wholly effective in relieving the anemia. The effectiveness of splenectomy is attributed to the removal of the organ in which red cells, coated with antibody, are selectively trapped and destroyed.

The mechanism underlying the development of autoimmune hemolytic anemia is mystifying because the substance that causes development of the hypothetical autoantibodies—the antigen—is unknown. Trapping of the red cells by the spleen is thought to depend on the fact that red cells coated with incomplete (nonhemolytic) antibody, when brought into contact with reticuloendothelial cells, adhere, become spherical, and break down.

Other varieties of hemolytic anemia include that associated with mechanical trauma, such as that produced by the impact of red corpuscles on artificial heart valves, excessive heat, and infectious agents (*e.g.*, the organism causing malaria).

Types of normal and of abnormal hemoglobin

Thalassemia and the hemoglobinopathies. Hemoglobin is composed of a porphyrin compound heme and globin. Normal adult hemoglobin (Hb A) consists mainly (97 percent) of globin containing two pairs of chains of amino acids, of which the alpha chain consists of 141 amino acids, the beta chain 146. (A chain of amino acids is called a peptide or, alternatively, when many amino acids make up the chain, a polypeptide.) A minor fraction of normal adult hemoglobin (Hb A_2) possesses a different second set of polypeptide chains, delta chains in place of beta chains. In fetal life, a different hemoglobin is present (Hb F); this progressively decreases in amount during infancy. Like Hb A_2, fetal hemoglobin possesses a pair of the same alpha chains as in Hb A, but the second set is different (gamma chains).

In normal hemoglobin, the order in which the amino acids follow one another in the chain is always exactly the same. Studies by a number of investigators have now shown that qualitative and quantitative abnormalities in the globin chains can lead to disease. In thalassemia it is thought that the primary defect is a reduction in the rate at which alpha, beta, or delta chains are manufactured, the chains being otherwise normal. In sickle-cell anemia and in other abnormalities affecting hemoglobin, the substitution of one amino acid for another at a particular site in the chain is the underlying fault. The substitution of valyl for glutamyl in the sixth position of the beta chain, for example, results in the formation of Hb S (that of sickle-cell disease) instead of Hb A. The defect is inherited as a Mendelian recessive: If only one parent transmits the abnormality, the offspring inherits the trait but is harmed relatively little. If the trait is inherited from both parents, the serious and ultimately fatal disease sickle-cell anemia is the consequence. Since the first characterization of the nature of the abnormality in Hb S by Linus Pauling and his associates (1949) more than 100 abnormal hemoglobins have been identified, and other forms of "molecular disease" have been recognized as well. Fortunately, most abnormal hemoglobins are not sufficiently affected to alter their function, and therefore no observable illness occurs. Several of these are extremely rare.

Sickle-cell trait and sickle-cell anemia

Sickle-cell anemia occurs almost exclusively in Negroes, about 7 percent of whom, in the United States, carry the sickle-cell trait. The actual disease, sickle-cell anemia, is less common (0.3 to 2.5 percent). In this condition most of the red cells of a sample of fresh blood look normally disc-shaped—discoidal—until deprived of oxygen, when the characteristic sickle- or crescent-shaped forms with threadlike extremities appear. Re-exposure to oxygen causes immediate reversion to the discoidal form. Sickle-cell anemia is characterized by severe chronic anemia, punctuated by painful crises, the latter being due to blockage of the capillary beds in various organs by masses of sickled red cells. This gives rise to fever and episodic pains in the chest, abdomen, or joints that are difficult to distinguish from the effects of other diseases. Death from anemia, from infections, or, ultimately, from heart or kidney failure often occurs before the age of 35 to 40 years. No treatment other than transfusion has been found to be effective.

Thalassemia

Thalassemia (Greek: "sea blood") is so called because it was first discovered among peoples around the Mediterranean Sea, among whom its incidence is high. This condition, when inherited from one parent, is called thalassemia minor; it causes serious disease only when inherited from both parents (thalassemia major, Cooley's anemia). Thalassemia now is known also to be common in Thailand and elsewhere in the Far East. The red cells in this condition are unusually flat with central staining areas and for this reason have been called target cells. In the mild form of the disease, thalassemia minor, there usually is only slight or no anemia, and life expectancy is normal. Thalassemia major is characterized by severe anemia, great enlargement of the spleen, and body deformities associated with enlargement of the bone marrow. The latter presumably represents a response to the need for greatly accelerated red cell production. The skull may be so deformed as to produce a mongoloid appearance, and body growth is impaired. Many victims die in childhood. Transfusions are of only temporary value and lead to excessive iron in the tissues. Removal of the spleen is not usually beneficial; apparently the short-lived red cells are destroyed elsewhere.

The defect in thalassemia may involve the beta chains of globin (beta-thalassemia), the alpha chains (alpha-thalassemia), the delta chains (delta-thalassemia), or both delta- and beta-chain synthesis. In the last (delta–beta-thalassemia), Hb F concentrations usually are considerably elevated. Beta-thalassemia comprises over 90 percent of all thalassemias. Alpha-thalassemia inherited from both par-

ents results in intra-uterine fetal death or severe disease of the newborn child (hydrops fetalis).

In most forms of hemoglobin abnormality only a single amino acid substitution occurs, but there may be combinations of hemoglobin abnormalities, or a hemoglobin abnormality may be inherited from one parent and thalassemia from the other. Thus, sickle-thalassemia, Hb C-S disease, and Hb E-thalassemia are relatively common.

Not only may anemia result from the inheritance of a hemoglobin abnormality but in certain forms of hemoglobinopathy there may be polycythemia—overproduction of red cells—resulting from the increased oxygen affinity of the abnormal hemoglobin (*e.g.*, Hb Chesapeake); impaired reversible association of the hemoglobin with oxygen, thereby interfering with oxygen transport and resulting in cyanosis (a bluish coloration of the skin because the blood lacks sufficient oxygen) and methemoglobinemia (abnormal amounts of methemoglobin in the blood; methemoglobin is a form of hemoglobin that combines lastingly rather than temporarily with oxygen); diminished oxygen binding (Hb Kansas); and normal heme–oxygen interaction, but with formation of a hemoglobin that precipitates readily, resulting in hemolytic anemia, abnormal structures (Heinz bodies) in the red cells, and abnormal pigment in the urine.

THE POLYCYTHEMIAS

The term polycythemia signifies an increase above the normal in the number of red corpuscles in the circulating blood. This increase is usually accompanied by a corresponding increase in the quantity of hemoglobin and in the volume of packed red corpuscles. The increase may be associated with an increase in the total quantity of red blood cells in the body (absolute polycythemia) or may be only relative. The latter situation occurs when, through loss of blood plasma, the concentration of red cells becomes greater than normal in the circulating blood. Relative polycythemia may be the consequence of abnormally lowered fluid intake or of marked loss of body fluid, such as occurs in persistent vomiting, severe diarrhea, or copious sweating or when water is caused to shift from the circulation into the tissue cells.

Absolute polycythemia occurring in response to some known stimulus is usually termed erythrocytosis, in contrast to erythremia, or polycythemia vera, which is a disease of unknown cause.

Erythrocytosis develops in the presence of defective saturation of the arterial blood with oxygen. This may result from decreased atmospheric pressure, as at high altitudes or from impaired pulmonary ventilation. The sustained increase in red cells in persons who reside permanently at high altitudes is a direct result of the diminished oxygen pressure in the environment. Chronic pulmonary disease may produce chronic hypoxemia (reduced oxygen tension in the blood) and lead to erythrocytosis; *e.g.*, emphysema—abnormal distension of the lungs with air. Extreme obesity also may severely impair pulmonary ventilation and thereby cause erythrocytosis (Pickwick syndrome).

Congenital heart disorders that permit shunting of blood from its normal path through the pulmonary circuit, thereby preventing adequate aeration of the blood, can cause erythrocytosis, as can a defect in the circulating hemoglobin. The latter defect may be congenital because of an enzymatic or a hemoglobin abnormality, as mentioned above; or it may be acquired as the result of the excessive use of coal-tar derivatives, such as phenacetin, which convert hemoglobin to pigments incapable of carrying oxygen (methemoglobin, sulfhemoglobin). Lastly, erythrocytosis can develop in the presence of certain types of tumours and as the result of the action of adrenocortical secretions. Treatment of polycythemia due to any of these causes involves the correction or alleviation of the primary abnormality.

Erythremia, or polycythemia vera, is a condition of unknown causation in which the numbers of red cells, and often also the numbers of white corpuscles and platelets, are increased and the spleen usually is enlarged. Persons affected with this disease have an exceptionally ruddy complexion and may complain of headaches, dizziness, a feeling of fullness, and other symptoms. Because of the excessive quantities of red cells, the blood is usually thick, and its flow is retarded; it sometimes clots in the vessels (thrombosis) of the heart, the brain, or the extremities with extremely serious consequences. The simplest method of treatment is to remove the blood, one pint at a time, from a vein until the cellular level approaches normal and the symptoms disappear. Another useful method involves the injection of radioactive phosphorus; this material is carried to the bones, where the radioactivity acts directly upon the blood-producing elements in the bone marrow and impairs their productive capacity. The administration of the phosphorus is repeated, when necessary, at intervals of three months or longer. The possibly untoward effects of ionizing radiation must always be kept in mind when this material is used. Small doses of such drugs as busulfan have been used in place of radioactive phosphorus.

Diseases related to leukocytes (white cells)

White cells include granulocytes, lymphocytes, and monocytes. Normally, the white cell count in an adult, in round numbers, ranges from 5,000 to 10,000 cells per cubic millimetre of blood.

The granulocytes as a group make up between 50 percent and 65 percent of the white cell total. There are three subgroupings of granulocytes based on the type of dye absorbed by the granules in the cytoplasm when the cells are stained in preparation for microscopic examination. Neutrophils, which absorb neutral dyes, are the largest group, numbering 3,000 to 5,500 per cubic millimetre of blood in an adult, or about two-thirds of all the white cells. Eosinophils, which take up acidic dyes such as eosin, number from 50 to 400 in an adult, or 1–4 percent of the total. Basophils, which take up basic stains, number from 0 to 40 in an adult, or up to 1 percent.

Lymphocytes, of which there are generally from 2,000 to 3,000 in an adult (28 percent to 42 percent of all white cells), possess a round nucleus that occupies most of the cell. Monocytes, which number from 300 to 700 (4 percent to 8 percent), have a lobulated nucleus with rounding projections and are the largest of all the white cells.

CHANGES IN NUMBER OF WHITE CELLS

Variations in the number of white cells in humans occur normally from hour to hour, the highest counts being recorded in the afternoon and the lowest in the early morning. Temporary increases also normally occur during muscular exercise, menstruation, pregnancy, and childbirth, as well as in emotional states.

Abnormal changes in the count, appearance, or proportion of the various white cells are indicative of pathological conditions in the body.

Leukocytosis. The condition in which white cells are present in greater numbers than normal is termed leukocytosis. It is usually caused by an increase in the number of granulocytes (especially neutrophils), some of which may be immature (myelocytes). Most often leukocytosis is the result of the presence of an infection, usually caused by pyogenic (pus-producing) organisms such as streptococcus, staphylococcus, gonococcus, pneumococcus, or meningococcus. Leukocyte counts of 12,000 to 20,000 per cubic millimetre during infections are not unusual. As the number of cells increases, the proportion of immature cells usually rises, perhaps because the demands on the leukocyte-producing tissues in the bone marrow have increased to the point at which there is an insufficient number of mature cells for delivery into the circulation. This picture of immaturity is referred to as a "shift to the left." As the infection subsides, the number of younger forms and the total white cell count decrease and ultimately return to normal. During the period of repair following an inflammatory reaction, the monocytes may increase in number, and subsequently the lymphocytes will become more numerous.

Certain types of infection are characterized from the

Lymphocytosis; infectious mononucleosis

beginning by an increase in the number of small lymphocytes unaccompanied by increases in monocytes or granulocytes. Such lymphocytosis is seen in whooping cough and in infectious lymphocytosis, a rare disorder probably of viral origin. Moderate degrees of lymphocytosis are encountered in certain chronic infections such as tuberculosis and brucellosis.

Infectious mononucleosis is a unique form of lymphocytosis in that the lymphocytes are larger than normal and often contain vacuoles—empty spaces—in their cytoplasm. Leukocyte counts of 15,000 to 30,000 per cubic millimetre are common, but a deficiency in the number of leukocytes can occur. As a rule there is no accompanying anemia or alteration in platelets; rarely the latter may be decreased in number, and extremely rarely hemolytic anemia develops. Infectious mononucleosis occurs predominantly in persons from 10 to 30 years of age and probably is due to a virus. It is thought to be transmitted by oral contact with exchange of saliva. Discomfort, fever, sore throat, and grippelike symptoms, together with enlargement of lymph nodes and spleen, characterize the condition. The blood serum contains an antibody (sheep cell or heterophil agglutinin) that is characteristic of the disease. The symptoms of this disease vary in severity in different persons, but often they are mild. Recovery takes place, as a rule, within several weeks.

Monocytosis, an increase in the number of monocytes in the blood, occurs in association with certain infectious processes, especially subacute bacterial endocarditis—inflammation of the lining of the heart—and malaria.

Eosinophilia, an increase in the number of eosinophilic leukocytes, is encountered in many allergic reactions and parasitic infections. It is especially characteristic of infestation by trichina larvae, which are ingested when infected, poorly cooked pork or pork products are eaten.

Leukopenia. The term leukopenia refers to leukocyte counts that are abnormally low (below 5,000 per cubic millimetre). Like leukocytosis, which is usually due to an increase of neutrophils (neutrophilia), leukopenia usually is due to a reduction in the number of neutrophils (neutropenia). Of itself, neutropenia causes no symptoms, but persons with neutropenia of any cause may suffer from frequent and severe bacterial infections. The term agranulocytosis refers to an acute disorder characterized by severe sore throat, fever, and marked prostration associated with extreme reduction in the number of neutrophilic granulocytes or even their complete disappearance from the blood.

Formerly, agranulocytosis often ended in death. It was later recognized that the disorder was due to sensitivity to a coal-tar product, aminopyrine (Pyramidon). It is now known that a number of drugs used therapeutically may cause neutropenia in sensitive persons, and, if this is not discovered and exposure to the offending agent is not stopped, sore throat, fever, and other signs of infection may develop and death can ensue. The severity of the infection is due to the lowered resistance that results from absence of granulocytes. Drugs that cause neutropenia include pain relievers, antihistamines, tranquillizers, anticonvulsants, antimicrobial agents, sulfonamide derivatives, and antithyroid drugs. In persons with unusual sensitivity, agranulocytosis may result from a single dose of a drug.

Neutropenia also is associated with certain types of infections (*e.g.*, typhoid, brucellosis, measles) and with certain diseases involving the bone marrow (*e.g.*, aplastic anemia) or the spleen. In addition, sufficiently high doses of radiation will cause neutropenia, as will certain antitumour agents.

Treatment includes removal of the cause of the neutropenia and the use of antibiotics appropriate for the existing infection.

LEUKEMIAS

Nature, types, and cause of leukemia. The term leukemia means "white blood" and arose from the discovery of extremely large numbers of white blood cells in the blood of certain persons; counts as high as 500,000 per cubic millimetre and even 1,000,000 per cubic millimetre may be found in some instances.

Leukemia is a fatal disease of the blood-forming organs that is encountered at all ages and in both sexes. The frequency of occurrence of this disease throughout the world has increased substantially during the 20th century. There are two main varieties of leukemia, myelogenous, or granulocytic, and lymphatic. These terms refer to the types of cells that are involved. Each of these types is further subdivided into acute and chronic categories, and additional, less common varieties are recognized, as will be mentioned below. The excessive cellular accumulation observed in all forms of leukemia is, in the acute leukemias and in chronic myelogenous leukemia, no doubt due to an overall increase in the rate of cell proliferation. This may not be true of chronic lymphatic leukemia. In that condition the primary defect may be a decreased rate of destruction of small lymphocytes.

The development of spontaneous or induced leukemia in mice as well as in certain other animals has been associated with a filterable virus. That such a virus is related to the disease in man has not been definitely proved. In mice, a variety of factors other than the virus determine whether the animal will or will not develop leukemia; certain strains of mice harbour a leukemia-causing agent and yet rarely develop leukemia. Thus genetic factors, as well as external factors such as ionizing irradiation, contribute to the development of leukemia. In man, ionizing irradiation is the one unequivocally proved leukemia-inducing agent. Survivors of the atomic bomb in Hiroshima and Nagasaki, the pioneering radiologists who used inadequately shielded apparatus, and certain patients receiving irradiation are known to have developed leukemia with a frequency far exceeding that of the general population. Noteworthy is the fact that almost all radiation-induced leukemia has been of the granulocytic variety.

There is suggestive evidence that certain chemicals and drugs, notably benzol and phenylbutazone, may cause leukemia. That genetic factors lead to an increased frequency of leukemia in certain selected instances is suggested by the higher concordance rate for acute leukemia observed in identical as compared with fraternal twins and the frequency of development of acute myeloblastic leukemia in children with Down's syndrome (mongolism), where there is a recognized chromosome defect. Evidence for the role of trauma, hormones (especially estrogens), infections, and psychological and other influences as factors leading to the development of leukemia is unconvincing.

Clinical manifestations. The terms acute and chronic refer to the duration of the untreated disease. Before the advent of modern chemotherapy, patients with acute leukemia would usually die within weeks or months of the first manifestations of the disease. The life-span of patients with chronic leukemia is measured in years.

Early signs of leukemia

Leukemia primarily involves the bone marrow, lymph nodes, and spleen. The lymph nodes and spleen usually are enlarged. Changes also take place in the leukocytes, red cells, and platelets, with consequent anemia and bleeding manifestations. The first symptoms may be weakness and an increased tendency to become fatigued because of anemia; or hemorrhages into the skin and nosebleed or gum bleeding due to a decrease in the number of platelets. In the acute leukemias these symptoms may be severe, the anemia may progress rapidly, and there may be fever. Chronic leukemia is more insidious in development, and the early manifestations may be overlooked until enlargement of the spleen or lymph nodes is discovered. Leukemia is recognized by examination of the blood, supplemented in many instances by examination of the bone marrow.

Acute leukemia is marked by the presence in the blood of immature cells normally not found there. In the lymphoblastic variety, most frequently seen in children, the cells are immature forms of the lymphatic series of cells. In myeloblastic leukemia, the predominant cells are the youngest recognizable precursors (myeloblasts) of the

neutrophils of the blood. In a third and the least common variety, acute monocytic leukemia, the immature cells appear to be precursors of the monocytes of the blood. Myeloblastic and monocytic leukemia occur more commonly in adults and adolescents than in young children. In general, acute leukemia occurs in young persons, but no age group is exempt.

The total white cell count usually is increased but not uncommonly is normal or lower than normal (leukopenic). In such cases, abnormal immature cells may nevertheless be seen in the blood. In all forms of acute leukemia, the typical cells are found in abundance in the bone marrow. "Aleukemic" leukemia refers to those instances of leukemia in which no abnormal cells are found in the blood; in these instances the leukemia is identified by examinations of the bone marrow.

Chronic granulocytic (myelogenous) leukemia is characterized by the appearance in the blood of large numbers of immature white cells of the granulocytic series in the stage following the myeloblast, namely, myelocytes. The spleen becomes enlarged, anemia develops, and the affected person may lose weight. The platelets may be normal or increased in number, abnormally low values being found only in the late stages of the disease or as an unintended result of therapy. The disease is most commonly encountered in persons between the ages of 20 and 45 years. With treatment, the leukocyte count falls to normal, anemia is relieved, the size of the spleen is greatly reduced, and a sense of well-being returns. When the leukocyte count rises again, treatment is re-instituted. Such cycles of treatment, remission, and beginning relapse with rise of leukocyte count can be repeated many times. Unfortunately, a stage ultimately is reached when treatment no longer is effective. The disease then often terminates in a form resembling acute leukemia ("blastic crisis"). There is considerable variation in the duration of the disease. Although in various series mean life-span has been about 3¼ years, many affected persons live in good general condition for five to ten years and sometimes longer.

Chronic lymphocytic leukemia differs in many ways from other forms of leukemia. It occurs most often in people over 50 years of age, and its course usually is rather benign. It is mainly characterized by an increase in the number of lymphocytes in the blood, often accompanied by more or less generalized enlargement of lymph nodes and the spleen. Affected persons may carry on for many years, without treatment and without any other manifestations. There may be no anemia and no loss of weight. Life-span in this disease is measured in terms of 5, 10, and even 15 years, occasionally even longer. Two events mark a change in the state of relative good health. One is the development of anemia, sometimes hemolytic in type, often accompanied by some decrease in the number of platelets. The other is impairment of immune mechanisms, resulting in great susceptibility to bacterial infections.

Treatment and chances of survival

Treatment differs according to the type of leukemia. Consequently, proper classification of the leukemia is the first step, once the diagnosis of leukemia has been made. Treatment of all types of leukemia reduces illness and, in acute leukemia, prolongs life. Cure of leukemia is not yet a realistic eventuality, but there are a few persons with acute leukemia who may have been cured; more than 100 affected persons, a small fraction of the total number affected, have shown no evidence of the disease for 5 to more than 15 years. Cure of chronic leukemia is as yet unknown, but spontaneous improvement may occur in any of the forms of leukemia.

A number of drugs are used for the treatment of leukemia and, now less frequently than before chemotherapy was available, various forms of irradiation. The therapeutic agents are all myelotoxic; *i.e.*, they injure all the cells of the bone marrow, normal cells as well as leukemic cells. Their mode of action is through direct damage to the dividing stem cell (unspecialized cell from which specialized cells develop) or by slowing or cessation of cell division. These effects may be accomplished by antimetabolites, substances that interfere with the synthesis of deoxyribonucelic acid (DNA, a constituent of the chromosomes in the cell nucleus), by blocking DNA strand duplication through the binding of drugs such as the nitrogen mustards with the base groups of DNA, by disruption of the mitotic spindle during cell division, or by interfering with the formation or functioning of ribonucleic acid (RNA, a substance that is manufactured in the cell nucleus and that plays an essential role in the production of protein and in other cell functions).

Much skill and experience are needed in steering the narrow path between maximum possible kill of the leukemic cells and tolerable injury to the normal cells of the host. In the process of treatment anemia may increase; the body defenses, through the decrease in the number of neutrophils, may be impaired, and the platelets may be greatly reduced in number. Anemia can be treated with blood transfusions, and serious reductions in platelets can be met for a time with platelet transfusions. Attempts have been made to prevent infection by isolation of patients and by sterilizing the gastrointestinal tract with appropriate antibiotics. When serious infections develop, as they often do, they are treated with antibiotics and by the introduction of leukocytes; the effectiveness of these measures is limited.

Acute lymphoblastic leukemia is more successfully treated than are other forms of acute leukemia. At least one complete remission can be brought about in almost all patients, and the average survival exceeds two years. Certain drugs are used to bring about remission; if the remission is complete, the patient becomes well, and no signs of the disease are demonstrable in the blood or bone marrow; drugs other than those used to induce remission often are more useful in maintaining the remission than the remission-inducing drugs.

Acute myeloblastic leukemia and acute monocytic leukemia are less effectively treated by available drugs than is acute lymphoblastic leukemia. Transplantation of normal bone marrow, following total irradiation of the patient to destroy all his normal bone marrow cells as well as the leukemic cells, has shown promise only when an identical twin was available as the source of marrow for transplantation.

Chronic myelocytic leukemia is treated with a drug, busulfan, in daily doses until the leukocyte count has returned to normal; by this time the patient usually feels well. Treatment then is interrupted until the leukocyte count has risen to about 50,000 cells per cubic millimetre, when treatment is resumed. This can be repeated many times, and thus the affected person is maintained in good health for years. Not infrequently the intervals between treatments are six months in duration or longer. Busulfan, however, like other antileukemic agents, can injure the bone marrow, and other adverse effects may occur. Other drugs and X-ray therapy also have been used but are somewhat less valuable than busulfan.

Chronic lymphocytic leukemia seems best untreated, as long as anemia is not present or glandular enlargement is not too troublesome. Many of the chemotherapeutic agents increase the rate of infection, a risk this disease carries already. The high leukocyte counts in themselves are not harmful. When there is severe anemia, however, or when the platelet count is very low and bleeding manifestations are severe, adrenocorticosteroid hormones are given, preferably for only short periods.

The lymphomas, malignant tumours of the blood-forming tissues, include lymphosarcoma, reticulum-cell sarcoma, and Hodgkin's disease. In these disorders there may be no striking abnormalities in the blood, even though large tumours develop in the lymph nodes, spleen, bone, and other tissues and in organs such as the stomach and liver. In general, the course of reticulum-cell sarcoma is brief, the affected person succumbing in months or a year or two. The course of lymphosarcoma may sometimes be slower than this, while the course of Hodgkin's disease varies greatly, ranging from six months in the rare case to several years in the majority of cases and many years in a few. Like chronic leukemia, these

conditions are treated by irradiation as well as by certain chemical agents.

Hodgkin's disease; multiple myeloma

Hodgkin's disease, the most common of these conditions, differs from the others in that its course sometimes resembles that of an infectious disease rather than a tumour. There may be fever as well as lymph-node enlargement; leukocytosis often occurs. The disease generally starts as a painless lump, often in the neck. Lymph nodes in many locations in the body, such as the chest cavity and the abdominal cavity, ultimately enlarge, however, as does the spleen. Anemia and weight loss ensue, and death ultimately occurs.

Another malignant disease, probably related to the above conditions, is multiple myeloma, which is characterized by a malignant overgrowth of plasma or plasma-like cells within the bone marrow. This severely painful disorder causes defects in the bone of the skull, the ribs, the spine, and the pelvis that ultimately result in fractures. As the bone marrow becomes more and more involved, anemia develops and hemorrhages occur: the number of leukocytes may be low, and abnormal myeloma or plasma cells are found in the bone marrow. This disorder is associated with a peculiar disturbance in protein metabolism. Certain blood proteins—globulins—may be found greatly increased, and the urine often contains a unique protein called the Bence-Jones protein. A type of chronic kidney disease often develops, probably as a result of the high concentration of Bence-Jones protein in the kidney tubules; this frequently is the ultimate cause of death. In some cases the condition remains quiescent for a time, but death is inevitable. Adrenocorticosteroid hormones and chemotherapeutic agents have been used in the treatment of multiple myeloma.

Diseases related to platelets

BLEEDING DISORDERS

Several different agents keep blood from flowing out of the blood vessels and into the tissues: the lining of the blood vessels, the blood platelets (thrombocytes), and certain compounds and enzymes in the blood that tend to promote the formation of a clot.

The capillaries (minute blood vessels in intimate contact with the tissues) are elastic and have smooth linings; when the capillaries are cut, the linings normally retract about the injured area, serving to prevent further blood loss. The blood platelets are very small cellular particles (not actually cells) derived from the largest cells of the bone marrow—the megakaryocytes. The platelets number normally 200,000 to 400,000 per cubic millimetre of blood. They plug small leaks in blood vessels and participate in the clotting process. Disintegrating platelets release substances that take part in a series of complex chemical reactions in the blood that result in the production of a firm blood clot.

Causes of bleeding disorders. *Vascular weakness.* In cases of vitamin C (ascorbic acid) deficiency, capillary integrity is lost, and blood oozes into the tissues; under these conditions—*i.e.*, widespread capillary injury—the normal number of platelets (which tend to plug small breaks in blood vessels) is not sufficient.

Thrombocytopenia. Reduction in the number of blood platelets, termed thrombocytopenia, may come about as the result of impaired production of platelets or of increased destruction. Thrombocytopenia associated with such blood diseases as aplastic anemia and leukemia is attributed to impaired production, as is that associated with excessive irradiation and that which follows exposure to certain chemical agents such as benzol or the chemicals used in the treatment of leukemia. In sensitive persons, drugs such as quinidine, chlorothiazides, and some sulfonamides provoke platelet antibodies and platelet destruction, with resulting thrombocytopenia. In the case of still other drugs such a mechanism has not been as clearly demonstrated. Thrombocytopenia may also temporarily accompany certain infections, such as measles, and various systemic disorders, such as systemic lupus erythematosus. The list of circumstances in which thrombocytopenia may occur is long, and the mechanism is not always understood. There is a form of thrombocytopenia that has not been associated with any underlying cause—idiopathic thrombocytopenic purpura.

Thrombocytopenia, if sufficiently severe, is accompanied by spontaneous bleeding from the capillaries; this causes either tiny purplish spots (petechiae) or larger black and blue areas (ecchymoses) to appear in the skin; bleeding occurs commonly from the nose and gums and occasionally from other sites such as the urinary tract and the intestines. When hemorrhage occurs in the brain it is usually fatal.

Faulty coagulation. Deficiencies in any of the chemical and enzymatic factors involved in coagulation result in hemorrhages following minor injuries. In some of these disorders a specific deficiency is due to a heritable defect (*e.g.*, hemophilia); in others, an acquired pathological condition may be responsible for the deficiency (*e.g.*, conditions interfering with absorption of vitamin K). Bleeding also may be the result of faulty platelet function (thrombasthenia and thrombocytopathies).

Thrombocytosis. Substantial increases in numbers of platelets, presumably due to increased production, are seen (1) in diseases in which cells normally formed in the bone marrow are produced in excess (chronic myelocytic leukemia, polycythemia vera), (2) after removal of the spleen, and, (3) rarely, in the absence of associated diseases. Of itself, thrombocytosis rarely causes any ill effects.

Varieties of purpura

Purpura. The term purpura refers to the presence of tiny, purple spots in the skin (petechiae) resulting from the escape of some red cells out of the smallest blood vessels, the capillaries, or larger hemorrhagic areas (ecchymoses). Purpura may or may not be associated with thrombocytopenia. Some of the causes of thrombocytopenia were mentioned above. A form that is relatively common and of unknown cause is called idiopathic thrombocytopenic purpura (ITP). It usually occurs in children, adolescents, and young adults. It may follow an infection and often disappears without treatment. This condition probably represents an autoimmune process. If it does not resolve spontaneously or after treatment with adrenocorticosteroid hormones, removal of the spleen will cure more than two-thirds of cases.

There are many varieties of nonthrombocytopenic purpura. These usually are the result of some vascular weakness (vascular purpura) and may be mild and of no consequence, as in the minor bruises occurring in women and those seen in the elderly. Other instances may be troublesome and even painful. They may be associated with exposure to various infectious or chemical agents and occur with various diseases. A form that may be serious is Henoch–Schönlein purpura, which has been attributed to sensitivity to certain foods, drugs, or infectious agents, but is poorly understood. In this condition, there may be serious bleeding from the bowel, abdominal pain or joint pain and swelling.

COAGULATION DISORDERS

Hemophilia

Among coagulation disorders are included a number of bleeding disorders that are related to defects in the clotting of blood. The best known of these is hemophilia, which is due to an inherited defect transmitted by the female but manifested only in the male. The abnormal bleeding, sometimes spontaneous but often the result of slight injuries, occurs as a consequence of a lack of antihemophilic globulin (factor VIII; there are thirteen numbered "factors," or substances in the blood, that participate in the complex process of blood clotting) and may threaten the life of the victim or cripple him as a result of hemorrhages into joints.

The term bleeder once was considered to be more or less synonymous with hemophilia. Now that many other causes of abnormal bleeding besides deficiency of antihemophilic globulin have been discovered, the term bleeder has no specific meaning.

A closely related disorder is PTC (plasma thromboplastin component; coagulation factor IX) deficiency, or "Christmas disease." The inheritance of this bleeding dis-

order is like that of hemophilia, but a different substance is lacking. As more knowledge has been gained of the intricate process of blood coagulation, more and more bleeding disorders have been discovered. These include PTA (plasma thromboplastin antecedent, factor XI), factor V, factor VII, factor X, and prothrombin deficiency. Many of these conditions are due to inherited defects, but some occur as the result of action by toxic agents.

Hypoprothrombinemia, a deficiency in prothrombin, coagulation factor II, occurs most commonly in cases of obstructive jaundice, in which the flow of bile into the bowel is interrupted. Bile is necessary for the absorption of vitamin K, which is needed in prothrombin formation. When biliary obstruction is prolonged, vitamin K deficiency occurs, and hypoprothrombinemia develops. Similar changes may take place when absorption of vitamin K is impaired by conditions such as chronic diarrhea. Hypoprothrombinemia also occurs in the newborn infant as hemorrhagic disease of the newborn. This form of prothrombin deficiency can be prevented by administration of vitamin K to the mother during labour. Hypoprothrombinemia also can be produced deliberately by the use of substances such as bishydroxycoumarin when it is necessary to prolong clotting time.

Fibrinogenopenia refers to a reduction in the amount of the clotting factor fibrinogen (a plasma protein) in the blood. Fibrinogenopenia is seen in rare instances as a heritable disorder, but more commonly it is found as a complication of labour following pregnancy (defibrination syndrome).

Von Willebrand's disease (pseudohemophilia, vascular hemophilia) in some areas is second in frequency only to classical hemophilia. Transmitted as an autosomal dominant trait, this disorder appears to be due to deficiency of a plasma factor that is essential for normal platelet or vascular function and is also required for normal synthesis of the antihemophilia factor, VIII. The symptoms consist mainly of bleeding through the skin and mucous membranes, as in the purpuras, but symptoms similar to those of hemophilia may also occur.

In coagulation disorders, bleeding can be stopped if adequate amounts of the deficient factor can be given. Normal blood plasma supplies these, but the various types of factor concentrates that are being made available are preferable because they permit the administration of much larger quantities of the missing substances than can be supplied by blood or plasma transfusions (see BLEEDING AND BLOOD CLOTTING).

BIBLIOGRAPHY. M.M. WINTROBE, *Clinical Hematology*, 6th ed. (1967), an authoritative text dealing with the whole field of hematology, with emphasis on the clinical, diagnostic, and therapeutic aspects; J.V. DACIE, *The Haemolytic Anemias: Congenital and Acquired*, 2nd ed. (1960–67), a classic text with detailed bibliography; I. CHANARIN, *The Megaloblastic Anemias* (1969), a comprehensive work with numerous references; D.J. WEATHERALL, *The Thalassaemia Syndromes* (1965); H. LEHMANN and R.G. HUNTSMAN, *Man's Haemoglobins* (1966); F.B. LIVINGSTONE, *Abnormal Hemoglobins in Human Populations* (1967), the world distribution of abnormal hemoglobins; A.M. MAUER, *Pediatric Hematology* (1969), a comprehensive text on blood diseases in children; W. DAMESHEK and F. GUNZ, *Leukemia*, 2nd ed. rev. (1964); O.D. RATNOFF, *Bleeding Syndromes* (1960), an authoritative reference source; R.P. BIGGS and R.C. MACFARLANE, *Human Blood Coagulation and Its Disorders*, 3rd ed. (1962), a classic text on clotting of blood; P.L. MOLLISON, *Blood Transfusion in Clinical Medicine*, 4th ed. (1967), an authoritative text on all aspects of blood transfusion.

(M.M.W.)

Blood Groups

Blood groups are classifications of blood, based on the properties of red cells as determined by antigens. Knowledge of these groups developed during the course of centuries of experience with the introduction of blood from one person or other animal into another.

GENERAL SURVEY OF BLOOD GROUPING

Historical background. William Harvey announced his observations on the circulation of the blood in 1616, and his famous monograph entitled *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* (*On the Motion of the Heart and Blood in Animals*, 1653) was published in 1628. This discovery that blood circulates around the body in a closed system was an essential prerequisite of the concept of transfusing blood from one animal to another of the same or different species. In England experiments on the transfusion of blood were pioneered in dogs in 1665 by Richard Lower. A later experiment, in which dogs were also used, is described in the diary of Samuel Pepys in the entry for November 14, 1666. In November 1667, Lower transfused the blood of a lamb into a man; this event also is recorded in Pepys' diary. Meanwhile, in France, Jean-Baptiste Denis had also been transfusing lambs' blood into human subjects. After a fatality, Denis was arrested, and the procedure of transfusing the blood of other animals into man was prohibited by an act of the Chamber of Deputies in 1668. Little advance was made in the next 150 years. In 19th-century England, interest was reawakened by the activities of James Blundell, whose humanitarian instincts had been aroused by the frequently fatal outcome of hemorrhage occurring after childbirth. He insisted that it was better to use human blood for transfusion in such cases. By 1875, 347 examples of transfusions had been recorded. During the years 1875–1900, another hiatus occurred as the result of the introduction of the treatment of shock by injecting physiologic saline solution (salt solution of the same osmotic pressure as the blood).

Early transfusions

The German physiologist Leonard Landois had recorded, in 1875, that if the red blood cells of an animal belonging to one species were mixed with serum taken from an animal of another species, clumping of the red cells usually occurred and that sometimes the red cells burst—*i.e.*, hemolyzed. (Blood plasma is the whole blood except for its formed elements—the red and white cells and the platelets; blood serum is the liquid after the formed elements and the clotting factors have been removed.) He attributed the appearance of black urine after transfusion of heterologous blood (blood from a different species) to the hemolysis of the incompatible red cells. Thus, the dangers of transfusing blood of another species to man were established scientifically. Safe transfusion as it is known today rests primarily on two more advances, the first of which was the discovery of the human ABO blood groups by the biologist Karl Landsteiner in 1900. Landsteiner labelled with the letter O blood the red cells of which he found not to be clumped by serum from other persons; cells from a second lot of blood samples were clumped by serum from other persons and were labelled A. The red cells from blood the serum of which clumped A red cells were themselves clumped by A serum and were labelled B. Thus, the red cells of group A blood, for example, have a particular substance, called an antigen, on their surfaces that reacts with the serum from people of groups B and O. The ABO blood groups and other groups discovered later, such as Kell, Diego, Lutheran, Duffy, and Kidd, are determined in this manner, and it is in this way that blood groups are defined. The second advance involved the use of substances such as sodium citrate to prevent clotting of the blood being used in transfusion. These substances are innocuous when given by infusion to man. World War I occasioned many practical advances in the techniques of blood transfusion.

Antigens and antibodies. As has been stated, blood groups are determined by the presence of antigenic substances on the surfaces of the red cells. An antigen has been defined as a substance that can, in certain circumstances, excite the production of the corresponding antibody. (An antigenic substance that can excite antibody production is also correctly termed an immunogen.) An antibody is a substance capable of reacting specifically with particular antigens. The human red cells carry many such antigens. Because the reaction between red cells and corresponding antibody usually results in clumping—agglutination—of the red cells, the antigens are often referred to as agglutinogens. Antibodies are part of the

Definitions of antigen and antibody

circulating plasma proteins known as immunoglobulins, distinguished into several classes by molecular size and weight and other biochemical properties. Most blood group antibodies are located either in the class of immunoglobulin known as IgG or that known as IgM.

It has been customary to distinguish between naturally occurring antibodies and immune antibodies. So-called naturally occurring antibodies are invariably present in individuals lacking the corresponding antigen—for example, anti-A in the plasma of people of blood group B and anti-B in the plasma of people of blood group A. The view is gaining ground, however, that these antibodies are induced by environmental exposure, because substances with A and B activity are widely distributed throughout the plant and animal kingdoms. Immune antibodies—for example, rhesus antibodies (see below)—occur only if they are invoked by exposure to the corresponding antigen. Immunization against (*i.e.*, the production of antibodies against) blood group antigens in human subjects can occur as a result of pregnancy (see below *Blood groups and disease*), therapeutic transfusion, or deliberate immunization. The combination of pregnancy and transfusion is a particularly potent stimulus. Individual blood group antigens vary in their antigenic potential; for example, some of the antigens belonging to the rhesus system are strongly immunogenic, whereas the antigens of the Kidd and Duffy blood group systems are much weaker immunogens.

The individual blood groups. In Table 1 the well-established human blood group systems are listed in chronological order. The discovery of the ABO blood group system in 1900 was the result of a deliberate search for intra-species differences in man. The application of knowledge of the ABO system in blood transfusion practice is vital, since mistakes can have fatal consequences. Twenty-seven years after the discovery, Landsteiner, with his colleague Philip Levine, discovered the MN and P blood group systems because they had the idea of immunizing another species with human blood. The discovery of the rhesus system by Landsteiner and Alexander S. Wiener, in 1940, was made because they tested with human red cells antisera developed in rabbits and guinea pigs by immunization of the animals with the red cells of the rhesus monkey, *Macaca mulatta*. A few months elapsed before it was realized that antibodies indicating antigens of the rhesus system could be found in human beings as a result of pregnancies or blood transfusions. All the other blood group systems were first described by use of antibodies found in patients. Frequently, such discoveries resulted from the search for the explanation of unexpected unfavourable reaction in a patient after a transfusion with formerly compatible blood. In such cases the antibodies in the recipient were produced against hitherto unidentified antigens in the donor's blood. In the case of the rhesus system the presence of antibody in the maternal serum directed against antigens present on the child's red cells can have most serious consequences because of antigen–antibody reactions that produce the disease entity known as hemolytic disease of the newborn, or erythroblastosis fetalis. Some of the other blood group systems —for example, the Kell and Kidd systems—were discovered because an infant was found to be suffering from hemolytic disease of the newborn when mother and child were compatible as far as the rhesus system was concerned.

Table 1: Major Human Blood Group Systems

system	date of discovery	main antigens
ABO	1900	A_1, A_2, B, H
MNSs	1927	M, N, S, s
P	1927	P_1, P_2
Rh	1940	D, C, c, E, e
Lutheran	1945	Lu^a, Lu^b
Kell	1946	K, k
Lewis	1946	Le^a, Le^b
Duffy	1950	Fy^a, Fy^b
Kidd	1951	Jk^a, Jk^b
Diego	1955	Di^a, Di^b
Yt	1956	Yt^a, Yt^b
I	1956	I, i
Xg	1962	Xg^a
Dombrock	1965	Do^a

Blood group antigens in tissues. The blood group antigens are inherited characteristics; they are present primarily on the red cells as a property of the cell membrane. The antigens of almost all the systems have been detected in fetal red cells and remain present constantly throughout life. The I system is an exception in that the I antigen is much weaker and the i antigen much stronger in the cells of infants than in those of adults.

Presence of ABO antigens in tissues

There is no doubt that the antigens of the ABO system are widely distributed throughout the tissues. These antigens have been unequivocally identified on the other formed elements circulating in the blood besides the red cells; *i.e.*, on the platelets and the white cells (both lymphocytes and polymorphonuclear leukocytes). They are present in skin, the epithelial (lining) cells of the gastrointestinal tract, the kidney, the urinary tract, and in the lining of the blood vessels. Evidence for the presence of the antigens of other blood group systems on cells other than red cells is less well substantiated, and sometimes the opinions of different workers are at variance. Among the red cells antigens, only those of the ABO system are regarded clinically as tissue antigens and therefore need to be considered in transplantation.

As they occur on the red cells the blood group substances are in an alcohol-soluble form. In the ABO and Lewis systems, the blood group specific substances also occur in tissue fluids and secretions in a water-soluble form. In the case of the Lewis system the antigens are primarily in the secretions, and their presence on the red cells is dependent on passive absorption onto the cells from the plasma.

Chemistry of the blood group substances. Knowledge of the chemistry of the ABH and Lewis blood group substances has accrued not from extraction of red cells but by analyses of the water-soluble products. (The specificity of the only ABO blood group substance secreted by persons of group O is called H.) The substances are glycoproteins (that is, substances the molecules of which have sugar and protein fractions) containing a high percentage of carbohydrate. The specificity is associated with the carbohydrate part of the molecule.

Much less is known about the chemical composition of the other blood group systems. Some success has been obtained in extracting MN substances from red cells; P_1 substance has been identified in the fluid inside hydatid cysts in sheep, and a diversity of compounds have been found that inhibit the reactions between the Rh antigens of red cells and the corresponding antibodies. (Hydatid cysts are the sacs containing tapeworm larvae in the embryonic stage.) Some substances with blood group specificity are listed in Table 2.

Table 2: Substances with Blood Group Activity

antigenic determinant	specific carbohydrate
A	*N*-acetyl-D-galactosamine
B	D-galactose
H	L-fucose
	N-acetyl-D-glucosamine
Le^a Le^b	L-fucose
M	*N*-acetylneuraminic acid
N	*N*-acetylneuraminic acid

METHODS OF BLOOD GROUPING

Identification of blood groups. *The agglutination test.* The basic technique in identification of blood groups is the agglutination test. In its simplest form, a volume of serum containing antibody is added to a thin suspension (2–5 percent) of the red cells suspended in physiological saline solution in a small tube with a narrow diameter. After incubation for an adequate time, at the appropriate temperature, the red cells will have settled to the bottom

of the tube. These sedimented red cells are examined macroscopically (with the naked eye) for agglutination, or they may be spread on a slide and looked at through a low-power microscope. An antibody that agglutinates red cells when they are suspended in saline solution is called a complete antibody. With powerful complete antibodies, such as anti-A and anti-B, agglutination reactions visible to the naked eye take place when a drop of antibody is placed on a tile together with a drop containing red cells in suspension. After stirring, the tile is rocked, and agglutination is visible in a few moments. It is always necessary in blood grouping to include a positive and negative control for each test that is done.

Incomplete antibody

An antibody that does not clump red cells when they are suspended in saline solution is called incomplete. Such antibodies block the antigenic sites of the red cells so that subsequent addition of complete antibody of the same specificity does not result in agglutination. Incomplete antibodies will agglutinate red cells carrying the appropriate antigen when the cells are suspended in media containing protein. Albumin from the blood of cattle is a substance that is frequently used for this purpose. Red cells may also be rendered specifically agglutinable by incomplete antibodies after treatment with such enzymes as trypsin, papain, ficin, or bromelin.

The Coombs test. When an incomplete antibody reacts with the red cells in saline solution, the antigenic sites become coated with antibody globulin (gamma globulin). This means that, although no visible reaction has taken place, gamma globulin is bound to the cell membrane. The presence of this antibody globulin on cells that have been thoroughly washed in saline solution can be detected by a test that is called the indirect Coombs test after its inventor, the English immunologist Robert Royston Amos Coombs. The Coombs reagent is made by immunizing rabbits with human gamma globulin. The rabbits respond by making anti-human globulin that must be purified by removal of unwanted components before use. The Coombs reagent is then added to the washed cells on a tile, and if they are coated by antibody globulin, after gentle rocking, agglutinates will form.

In certain diseases, anemia may be caused by the coating of red cells with antibody globulins. This can happen when a mother has made antibodies against the red cells of her newborn child, or if a person makes an auto-antibody against his own red cells. The presence of this antibody can be detected by the direct Coombs test, in which the patient's red cells are washed thoroughly and tested on a tile with the Coombs reagent.

Absorption, elution, and titration. If a serum contains a mixture of two antibodies, it is possible to prepare pure samples of each by a technique that is called absorption. By this technique an unwanted antibody is removed by mixing the antiserum with red cells carrying the appropriate antigen, which interacts with the antibody and binds it to the cell surface. These red cells are washed thoroughly and spun down tightly by centrifugation, all the fluid above the cells is removed, and the cells are then said to be packed. The cells are packed to avoid dilution of the antibody being prepared. The purification of the Coombs reagent, mentioned earlier, is done in the same way.

If red cells have absorbed antibody globulin onto their surfaces, the antibody can sometimes be recovered by a process known as elution. One simple way of eluting from washed red cells is to heat them at 56° C (133° F) in a small volume of saline solution. This technique is sometimes useful in the identification of antibodies.

The technique used to determine the strength of an antibody is called titration. Doubling dilutions of the antisera are made in a suitable medium, in a series of tubes. Cells carrying the appropriate antigen are added, and the agglutination reactions are read and scored for the degree of positivity. The actual titre (concentration) of the antibody is given by the dilution at which some degree of agglutination, however weak, can still be seen. This would not be a safe dilution to use for blood grouping purposes. If an antiserum can be diluted, the dilution chosen must be such that strong positive reactions occur with selected positive control cells.

Uses of inhibition tests

Inhibition tests. Inhibition tests are used to detect the presence of substances with blood group specificity in solutions. If an active substance is added to antibody, neutralization of the antibody's activity prevents agglutination when, subsequently, red cells carrying the antigen in question are added to the mixture. This technique was of prime importance in the elucidation of the chemistry of the ABO and Lewis blood group substances. Inhibition tests are also used to find the substances normally secreted by an individual, as in the saliva. Finally, the inhibition test has applications in forensic medicine as a means of identifying antigens in bloodstains.

Hemolysis. Laboratory tests in which hemolysis (destruction) of the red cells is the end point are not used frequently in blood grouping. For hemolysis to take place, a particular component of fresh serum called complement must be present. Complement must be added to the mixture of antibody and red cells. It may sometimes be desirable to look for hemolysins that destroy group A red cells in mothers whose group A children are incompatible, or in individuals, not belonging to groups A or AB, who have been immunized with tetanus toxoid that contains substances with group A specificity.

Hemolytic reactions may occur in patients given transfusions of blood that is incompatible or has already hemolyzed. The sera of such patients require special investigations for the presence of hemoglobin that has escaped from red cells destroyed within the body and for the breakdown products of other red cell constituents.

Sources of antibodies and antigens. Normal donors are used as the source of supply of the so-called naturally occurring antibodies, such as those of the ABO, P, and Lewis systems. These antibodies work best at temperatures below that of the body (37° C, or 98.6° F); in the case of cold agglutinins, such as anti-P_1, the antibody is most active at 4° C. Most antibodies used in blood grouping must be searched for in immunized donors.

Antibodies for MN typing are usually raised in rabbits —similarly for the Coombs reagent. Antibodies prepared in this way have to be absorbed free of unwanted components and carefully standardized before use.

An unlikely source of substances with specific blood group activity has been found in certain plants. Plant agglutinins are called lectins. Some useful reagents extracted from seeds are anti-H from *Ulex europaeus* (common gorse); anti-A_1, from another member of the pulse family Fabaceae (Leguminosae), *Dolichos biflorus;* and anti-N from a South American plant, *Vicia graminea.* Agglutinins have also been found in animals; for example, the fluid pressed from the land snail *Octala lactea.*

Difficulties in blood grouping. Suspensions of cells suitable for blood grouping can be prepared from blood taken with or without an anticoagulant. Serologists have individual preferences. After such infections as pneumonia, red cells may become agglutinable by almost all normal sera because of exposure of a hitherto hidden antigenic site (T), as a result of the action of bacterial enzymes. When the patient recovers, the blood also returns to normal with respect to agglutination. It is unusual for the red cells to reflect antigenicity other than that determined by the individual's genetic makeup. The presence of an acquired B antigen on the red cells has been described occasionally in diseases of the colon.

Blood group antibodies are usually stored frozen solid, and in this condition they will keep for years. Contaminated sera or any that have lost activity can give false positive or negative results. The serologist is unlikely to be misled by this because all the tests are controlled. Particularly potent antibodies are sometimes formed by the deliberate immunization of men or of women past the childbearing stage. Before they are issued for routine use, the antibodies must be standardized with a selection of red cells from different donors who have been tested for all the known antigens. The cells can be taken fresh from volunteers or from a bank of stored frozen cells. Blood samples with rare antigens may be exchanged on an international basis.

Identification of antibodies. When an antibody is found that apparently does not correspond to any that are already known, it must first be tested with a large panel of typed red cells, which should include representation from different races. (An antigen common in one population may be rare in a different racial group.) Statistical evaluation can then be made to see whether the reactions of the new antibody are independent of all the systems known at the time of testing. Finally, the pattern of inheritance of a new blood group antigen can be established by testing families whose members have been typed for all the known systems. This is the way in which the inheritance patterns of a new system are demonstrated.

USES OF BLOOD GROUPING

Tests of blood donors

Medical uses. *Transfusion.* Blood donors must be healthy. A sample of their blood is tested to ensure that the level of hemoglobin is satisfactory and that there has been no previous venereal disease or liver infection. Correct matching for the ABO system is vital. Compatible donors on the basis of their possessing A, B, or O blood are shown in Table 3.

Table 3: The ABO Groups in Transfusion

group	antigen on red cells	antibody in serum	possible donors	forbidden donors
A	A	anti-B	group A, O	group B, AB
B	B	anti-A	group B, O	group A, AB
O	neither	anti-A and anti-B	group O	group A, B, AB
AB	A and B	neither	all groups	none

Potential donors are also tested for some of the antigens of the Rh system, since it is essential to know whether they are Rh-positive or Rh-negative. This involves testing for the D (Rh_o) antigen. Rh-positive blood must never be given to Rh-negative females before or during the childbearing age. If such a woman subsequently became pregnant with an Rh-positive fetus, she might form anti-Rh antibody, even though the pregnancy was the first, and the child might suffer from hemolytic disease of the newborn.

Direct match test

Care is taken not to give a transfusion unless the cells of the donor have been tested against the recipient's serum. If this compatibility test indicates the presence of antibody in the recipient's serum for the antigens carried by the donor's cells, the blood is not suitable for transfusion because an unfavourable reaction might occur. The test for compatibility is called the direct match test. It involves testing the recipient's serum with the donor's cells, suspended in saline solution at 20° C and 37° C, in albumin at 37° C, and by the indirect Coombs test at 37° C. These are adequate screening tests for most naturally occurring and immune antibodies.

If, in spite of all the compatibility tests, a reaction does occur after the transfusion is given (the unfavourable reaction often takes the form of a fever), an even more careful search must be made for any red cell antibody that might be the cause. A reaction after transfusion is not necessarily due to red cell antigen–antibody reactions. It could be caused by the presence of antibodies to the donor's platelets or white blood cells. Transfusion reactions are a particular hazard for persons requiring multiple transfusions.

Organ transplants. The ABO antigens are widely distributed throughout the tissues of the body. Therefore, when organs, such as kidneys, are transplanted, most surgeons prefer to use organs that are matched to the recipient's with respect to the ABO antigen system, although the occasional survival of a grafted ABO-incompatible kidney has occurred. The remaining red cell antigen systems are not relevant in organ transplantation.

Blood groups and disease. The question whether the blood groups merely have nuisance value in clinical practice has led to a search for evidence of associations with disease. The possibility that people with particular antigens are more or less likely to get certain diseases has inspired a great deal of work, but convincing positive correlations are few in number. Cancer of the stomach is more common in people of group A than in those of groups O and B. Duodenal ulceration is more common in nonsecretors of ABH substances than in secretors.

Hemolytic disease of the newborn

On the other hand, hemolytic disease of the newborn is undoubtedly due to blood group incompatibility between mother and child. Only certain of the blood group systems—Rh, Kell, Kidd, Diego, and, rarely, ABO, Duffy, and MNS—have been implicated in this disease. The most common cause of hemolytic disease of the newborn is incompatibility for the Rh antigen D (Rh_o) between mother and child. If the mother is Rh-negative (does not have the Rh antigens) and the child is Rh-positive, then the mother may form anti-D, which can react with the child's cells. The antibody molecules are small enough to pass through the placental barrier and enter the fetal circulation. The antibody can then coat the fetal red cells and ultimately destroy them. Thus, the baby may be anemic at birth. This anemia can be treated by giving the baby transfusions of Rh-negative blood. Often the blood is exchanged, and the child's own Rh-positive red cells coated with antibody are removed. The supply of maternal antibody lasts for only a few weeks. The transfusion tides the child over the time while the antibody persists; once the antibody has disappeared, the child's own genetically constituted Rh-positive cells will survive normally.

It is most unusual for Rh antibody to be formed in a first pregnancy with an Rh-positive fetus unless the mother has at some time received a transfusion of Rh-positive blood. This is because the mother does not become immunized until fetal red cells carrying the D (Rh_o) antigen have entered her circulation. This does not happen to any great extent during pregnancy, but while the baby is being born significant numbers of fetal cells can enter the maternal circulation. These fetal cells are the stimulus for immunization. The effect of this does not become apparent until the mother becomes pregnant with her second Rh-positive fetus.

Hemolytic disease of the newborn is now a preventable disease. The observation that pointed the way to this medical advance was that Rh-negative mothers who gave birth to ABO-incompatible Rh-positive babies had a smaller risk of having diseased children than similar mothers with ABO-compatible ones. This protection afforded by ABO incompatibility results from the elimination of ABO-incompatible fetal cells from the maternal circulation by the corresponding anti-A or anti-B antibody. Thus, the cells do not survive long enough to immunize the mother against the Rh antigen. The prophylactic treatment for hemolytic disease consists of giving an Rh-negative mother an injection of powerful anti-D antibody soon after she has given birth. This antibody mops up the offending red cells, and the mother is prevented from becoming immunized.

Evidence with respect to paternity. Although blood group studies cannot be used to prove paternity, they can provide unequivocal evidence that a man is not the father of a particular child. Since the blood groups are inherited, a child cannot have a blood group antigen that is not present in one or both parents. For example, if the child in question belongs to group A and both the mother and the putative father are group O, the man is excluded from paternity. Table 4 shows the phenotypes (observed characters) of the offspring that can and cannot be produced in the matings on the ABO system, considering only the three alleles (alternative genes) *A*, *B*, and *O*. Furthermore, if one parent is genetically homozygous for a particular antigen—that is, has inherited the gene for it from both the grandfather and grandmother of the child—then that antigen must appear in the blood of the child. For example, on the MN system, a father whose phenotype is M and whose genotype is *MM* (in other words, a father who is of blood type M and has inherited the characteristic from both parents) must transmit an *M* allele to all his progeny. The genotypes of possible children in the various matings of the MN system are shown in Table 5.

Evidence against paternity

Table 4: Exclusions of Paternity on the ABO System

matings	possible children	impossible children
O × O	O	A, B, AB
O × A	O, A	B, AB
O × B	O, B	A, AB
O × AB	A, B	O, AB
A × A	O, A	B, AB
A × B	O, A, B, AB	
A × AB	A, B, AB	O
B × B	O, B	A, AB
B × AB	A, B, AB	O
AB × AB	A, B, AB	O

In medicolegal work it is important that the blood samples are properly identified. If the ABO (including the subdivision of A into A_1 and A_2), MNS, Rh, and at least three of the following antigens, Fy^a, K, Lu^a, and Jk^a, are tested for, the chance of proving nonpaternity is about 62 percent.

Blood groups as genetic markers

Blood groups and genetic linkage. The blood group genes act as markers of the chromosomes that carry them. The site of a particular genetic system on a chromosome is called a locus. Each locus may be the site of several alleles (alternative genes). In an ordinary cell of the body the human chromosome set consists of 46 pairs, 22 pairs of autosomes (chromosomes other than sex chromosomes), and two sex chromosomes, designated XX in females and XY in males. The loci of the blood group systems are on the autosomes, except for Xg, which is unique among the blood groups in being located on the X-chromosome. Genes carried by the X-chromosome are said to be sex-linked. Since the blood groups are inherited in a regular fashion, they can be used as genetic markers in family studies to investigate whether any two particular loci are sited on the same chromosome; *i.e.*, are linked. The genes sited at loci on the same chromosome travel together from parent to child, and, if the loci are close together, the genes will rarely be separated.

Loci that are further apart can be separated by recombination. This happens when material is exchanged between homologous chromosomes (pair of chromosomes) by crossing over during the process of formation of the sex cells. The reproductive cells contain half the number of chromosomes of the rest of the body; ova carry an X-chromosome and spermatozoa an X or a Y. The characteristic number of 46 chromosomes is restored at fertilization. In a classical pedigree linkage study, all the members of a family are examined for a test character and typed for all available genetic markers that are not confined to the antigens of the red cells. The pedigree is then inspected for evidence of the nonindependent segregation of pairs of characters. The results must be assessed statistically, and computer programs are also available for this type of analysis.

Advances in seeing and identifying chromosomes

Since 1956 it has been feasible to culture human cells and to make microscopic preparations in which the chromosomes can be seen. Furthermore, since 1970 it has been possible to identify each human chromosome individually. This means that it is possible not only to find linkage groups in man but also to assign genetic markers to the chromosome that carries them.

Genetic linkages and a location concerning blood group systems are shown in Table 6. The locus of the Lutheran blood group system is on the same chromosome as the locus that controls the secretion of ABH substances. These two loci are on the same chromosome as that for an inherited wasting disease of the muscles, myotonic dystrophy.

This knowledge can be applied to the problem of antenatal diagnosis—that is, the determination of whether or not the disease is present in the fetus. The amniotic fluid contains secretions of the fetus, and this fluid can be sampled during pregnancy, a procedure known as amniocentesis. If the family of the unborn child is available for testing, then by determining the antigens present in the amniotic fluid, there is sometimes enough evidence to predict the chances that the child will be affected with the disease.

The linkage group comprising the locus for the nail–patella syndrome, an inherited abnormality of certain bones and the nails, for the enzyme system adenylate kinase, and for the ABO blood group system, is more of academic interest, as is the linkage quartet that includes the Rh blood group system.

Table 5: The MN System, Genotypes of Possible Children

genotypes of matings	genotypes of possible children
MM × *MM*	*MM*
MM × *MN*	*MM*, *MN*
MM × *NN*	*MN*
MN × *MN*	*MM*, *MN*, *NN*
MN × *NN*	*MN*, *NN*
NN × *NN*	*NN*

Table 6: Genetic Linkages and Location Involving Blood Groups

linkages			location
myotonic dystrophy disease	nail–patella syndrome disease	phosphoglucomutase, enzyme	
secretion of ABH	adenylate kinase enzyme	elliptocytosis (oval red cells) anomaly	chromosome number 1
Lutheran blood group	ABO blood group	Rh blood group	congenital cataract disease
		6-phosphogluconate dehydrogenase	Duffy blood group

The first blood group system to be sited was the Duffy locus on chromosome 1. Since one type of congenital cataract is genetically linked to Duffy, the locus for the cataract must also be on chromosome 1.

The Xg blood group system has been used in attempts to map the genes on the sex chromosome. Two clusters of genes have been identified: one around a locus for colour blindness and the other around Xg. These two loci are not within a measurable distance of each other by the technique of pedigree analysis. Unfortunately, it appears that the Xg locus is sited at the far end of the short arm of the X-chromosome. It would be a more useful marker if it were sited in the middle of the chromosome.

Genetic pathways. In some of the blood group systems the amount of antigen produced depends on the genetic constitution. For example, the red cells of a person whose genotype is MM show more M antigen than do MN red cells. The amount of antigen produced can also be influenced by the position of the genes. Such effects within a genetic complex can be due to determinants on the same chromosome—they are then said to be in cis—or to determinants on the opposite chromosome of a chromosome pair—in trans.

Antigen production and the position of the genes

In the Rh combination cdE/cde, more E antigen is produced than in the combination cDE/cde. This can be interpreted as a suppressing effect, on the E antigen, of the presence of D in cis. An example of suppression in the trans situation is that more C antigen is detectable on the red cells from CDe/cde donors than on those of CDe/cDE people.

The cis/trans test can also be used to see whether genetic determinants are in the same or different cistrons. A cistron is a smaller unit of genetic material than the gene as defined by classical recombination in multicel-

lular organisms. It can be defined in bacteria and fungi, in which it has been shown that genetic sites within the same cistron can complement each other. This type of analysis can be applied to the study of a complex genetic locus such as Rh. Various compound antigens are produced when the determinants are aligned in cis but not when they are in trans. Examples of such compound antigens are ce, Ce, and CE. Observations on the Rh system lead to the conclusion that the Ce and Ee sites may be in the same cistron but that the D genetic site probably is not.

During the 1940s there evolved two different schools of thought about the genetics of the Rh system. Alexander Wiener maintained that the inheritance is controlled by a single complex locus, and this view was reflected in his nomenclature. The view of Sir Ronald Aylmer Fisher and Robert Russell Race was that the system is controlled by three pairs of closely linked genes, *C-c*, *D-d*, and *E-e*. The changed conception of the gene has tended to reconcile the two viewpoints.

As knowledge of the blood groups progressed, it became apparent that the pathway leading from the gene to the production of antigen on the red cells involved a series of biochemical changes. It is known that in some systems—for example, ABO and Rh—several gene loci become involved. One of the loci controls the production of a precursor substance, and the products of the other loci modify this substance sequentially. It is interesting in this context to consider that an antigenic idiosyncrasy within a family is reproduced so precisely in all of its members.

The detection of recombination (exchange of material between chromosomes) or mutation in human families is complicated by the difficulty of excluding illegitimacy. In spite of the large number of families that have been studied, there is only one example of recombination, within the Rh system, that stands up to criticism. As far as germ cell mutations are concerned, only one possible case has been described. This paucity of examples may indicate that the mutation rate for blood group genes is lower than that estimated for other human genes (of the order of one in 50,000).

Twins and chimeras. Blood groups can be used to find out whether twins, triplets, or quadruplets differ from each other. It is not possible to prove that a twin pair is identical—only to estimate the chance that this is so. The blood groups were used to show that quadruplets born in Australia in 1950 were all different and, furthermore, that four eggs were involved.

Differences between some identical twins

It is possible for twins arising from a single egg and sperm not to have the same complement of chromosomes. This can happen if an error occurs during cell division after the formation of the zygote (fertilized ovum) and a chromosome is lost from one twin but not from the other. Some recorded examples involve the loss of a sex chromosome, X or Y, from one twin, producing an XY, XO pair or an XX, XO pair. The XO sex chromosome complement is associated with a characteristic syndrome (complex of symptoms) of an incomplete female phenotype. Such twin pairs are identical for all the characters carried by the autosomes.

An individual may be constituted of a mixture of cells of different zygotic lineage. This can occur if there is an exchange of cells between a pair of nonidentical twins during embryonic life through a communal vascular channel. This phenomenon was first described in cattle, and such twins are called chimeras. Several examples are known in man. They come to light when blood group studies reveal a mixture of two populations of cells in a donor who has not received transfusions. Another way in which a dual population of cells may arise is when two sperm and one egg are involved in the formation of a single zygote. The paternal contributions to the offspring can be sorted out by studying the genetic markers, including blood groups.

Blood groups and anthropology. The different races of man have different frequencies of the various blood group antigens. There are a few phenotypes that are so peculiar to a certain population that they can be used to diagnose racial origin. These are listed in Table 7. The presence of the Di^a antigen indicates a Mongolian origin; the frequency of the antigen is about 10 percent among the Japanese and Chinese, but it has not been found in Eskimos. The Fy (a− b−) phenotype is characteristic of Negroes; Negro origin can also be recognized by the presence of the Js^a antigen, which forms part of the Kell system. The V antigen, listed in Table 7, is a compound antigen.

Table 7: Blood Group Antigens Peculiar to Different Races
(in percentage)

system	population		
	Caucasian	Mongolian	Negro
Diego			
Di^a antigen	0	36	0
Duffy			
Fy (a− b−)	0	—	68–90
Kell			
Js^a antigen	0	—	20
Rh system			
V antigen	0.5	—	27–40

When a population is being investigated, an adequate sample of the people is classified according to the phenotypes. The gene and genotype frequencies can be computed from the phenotype percentages. Among the blood group systems, most is known about the distribution of the ABO groups. This is true not only because the system has been under investigation longest (since World War I), but also because the typing reagents are easy to obtain.

Distribution of ABO groups

The frequency of the *A* gene is high in western Europe, western Asia, and among Australian Aborigines and certain American Indian tribes in North America. The *O* gene frequency is high in north and western Europe and south West Africa. Isolated populations such as the Indians of South and Central America may be entirely of group O. The maximum frequency of the *B* gene occurs in Central Asia and northern India. The *B* gene was probably absent from American Indians and Australian Aborigines before racial admixture occurred with the coming of the white man.

On the Rh system most northern and central European populations differ from each other only slightly and are characterized by a *cde* (*r*) frequency of about 40 percent. The frequencies of the most common Rh combinations of the English are given in Table 8; they are representative of all Caucasian populations. The peoples of Africa show a preponderance of the complex *cDe* (R^o), and the frequency of *cde* (*r*) is about 20 percent. In eastern Asia, *cde* (*r*) is almost absent, and, since everyone has the D (Rh_o) antigen, hemolytic disease of the newborn, due to the presence of maternal anti-D, is unknown in these populations.

Table 8: Frequency of the Most Common Rh Gene Complexes in the English Population
(in percentage)

gene complex	frequency
*CDe** (R^1)†	41
cde (*r*)	39
cDE (R^2)	14
cDe (R^0)	3
C^wDe (R^{1w})	1
cdE (*r''*)	1
Cde (*r'*)	1
CDE (R^z)	<1

*Fisher–Race notation. †Wiener notation.

The frequencies of M and N vary less on the whole throughout the world than the antigens of the ABO and Rh systems. There are some rare antigens belonging to

this system that are found only in Negroes. Another distinct feature is that about 1 percent of Negroes have neither S nor s on their red cells. Everyone else has one or both of these antigens, which belong to the MN system.

The uniqueness of the genetic characters of Africans probably results from their long isolation from the rest of mankind. The blood group frequencies in small inbred populations reflect the influences of genetic drift. In a small community an allele can be lost from the genetic pool if individuals carrying it happen to be infertile, while people with unusual fertility can promote an increase in frequency of a particular allele. It has been suggested, for example, that *B* alleles were lost by chance from American Indians and Australian Aborigines when these communities were small. There are pronounced discrepancies in blood group frequencies between the people of eastern Asia and the aboriginal people of America. On the other hand, Polynesians resemble American Indians and Eskimos more closely than any other people of a different continent. This does not prove that the ancestors of the Polynesians came from America, but it shows that aboriginal Americans have more recent ancestors in common with Polynesians than with Eastern Asiatics.

Blood groups in nonhuman primates. Nonhuman primates carry blood group antigens that can be detected with reagents used for typing human beings. The closer their evolutionary relationship to man, the greater is their similarity with respect to antigens. The red cells of the apes, with the exception of the gorilla, give reactions on the ABO system that are indistinguishable from those of human cells. Chimpanzees are most frequently group A, but group O is represented. Orangutans are usually A, with the occasional group B or AB. Gibbons can be of any group except O, and gorillas have a B-like antigen that is not identical with the human one in activity. In both Old and New World monkeys the red cells do not react with anti-A or with anti-B, but, when the secretions are examined, A and B substances can be detected. Furthermore, agglutinins, bearing a reciprocal relationship to the substance secreted, are present in the serum. The red cells of the apes have been shown to carry M-like and N-like antigens.

The ABO system and primates

As far as the Rh system is concerned, chimpanzees carry two Rh antigens D (Rh_o) and c (hr′) but not the others, whereas gibbons have only c (hr′). The red cells of monkeys do not give clear-cut reactions with human anti-Rh sera. The original Rh antibody (anti-Rh_o) was formed when rabbits and guinea pigs were immunized with the cells of rhesus monkeys, and the specificity appeared to be the same as the antibody formed in patients, which was also called anti-D (Rh_o). Later, it became apparent that the antibodies from the two different sources were not developing in response to (defining) the same antigen.

By the time this happened it was too late to change the name Rh, important in clinical medicine. It was therefore decided that the antigen defined by the nonhuman antibody should be called LW, thus honouring Landsteiner and Wiener. The gene locus for LW is separate from that for Rh, but the two systems are in competition for a precursor substance. Most people have the LW antigen, but a very few do not.

Blood groups and biology. In some situations blood group substances remain stable over hundreds of years. It has been shown that ABH substances are still present in bones and mummified tissue from ancient Egyptian tombs. The stability of the frequencies of the blood group genes is an example of balanced polymorphism (the existence of a variety of forms within a species). The selective forces favouring or disfavouring a particular allele are counterbalanced. Direct evidence of a selective value for any given allele is lacking. The genetic effect of fatal hemolytic disease due to anti-D (Rh_o) has been the elimination of heterozygotes (individuals inheriting a characteristic from one parent only), so this cannot have had a profound effect on the relative frequencies of gene complexes with and without D. Reports on the effects of the ABO group incompatibility on fertility have been conflicting. Nevertheless, the blood group genes may have effects that are not yet recognized.

The blood groups are an important means of identifying an individual. Their relevance in medicine lies in the practice of blood transfusion and the effects of incompatibility between mother and child. To date, studies of the inheritance of the blood groups have increased knowledge of genetic mechanisms in man; in the future they will probably play an important role in the mapping of the human chromosomes.

BIBLIOGRAPHY. F.H. ALLEN, JR., and L.K. DIAMOND, *Erythroblastosis Fetalis, Including Exchange Transfusion Technic* (1958), a concise monograph, with useful glossary, based on the authors' personal experience; K.E. BOORMAN and B.E. DODD, *An Introduction to Blood Group Serology*, 4th ed. (1970), a textbook of methods, with relevant theoretical explanations, commended for clarity; S.D. and L.J. LAWLER, *Human Blood Groups and Inheritance*, 3rd ed. (1971), a monograph designed for students of biology and interested lay readers; P.L. MOLLISON, *Blood Transfusion in Clinical Medicine*, 4th ed. (1967), a complete and authoritative work on blood transfusion; A.E. MOURANT, *The Distribution of the Human Blood Groups* (1954), a comprehensive account of blood groups in relation to anthropology; L.S. PENROSE, *Outline of Human Genetics* (1959), a clear, concise introduction to the subject; R.R. RACE and R. SANGER, *Blood Groups in Man*, 5th ed. (1968), the standard work on blood groups—scholarly, yet easy to read; CURT STERN, *Principles of Human Genetics*, 2nd ed. (1960), a textbook written for students of genetics concerned primarily with man; W.M. WATKINS, "Blood-Group Substances," *Science*, 152:172–181 (1966), an article that leads the reader gently along a difficult subject; A.S. WIENER, *Blood Groups and Transfusions*, 3rd ed. (1943, reprinted 1962), an important work concerned mainly with the ABO blood group system.

(S.La.)

Board and Tile Games

Board and tile games are games played with a number of pieces on a specially constructed or marked board or with marked pieces (tiles) on a tabletop or other flat surface. Primitive man appears to have invented games for divinitary purposes—through them the gods could reveal the future to those able to understand their signs. Individuals faced with alternatives would toss a marked stick into the air, allowing the throw to determine their course of action.

The historical boundary between divination and pastime is ill defined. Board games doubtless were part of the mystical equipment of sages and soothsayers, later adapted for relaxation and pleasure. As late as 1895, when the French were attacking the capital of Madagascar, the native queen and her advisors relied more on the supposed prophetic result of a game of Fanorona (see below) than on the actual performance of their army. Dice, cards, teetotums (spinning dice), and wheels have all been used in fortune telling.

A few cultures, including the Eskimo and the Australian aborigine, had no board games. The Maori of New Zealand had one board game, Mu Torere (see below), although this may have been an adaption of a game introduced by Europeans. The peoples of South America do not seem to have any indigenous board games; those played are variants of games introduced from Africa and Europe.

This article is intended for the general reader, who may have no knowledge of the games described. It discusses the history and main types of board and tile games and their principles of play. More detailed information on rules and how to play the various games can be found in the books listed in the bibliography.

BOARD GAMES

History—development of Backgammon. The oldest gaming boards known (Figure 1A) were found by the British archaeologist Sir Leonard Woolley in the royal tombs of Ur and date from about 3000 BC. Each player had seven marked pieces, and moves were controlled by the use of six pyramidal dice, two of the four corners

being tipped with inlay. Three white and three lapis lazuli dice made a set. No account of how to play the game has survived.

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

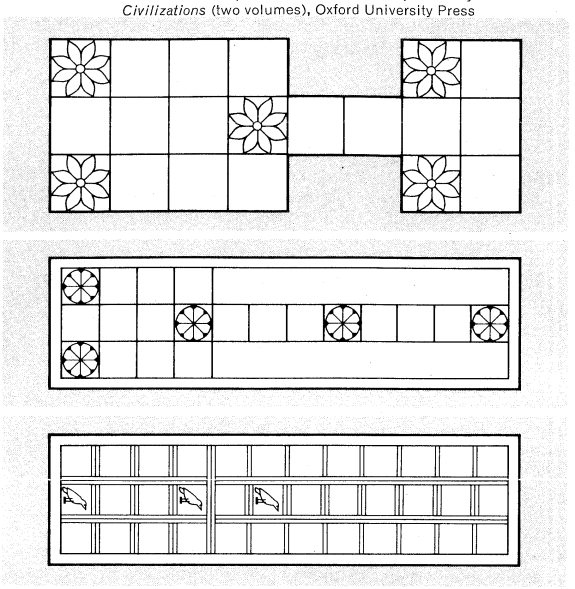


Figure 1: (Top) Game board found in the royal tombs of Ur. (Centre) Board for Egyptian game Senat. (Bottom) Game board found at Ak-hor, Egypt, for Game of Thirty Squares.

Later, the Egyptians used boards that appear to have been derived from those of Ur. Several have been found in tombs of the Empire age, about 1580 BC (Figure 1). Many of these boards were built as a box, containing dice and pieces within, and on the underside another game, known to archaeologists as the Game of Thirty Squares (Figure 1). The moves were controlled by four throwing sticks, two white and two black. Some sets contained two knuckle bones (ankle bones of sheep, marked on four sides), and a board from Ak-hor held a long die. (Long dice, with four marked faces, appear to be a natural progression from a split stick with two faces, one curved and the other flat. Cubic dice appear to be a relatively late development.)

The Game of Thirty Squares was adapted by the Romans for their Ludus Duodecim Scriptorum, with cubic dice replacing the earlier long die (Figure 2A). In the 1st century AD Ludus Duodecim Scriptorum was replaced by Tabula, a variant with only two rows of marked spaces (Figure 2B).

During the Middle Ages Tabula developed into Tables, from which Backgammon (Tric-trac) is derived (see DICE AND DICE GAMES). The development of Backgammon demonstrates how games passed from one civilization to another, undergoing modifications with the modern form bearing little resemblance to its ancient prototype.

Race games. Race games use boards with tracks marked off into spaces, some of which may be reward or penalty areas. Players race one or more pieces along the track according to throws of sticks or dice. The player whose pieces complete the course first is the winner.

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

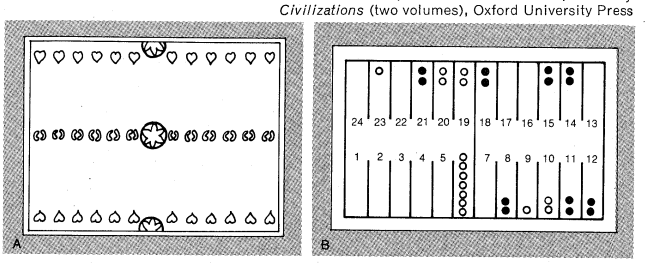


Figure 2: (A) Ludus Duodecim Scriptorum board. (B) Tabula board with game in progress.

Nyout. An early and widely disseminated group is based on a circle and cross, wholly surviving in the Korean game of Nyout (Figure 3A). Pieces, called "horses," move around the circle as directed by the throws of four sticks, which are flat on one surface and rounded on the other. If a horse lands on a cardinal point the player has the option of short circuiting home along the limbs of the cross, or continuing around the circle. If a horse lands on a square occupied by a rival's horse, the latter is sent back to start again. The game is won by the player whose horses first return to the starting point.

Circle and cross games

Pachisi. This game spread westward to India, where the shape of the board was modified by the circle being pressed inward along the sides of the cross. As Pachisi, (Figure 3B), the game became a race around the outside of the cross and up the centre of the limb nearest the player to the central square. The moves of the pieces were controlled by the throws of a long die, or by five cowrie shells, in which case the number of spaces moved depended upon the number of mouths facing upward. The Mughal emperors (16th–17th centuries) laid out courtyards in their palaces as Pachisi boards, and girls from the harem, dressed in different colours, moved about the board in response to the throws of the players gathered on a raised central dais. One of these courtyards survives at the palace in Āgra. About 1880 the game was taken to England, where it was modified and patented as Ludo. It is still a favourite with children at Christmas.

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

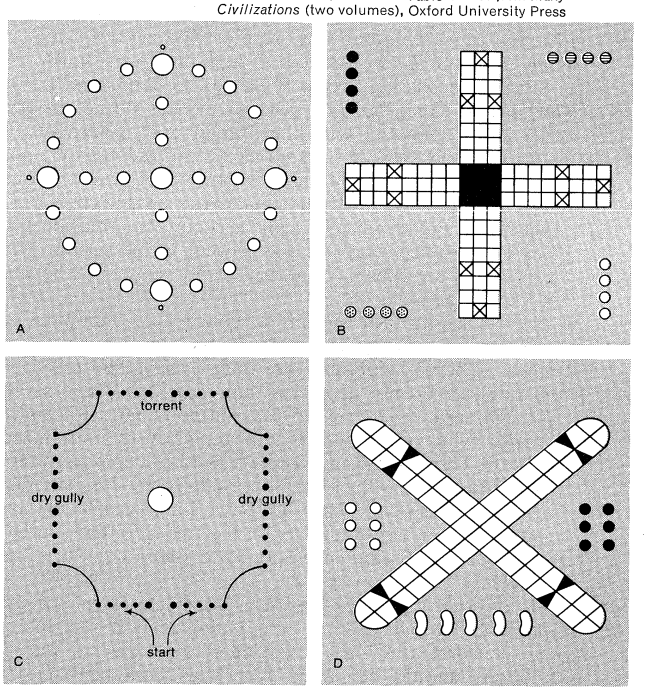


Figure 3: (A) Game board for Nyout. (B) Game board for Pachisi. (C) Zohn Ahl track. (D) Patolli cross with counters and beans.

Zohn Ahl. Circle and cross games also spread eastward into North America, and many forms survived among the Amerindians. One, Zohn Ahl, was played by the women of the Kiowa Indians in Oklahoma. The board was marked out on the ground with 40 small stones, the points being the intervals between the stones. In the centre was a flat stone, onto which dicing sticks similar to those used in Nyout were thrown. The wide gap at the north represented a river in flood, while those at the east and west were dry streams (Figure 3C). Each team had one runner and the moves of the runners were controlled by the throws of four sticks; the runners moving in opposite directions around the track. Each team started with four pebbles, which were used as counters.

A circuit of the track counted as a lap, and the winner

of a lap took a counter from the opposing team. If a runner fell into the torrent at the north or was hit by the opposing runner landing on the same space, the runner had to return to the beginning and the team forfeited a counter. A runner landing in a dry gully forfeited a turn of play. The game ended when one side held all the counters. In Zohn Ahl the circle was preserved, but the cross had disappeared except for the cardinal points, which had changed from points of advantage to penalty points.

Patolli. In the Aztec game of Patolli the circle disappeared, and the cross became a double track; the pieces were semiprecious stones, and the moves were controlled by the throws of five beans, with one side plain and the other hollowed out (Figure 3D). The Spanish Conquistadors found the nobles of Montezuma's court playing Patolli for high stakes.

Totolospi. The Hopi Indians of Arizona played Totolospi, in which the cross was degenerate, though still recognizable (Figure 4A). In another of their games only one limb of the cross had survived (Figure 4B). The moves of the stone markers were controlled by the throws of three staves, each with a flat and a curved face. The boards were scratched on slabs of sandstone.

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

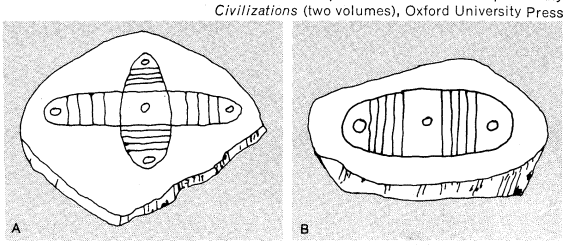


Figure 4: (A) Cruciform board for game Totolospi. (B) Alternate Totolospi board in which cross has become a single track.

Spiral track games

Game of Goose. Many race games use a spiral track. The Game of Goose was invented in Florence under the reign of Francesco de' Medici, 1574 to 1587, who sent one to Philip II of Spain. It spread rapidly to other parts of Europe, reaching England in 1597. Sixty-three points, consecutively numbered, are arranged as a spiral. Any number of players can take part, each having his own marker, which is entered on the first point and borne off from the last. Moves are determined by throws of two dice. If a higher number is thrown than that required to reach 63, the surplus is played backward from 63. Some points are marked with a goose; if a player's marker lands on such a point, that player may make another throw with the dice. Other points bear penalties, requiring the marker to go back, lose a turn, or pay counters as a penalty.

The Game of Goose was followed by many games of similar type, some designed to teach children history, geography, architecture, zoology, and moral improvement. Hundreds of race games of varying ingenuity and interest have been invented, only to be replaced by others.

War games. The games of this group simulate battles between opposing forces.

Shaturanga. Boards invented for one game have been used for others; the 64-square board for an ancient Indian race game, Ashtapada, was taken over about the 5th century AD for the four-handed war game Shaturanga, a forerunner of chess. Each player had eight pieces; a rajah, horse, elephant, boat, and four pawns, or foot soldiers representing the four corps of an army—cavalry, elephants, boatmen, and infantry (Figure 5A). The red and black pieces were loosely allied against the green and yellow. Each piece had a different power of movement and a long die determined the type of piece moved at each turn of play. On throwing a "two" the boat moved, a "three" the horseman; a "four" the elephant; and a "five" the rajah or one of the soldiers.

If a piece moved onto a square occupied by an enemy piece, the latter was removed from the board. Boatmen and soldiers were not permitted to capture major pieces; they could, however, capture each other. The object of the game was for a rajah to seize an enemy throne—the initial square of a rajah. On moving onto this square he won a stake from the defeated player. If a rajah mounted his ally's throne he assumed command of the allied forces as well as his own.

Chess-related games

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

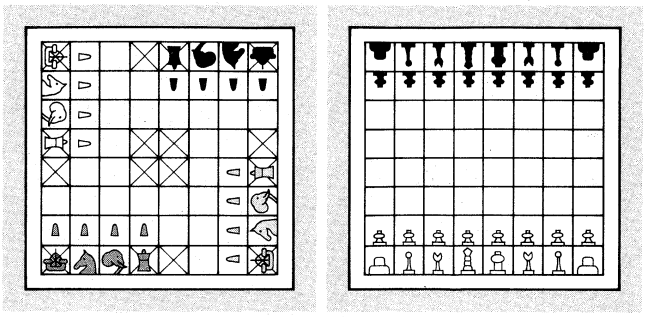


Figure 5: (A) Shaturanga pieces arranged for a game on an Ashtapada board. (B) Arrangement of pieces for Shatranj.

Shatranj. Gambling was forbidden in the ninth book of the Laws of Manu (about the first century BC), and Shaturanga players evaded trouble by discarding the die and turning the game into one of skill. Other changes followed, including the reduction of the four armies to two and making the second rajahs into prime ministers with only half their former power of movement. The new game, Shatranj (Figure 5B), traditionally was introduced into Persia about AD 550; it was played by the Arabs and later appeared at the Byzantine court. It was played in Europe by AD 700, but by then the weak prime minister had become the most powerful piece on the board, the queen. Shatranj also spread eastward from India, entering Burma, Thailand, China, and Japan, with modifications in each.

Chinese chess. In Chinese chess (Hsiang-ch'i) elephants, horsemen, infantry, cannon, and war chariots fight to capture the enemy general, who is confined within a fortress. The two armies are separated by a river, which is impassable to the elephants. The chessmen are cylindrical with the rank written on the upper face; one player's pieces in red and the other's in green. They are placed on the intersections instead of the squares, the board becoming a grid of 9 × 10 points. The moves differ from those in Shatranj—for instance, cannon can only capture by jumping over an intervening piece onto their target. They cannot jump over more than one piece in a move and do not jump unless making a capture. The earliest reference to Hsiang-ch'i was about AD 840.

Japanese chess. Japanese chess, or the General's Game (Sho-gi), is played on a board of 9 × 9 rectangles. The pieces are shaped like little coffins, with the rank written on the upper surface and a promoted rank on the underside. The pieces are all the same colour, and possession is indicated by the direction of the pointed end, a player's pieces being directed toward the enemy. When a piece enters enemy territory it is promoted and turned over, but promotion may be delayed at the player's option—for example, an Honourable Horse with its power of leaping over intervening pieces may be preferred to the more powerful Gold General, since check by an Honourable Horse cannot be covered.

A unique feature of Japanese chess is reintroduction of captured pieces into a player's own formation—instead of moving a piece, a player may enter a prisoner onto any unoccupied space, pointing it toward the enemy. The game ends when one of the Jewelled Generals is checkmate, or unable to move without being captured.

The Alquerque group

Alquerque. Boards similar to those for Alquerque have been discovered in Egypt dating back to about 1400 BC. Figure 6 shows an Alquerque board ready for play. The pieces move from any point to an adjacent point along a marked line. If the adjacent point is occupied by

an enemy piece and the next point beyond it on the line is empty, the player's piece may make a short jump over the hostile piece and remove it from the board. If another piece is then similarly exposed, it is taken in the same move by a second short jump, a change of direction being allowed. In this way, several pieces may be captured in one turn of play. If a piece can make a capture, it must do so, even to the player's disadvantage.

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

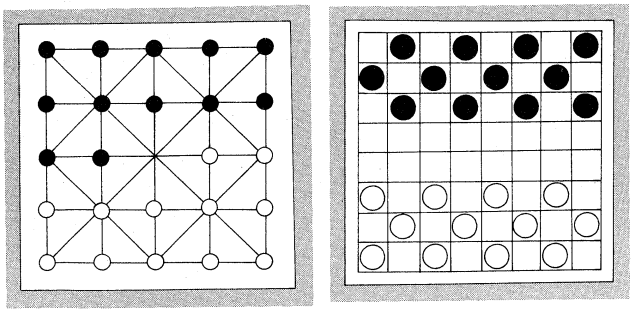


Figure 6: *The Alquerque Group.*
(Left) Starting position of pieces on an Alquerque board.
(Right) English Draughts board with pieces in place.

Draughts, or Checkers. About AD 1100 Alquerque was modified to Les Dames by being played on the black squares of a chessboard (Figure 6, right). Each player had 12 pieces, movement being one square diagonally forward, and capture by a short jump over an enemy piece onto an empty square beyond. Multiple captures were allowed in a turn of play. When a piece reached the opposite side of the board it was promoted to a king and permitted to move diagonally backward as well as forward. The 16th-century addition of compulsory capture under penalty of huffing, or loss of the offending piece, resulted in the modern game of Draughts, or Checkers.

Fanorona. Alquerque likewise sired Fanorona, when, about 1680, the board was doubled, the number of pieces increased to 44, and the method of capture changed. White starts and moves along any line to an adjacent point. If a move ends on a point nearest a point or points in the line of movement occupied by enemy pieces in unbroken sequence, these are captured by approach and removed. Capture may also be by withdrawal—a piece moving away from a point contiguous with a point or points occupied by enemy pieces in the line of movement captures those pieces. Captures are compulsory, but on the first move by each player only one sequence can be taken. On later turns a player may make several captures, either by approach or withdrawal, changing direction each time a capture is made.

The Tafl group

Hala-tafl (Fox and Geese). The war games already described have been between equal forces, but in northern Europe the Tafl group simulated miniature battles between unequal forces. The larger force tried to hem the smaller in, and the latter, equipped with a piece of special power tried to break out or destroy the aggressor.

In Icelandic Hala-tafl 13 geese are arranged as in Figure 7A, and a fox is placed on any vacant point. The fox and geese can move in any direction along a line to the next point. If a fox jumps over a goose and lands on an empty point beyond, the goose is killed and removed from the board. Two or more geese can be killed by a series of short jumps by the fox. The geese cannot jump but try to crowd the fox into immobility to win. If the fox kills enough geese, they become too weak to immobilize the fox and lose the game.

Officers and Sepoys. In later variants the geese were increased to 17 but deprived of the power of moving backward. During the Indian mutiny (1857–1858) the variant Asalto was renamed Officers and Sepoys (Figure 7B). The square represented a fort, and one player placed two officers on any two points within its walls. All pieces moved one point at a time along any marked line, but the sepoys moved only toward the fort. The officers captured by a short jump but were subject to huffing. The sepoys won if they occupied every point in the fort or if they immobilized the officers.

Tablut. In the game Tablut (Figure 7C), described by the Swedish botanist Linnaeus in 1732, eight Swedish soldiers and a Swedish king tried to break through a cordon of 16 Muscovites. The board was marked into 9×9 squares, the central one being a throne that only the king could occupy. All pieces could move any number of unoccupied squares orthogonally—straight along a row or column, not diagonally. A piece was captured and removed from the board when an opponent occupied both adjacent squares in a row or column. This was custodian capture. A piece could, however, move safely onto an empty square between two enemy pieces. The king was captured if surrounded on all four sides by enemy pieces, or on three sides and the throne. If the king was captured, the Muscovites were victorious; but, if he reached any square at the periphery of the board, he escaped from their trap and won the game.

Hnefatafl. Tablut seems to be a simplification of Hnefatafl, played by the Saxons in northern England in the 10th century AD. In this game, a black fleet of 24 ships and a kingship was attacked by a white fleet of 48 ships (Figure 7D). The rules were the same as in Tablut, the kingship being sunk when surrounded on four sides by enemy vessels or victorious when making its escape to the edge of the board.

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

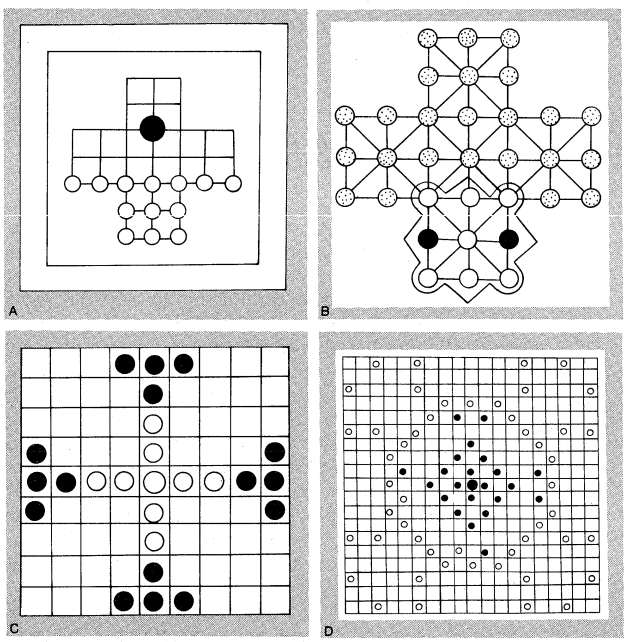


Figure 7: *Game boards.*
(A) Fox and Geese. (B) Officers and Sepoys. (C) Tablut.
(D) Hnefatafl.

Games of position. In these games the players try to place their pieces in winning formations.

Dara. The Dakarkari people in North Africa play Dara with a "board" of 30 small holes dug in the ground (Figure 8A). Each player has 12 distinctive pieces, which are placed one at a time in alternate turns of play into the holes. When all have been placed, or sown, the second phase begins. Each player alternately moves one of his pieces orthogonally (not diagonally) to the next hole, trying to form a line of three pieces in consecutive holes orthogonally. When such a three is formed, the player removes any one of his opponent's pieces from the board. Lines of four pieces do not count. The game ends when one player is unable to make lines of three.

Games of threes

Nine Men's Morris. Nine Men's Morris (Mill or Merelles) is more complex and widespread—board designs have been found cut into the roofing slabs of a temple at

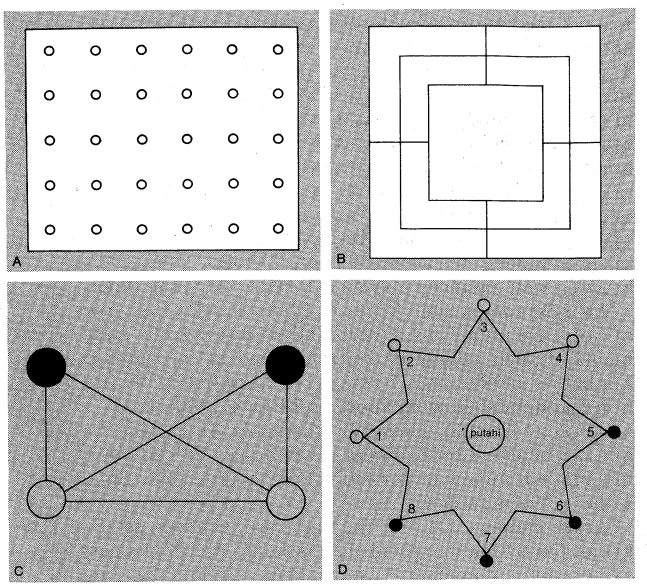


Figure 8: *Game boards.*
(A) Dara. (B) Nine Men's Morris. (C) P'ang ho ch'i.
(D) Mu Torere.
From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

Kurna in Egypt (*c.* 1400 BC), into the steps at Mihintale in Ceylon (AD 21), in prehistoric lake dwellings in Europe, and on the Gokstad Viking ship (*c.* AD 900). Each player has nine pieces and introduces them onto the points of the board (Figure 8B) in alternate moves of play. On forming a line of three pieces, called a mill, he removes one of his opponent's pieces from the board, but not one which is in a mill. When all the pieces have been entered, play continues by alternate moves of pieces to an adjacent vacant point along a line. Pieces cannot pass an opposing mill. A player wins the game by blocking his opponent's pieces or by reducing his opponent's pieces to two.

Blocking games

P'ang ho ch'i. In P'ang ho ch'i (Canton, China), or Onmoul-ko-no (Korea), the players each have two stones of distinctive colour, and they move one stone by alternate turns of play along any line to the next empty point, trying to block the opponent's stones (Figure 8C). The game can be started with the board empty, the stones being introduced in alternate turns of play.

Mu Torere. The Maori in New Zealand play Mu Torere on an eight-pointed star (Figure 8D). Each player has four pieces, which are placed on four adjacent points, the centre or *putahi* being empty. The players then move a piece alternately. Only one piece is allowed on each point, and jumping over pieces is illegal. There are three possible types of move: from one point to an adjacent empty point; from a point to the empty *putahi*, provided that one adjacent point is occupied by an enemy piece; and from the *putahi* to an empty point. The player blocking his opponent wins.

Games of territorial possession

Go. The game known as Go or I-go in Japan, Wei-ch'i in China, and Pa-tok in Korea is ranked by many as the greatest intellectual game. The leading exponents are Japanese, and Go has an enormous following. The board is marked into a grid of 19 × 19 lines forming 361 points. Nine of these are marked to help in orientation and handicapping. Each player has a bowl containing disk-shaped stones. Black has 181 stones, and White 180. At the beginning the board is empty. The weaker player takes the black stones and plays first, which gives him a slight advantage. He places a stone on any point, and play continues alternately, one stone being placed at each turn. Once played, stones remain until the end of the game unless they are captured and removed from the board.

The players try to control vacant points in such a way that they cannot be occupied by an opponent's stones, and the player who possesses the most vacant points at the end of the game wins. Stones completely surrounded by those of the opposite colour without empty points orthogonally adjacent to them are captured and removed from the board (Figures 9a and 9b), but this is not the primary object of the game. It is possible to win without making a single capture. A stone cannot be placed on a point completely surrounded by enemy stones unless it makes a capture by so doing (Figures 9c and 9d). A stone cannot occupy the last free point of one of its own groups unless enemy stones are captured by this action (Figure 9e).

Vacant points controlled by stones of one colour are called eyes, and a group with two eyes is impregnable (Figure 9f). A group of stones in diagonal contact may contain empty points, but the disconnected stones can be attacked and the formation killed. Such empty points are "false eyes" and may be mistakenly constructed by the novice instead of real eyes (Figure 9f).

When a player has just captured a stone in a repetitive position known as a *ko* (Figure 9g), his opponent must make one play elsewhere on the board. This rule prevents the likelihood of perpetual positions and a drawn game. If there are three *ko* on the board at once, however, the game is drawn.

If opposing formations are interlocked in an impasse, or *seki* (Figure 9h), they are left alone until the end of the game and then discarded.

At the end of the game any vacant points between opposing formations that are not controlled by either side

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

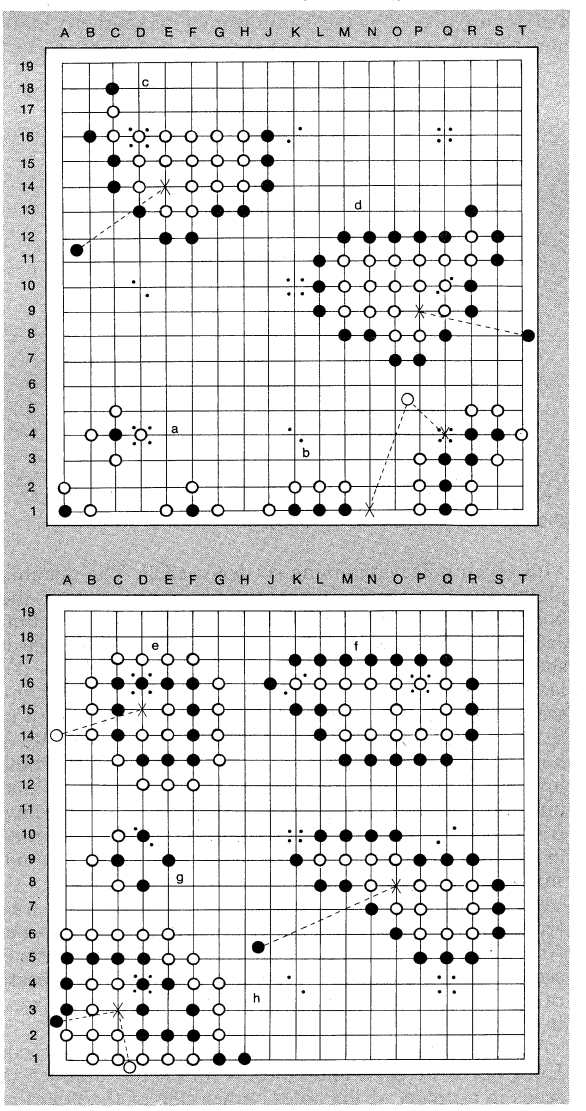


Figure 9: Game boards for Go, showing playing situations.

are filled in to help in adding up the final score. Stones that are not surrounded, but can inevitably be so, are "dead" and are removed at the end of the game without further play. When profitless points have been filled, and dead stones removed from each player's territories, each player places his opponent's captured stones on vacant enemy points, thus reducing the opposing score by the number of pieces held.

Players of equal skill play Black alternately, but if one wins three times consecutively he gives his opponent a handicap of two stones and increases this gradually until they meet on equal terms. Casual games may last about two hours, and professional matches up to three days.

Mancala games. These are games in which players distribute pieces into rows of holes under varying rules that allow accumulation of pieces by capture. They are widely played in Africa and Asia and areas influenced by those cultures. The games appear to have originated in ancient Egypt. Sets of deeply cut holes have been found in the roofing slabs of the Kurna temple in Egypt (*c.* 1400 BC), and other sets are cut into the summit of the damaged portion of the great pylon built in Ptolemaic times (305–30 BC) at the entrance to the temple of Karnak. These boards consist of two rows of six, seven, and eight saucer-shaped hollows. Boards have also been found in Arabia antedating Muhammad, whose followers carried variations of the game throughout Islām. Negro slaves took Mancala games to the West Indies, where forms are still played identical with those in remote villages in west Africa.

Pallanguli. One of the simplest forms of Mancala is Pallanguli, played by the Tamil women of southern India and Ceylon (now Sri Lanka) (Figure 10). Players start

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

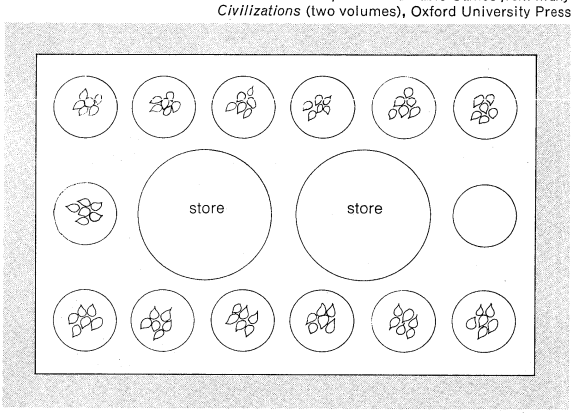


Figure 10: Game board for Pallanguli.

with seven holes containing six seeds each. The opening player lifts the seeds from any hole on her side of the board and moving counterclockwise sows one seed into each hole. If she reaches the end of her side of the board, she continues sowing in her opponent's holes. When the last seed of a lift falls into a hole she picks up all the seeds in the next hole and continues sowing as before in the same direction. If the last seed of a lift falls into a hole with an empty hole beyond, any seeds in the hole immediately beyond the empty hole are captured and put into the player's store hole; she then continues play from the next loaded hole beyond; but, if the last seed of a lift falls into a hole with two empty holes beyond, she wins nothing and her turn ceases.

Her opponent then plays by lifting the seeds or seed from any hole on her side of the board and plays as before. The game continues by alternate turns of play. Four seeds in a hole is called a cow; they become the property of the owner of the hole and are lifted at once and placed in the player's store while play continues.

At the end of the first round each player lifts the seeds from her store hole and restores to six the number of seeds in as many of the holes on her side of the board as possible, any remainder being returned to her store. The loser of the first round will not be able to fill all her holes, and these "rubbish holes" are marked with a little stick. The winner of the round fills all her seven holes, and the surplus is placed in her store. The player with first move in the first round has second move in the second round, each round being played in this way, and the rubbish holes are left empty. The game ends when one player is reduced to less than six seeds and is unable to fill even one hole at the beginning of a round.

During any round the losing player may win enough seeds to reopen one or more rubbish holes, and a game between well-balanced players can last a very long time.

The most complicated forms of Mancala with four rows of holes are found in southern and eastern Africa. Fragments of such boards have been found in the ruins of Zimbabwe in Rhodesia, remains of an African civilization that flourished from about AD 1400 to 1800.

Chisolo. Chisolo, played by the Baila of Zambia, requires a board of four rows of seven holes scooped out of the ground with a larger store hole at one end. Small stones serve as counters.

Career games. During the last half century several games have been marketed that simulate life in business, exploration, scholarship, journalism, diplomacy, and crime detection. Most games in this group suffer from a "point of no return"—initially there is an interesting struggle for power, but at a critical point one player outstrips the others to such an extent that they cannot defeat him, though victory may be long delayed.

Monopoly

One of the best known career games is Monopoly (copyright 1935). The players buy, rent, and sell properties, striving to become business tycoons. Each player starts with an equal amount of play money and a marker, which moves around the board according to the throws of two dice. Each player throws the dice to begin and the highest scorer leads by placing his marker on the corner marked "GO," throws the dice and moves his marker as directed. Play continues with the player on his left. More than one marker can rest on the same space at the same time.

When a marker lands on a space not already owned, the player may buy it from the bank. The board is marked out with spaces indicating building sites, railway stations, utilities, rewards, and penalties. The banker holds 32 green houses and 12 red hotels and two sets of cards for "chance" and "community chest" spaces. There are title cards for every property. The banker pays salaries and bonuses, sells properties to the players, and loans money when required on mortgages.

Word games. The popularity of crosswords has led to the invention of several word games played on boards.

Scrabble

One of the best is Scrabble, for two, three, or four players, first marketed in 1948. A set consists of a board marked with 15 × 15 squares, and 100 wood or plastic tiles, each bearing a letter and a number. The tiles are shuffled face downward and each player takes one. The holder of the tile nearest the beginning of the alphabet draws first and begins the game. The tiles are put back face downward and shuffled. Each player chooses seven tiles and places them on a rack. Play is clockwise.

The opening player combines two or more of his letters to form a word and places them on the board, with one letter on the centre square, to read either across or down. Diagonal words are not allowed. The player completes his turn by counting and announcing his score, and then draws as many new letters as he has played, thus retaining seven letters in his rack.

The next players add one or more letters to those on the board to form new words. Each player's score is recorded after each turn of play. When two or more words are formed in a single turn, each is scored, and the common letter is counted in the score for each word. Any letter of a word placed on an ordinary gray square earns the numerical value marked on the tile. Tiles placed on certain coloured squares earn bonus scores. At the end of the game each player's score is reduced by the sum of his unplayed letters, and if one player has used all his letters, his score is increased by the sum of all unplayed letters in his opponent's hands.

TILE GAMES

Dominoes. Dominoes appear to have originated in China, and the Chinese set represents the possible throws of two Chinese dice, the latter having six faces marked with a large red pip, two black pips, three black pips, four red pips, five black pips, and six black pips. The colour of the pips is important in some domino games. The sixes are half red pips and half black (Figure 11). A Chinese

From R.C. Bell, *Board and Table Games from Many Civilizations* (two volumes), Oxford University Press

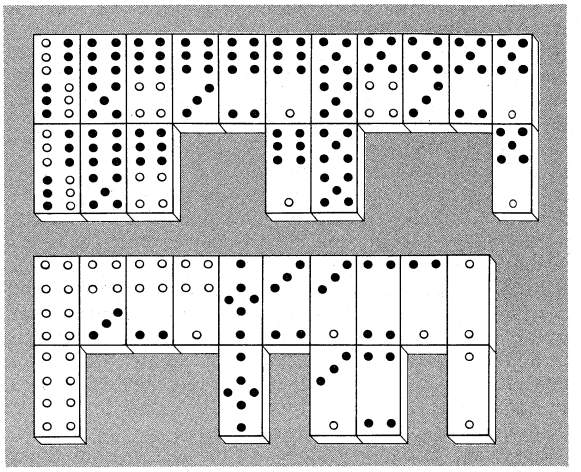


Figure 11: Set of Chinese Dominoes.

domino set consists of 32 tiles, some being duplicated and belonging to the civil series, the others for the military series. Identical tiles of the civil series pair together, while tiles of the military series pair by total count.

Dominoes appeared in Europe at the end of the 18th century, possibly imported from China, but with three important differences: there were no red pips, no duplicate tiles, and a blank sequence had been added. A set consisted of 28 tiles. Double-nine and double-twelve sets are obtainable but are less popular than the double-six. The best dominoes are made of ivory or bone backed with ebony.

Probably the simplest and most representative of the many games played with dominoes is the Block Game. Tiles are shuffled face downward and each player draws one; the player with the highest double, or if no double is drawn, the tile with most pips, becomes the leader. The dominoes are returned to the pool, reshuffled, and the players draw five dominoes each; if there are only two players, they draw seven each. The leader plays a domino, and the player on the left matches one end if possible. If he cannot he loses the turn and the next player tries. When no one can play, the game is blocked, and each player counts the pips on the tiles remaining in his hand. The player with the lowest number scores the total number of pips in his opponents' hands plus those in his own. The first player to reach 121 wins the game.

Other domino games include All Fives (Muggins), All Threes, Threes and Fives, Sniff, Domino Whist, and Matador.

Ma-jong (Mah-Jongg). Ma-jong became popular in China about 1900. It is played with a set of 136 tiles, 128 counters, two dice, a *tong* box, markers, and four stands for the tiles. There are four of each kind of tile, classed as honour and minor tiles. Honour tiles, which count double, are the red, white, and green Dragons; the East, South, West, and North Winds; and the ones and nines of the suits. Minor tiles are the twos through eights of the three suits—bamboos, circles, and characters. Many sets have eight additional tiles—four Flowers and four Seasons—but these are rarely used.

Play begins by building four walls seventeen tiles long and two high, all face down. The players then throw a pair of dice, the highest scorer becoming East Wind. East throws again, and starting with his own wall, counts counterclockwise around the walls, until he reaches the wall indicated by the throw of the dice. He then breaks this wall at the same number from the right end of the wall and places these two tiles on the first and third pair of tiles. These two "loose" tiles form, with the six pairs of tiles beneath them, the dead wall.

East takes the first four tiles to the left of the breach, South the next four, followed by West and North, continuing in this order until all players have drawn 12 tiles. Each then draws one more, except East, who draws two.

East starts by discarding an unwanted tile onto the centre of the table; and South follows by picking it up if he needs it to make a pung (three of a kind), or taking a tile from the wall, and then discarding a tile into the centre. This continues around the table, each player trying to collect four pungs and a pair; the first to do so calling "ma-jong." If a player holds a pair in his hand and a third of the same kind is discarded, he may call "pung" and take it out of the discard pile even if it is not his turn to play. Adding it to his pair, he places the three tiles face upward by his rack. This is an exposed pung and worth two points. Honour tiles count double.

If a player has a concealed pung in his hand, and the fourth tile is discarded, the player may call "quong," add it to his pung and expose them on the table, scoring eight points. A concealed quong is 16 points and is placed face downward on the table. On forming a quong the player takes one of the loose tiles off the wall into his hand. If the second loose tile is used, two more tiles are lifted from the dead wall to form new loose tiles.

To go ma-jong quickly, a player may include one non-scoring sequence of three suit tiles, called a chow. A tile for a chow may come from the wall or from a discard but may only be drawn in turn. If two players want a discarded tile, a player going ma-jong takes first priority, next one who needs it for a pung or quong, and lastly one who needs it for a chow.

The last 14 tiles in the wall, excluding the loose tiles are not used. If no one has completed his hand when this section of the wall is reached, the game is drawn.

There are several special hands not described here, and scoring is complicated. In Western Ma-jong hands must be cleared and contain tiles of only one suit or honours to go ma-jong.

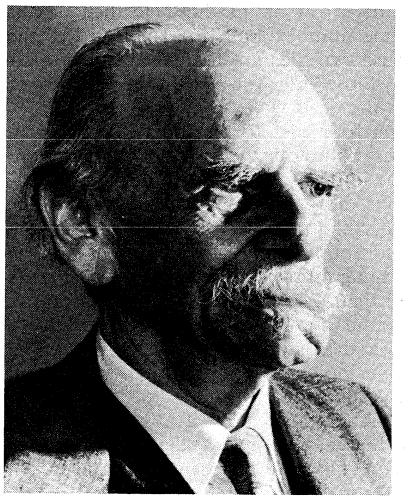
BIBLIOGRAPHY. R.C. BELL, *Board and Table Games from Many Civilizations*, 2 vol. (1969), a general review of the subject; S. CULIN, *Games of the Orient: Korea, China, Japan* (1958), original research by a distinguished anthropologist; C.T. DOBREE, *Gambling Games of Malaya* (1955), handbook written by a police commissioner for the use of lawyers, magistrates, and police officers; E. FALKENER, *Games Ancient and Oriental, and How to Play Them* (1961), highly speculative reconstructions of ancient games; E. HOYLE, *Hoyle's Rules of Games: Descriptions of Indoor Games of Skill and Chance*, 21st ed. (1963), a standard work for two centuries; T. LEGGETT, *Shogi: Japan's Game of Strategy* (1966), a clear description of a complicated game; H.J.R. MURRAY, *A History of Board-Games Other Than Chess* (1952), a scholarly work, but containing too much detail for the general reader; J. ORLEANS and E. JACOBSON, *How to Win at Scrabble* (1955), a description of basic points in play; M. ROBERTSON, *The Game of Mah Jong* (1964), a clear account of the Western game, and the forming of special hands; K. TAKAGAWA, *How to Play Go* (1956), a simple introduction by a former world champion, *The Vital Points of Go* (1958), more advanced advice from the same author; F.R.B. WHITEHOUSE, *Table Games of Georgian and Victorian Days* (1951), based on a large private collection formed by a director of an English game manufacturing firm.

(R.C.Be.)

Boas, Franz

Franz Boas was the founder of the kind of general anthropology—relativistic, culture-centred, nonracist—that became dominant in the 20th century. He was a specialist in the cultures and languages of American Indians. But he was also the purposeful organizer of a profession and the great teacher of most of two generations of scientists who developed anthropology in the United States.

Born in Minden, a small town in Westphalia, now in



Boas, 1941.
Wide World Photos

West Germany, on July 9, 1858, the son of a merchant, he was of delicate health as a child and spent much of his time with books. His parents were free-thinking liberals who held to the ideals of the revolution of 1848. Although Jewish, he grew up feeling completely German. From the age of five he took an interest in the natural sciences—botany, geography, zoology, geology, and astronomy. While studying at the *Gymnasium* in Minden he became deeply interested in the history of culture, although this was not taught in the school. He followed his various intellectual bents at the universities of Heidelberg, Bonn, and Kiel, taking his Ph.D. in physics and geography at Kiel in 1881.

Early years

After a year's military service Boas continued his studies in Berlin, then undertook a year-long scientific expedition to Baffin Island in 1883–84. Firmly interested now in human cultures, he took posts in an ethnological museum in Berlin and on the faculty of geography at the University of Berlin.

On his way back from a study of the Indians of Vancouver Island in 1886, he stopped in New York and decided to stay. He found a position as an editor of the magazine *Science*, married Marie A.E. Krackowizer, and began his long academic career.

Boas' first teaching position was at the newly founded Clark University in 1889. Next, he spent a period in Chicago, where he assisted in the preparation of the anthropological exhibitions at the 1893 Columbian Exposition and held a post at the Field Museum of Natural History. In 1896 he became lecturer in physical anthropology and in 1899 professor of anthropology at Columbia University. From 1896 to 1905 he was also curator of anthropology at the American Museum of Natural History in New York; in that capacity he directed and edited the reports submitted by the Jesup North Pacific Expedition, an investigation of the relationships between the aboriginal peoples of Siberia and of North America.

Career in America

From his earliest years in America, Boas was an innovative and prodigiously productive scholar, contributing equally to statistical physical anthropology, descriptive and theoretical linguistics, and American Indian ethnology, including important studies of folklore and art. His personal research contributions alone would have given him an important place in the history of science, but he also exerted enormous influence as a teacher. By the turn of the century, national leadership in anthropology was firmly in Boas' hands. In 1906, at the age of 48, he was presented with the *Festschrift* usually awarded by his colleagues to a scholar nearing retirement. The 36 years that followed were no less productive, influential, or honoured. Boas established the *International Journal of American Linguistics*, was one of the founders of the American Anthropological Association, president (1931) of the American Association for the Advancement of Science, and a member of many other scientific societies.

In 1911 Boas published *The Mind of Primitive Man*, a series of lectures on culture and race. It was often referred to in the 1920s by those who were opposed to new U.S. immigration restrictions based on presumed racial differences. In the 1930s the Nazis in Germany burned the book and rescinded his Ph.D. degree, which Kiel University had in 1931 ceremonially reconfirmed. Boas updated and enlarged the book in 1937; the revision had an influence in the civil rights struggle that took place in the U.S. during the 1950s.

Boas and his wife had six children, only three of whom survived them. Although Boas suffered a heart attack in 1931, he remained professionally active until December 22, 1942, when he died at the age of 84.

Significance

The revolutionary significance of Boas' work can only be understood in terms of the differing beliefs of anthropologists about man. Almost all anthropologists have almost always believed that the human species is one; but not as many of them believed in Boas' day that the races of mankind show equally the human capacity to develop cultural forms. It is partly because of Boas' influence that the proposition is now almost universally accepted that every surviving population large enough to have a distribution of individual differences shows equally the human capacity to develop cultural forms; and that differences in outcome are attributed by anthropologists to historic "cultural" rather than genetic factors.

Within this common framework there have sometimes been differences in view as to the actual attainments of particular peoples. Some anthropologists, often calling themselves "evolutionary," argue that some peoples have achieved "higher" states of culture, leaving behind—at least temporarily—other peoples. They believe that the differences between "civilized" and "primitive" peoples are the result of environmental, cultural, and historical circumstances. Other anthropologists, frequently called cultural relativists, argue that the evolutionary view is ethnocentric, deriving from a human disposition to characterize groups other than one's own as inferior, and that all surviving human groups have evolved equally but in different ways.

Franz Boas was of the second persuasion. Since British and U. S. anthropologists in the last third of the 19th century were not particularly disposed to this view, Boas' success in making it overwhelmingly dominant was all the more remarkable. While he had originally assumed as a natural scientist that universal laws must exist that would explain how different peoples have wound up with their characteristic ways of life, he concluded that the problem was too complex for any general solution. Laws of cultural causation, he argued, had to be discovered rather than assumed.

Boas' view requires the anthropologist to be capable of understanding all factors that might influence the histories of peoples. Thus, to assert that cultural differences are not the result of biological differences, one must know something of biology; and to see the interrelations of man and his environment, the anthropologist must understand such things as migration, nutrition, child-raising customs, and disease, as well as the movements and interrelations of peoples and their cultures. Anthropology then becomes holistic and eclectic, involved in any field of science or scholarship that appears relevant to a particular problem.

BIBLIOGRAPHY. There is no definitive biography of Boas. Useful discussions of the man and his work may be found in the following: GEORGE W. STOCKING, *Race, Culture and Evolution* (1968); WALTER R. GOLDSCHMIDT (ed.), *The Anthropology of Franz Boas: Essays on the Centennial of His Birth* (1959); JUNE HELM (ed.), *Pioneers of American Anthropology: The Uses of Biography*, pp. 83–222 (1966); MELVILLE J. HERSKOVITS, *Franz Boas: The Science of Man in the Making* (1953); ABRAM KARDINER and EDWARD PREBLE, *They Studied Man*, pp. 134–159 (1961); ALEXANDER LESSER in the *International Encyclopedia of the Social Sciences*, vol. 2, pp. 99–110 (1968); and *Dictionary of Scientific Biography*, vol. 2, pp. 207–213 (1970).

(So.T.)

Boating and Yachting

Boating and yachting for pleasure have been enjoyed for several centuries but have increased tremendously in popularity in the 20th century, particularly in North America, western Europe, and Australasia. Though boating is most often engaged in as a pleasure for its own sake, it is also enjoyed as a competitive sport.

The sports aspects of boating are dealt with in this article under the following headings:

I. Rowing
  History
    Early boat racing
    Development as an organized sport
  Competitive rowing
    Major events
    Course and equipment
    The race
    Technique
    Strokes and styles
    Women's rowing
II. Canoeing
  History
    Early canoes and kayaks
    Development of modern canoes and of the sport
    Development of organized canoeing
    Canoeing as a recreation
  Competitive canoeing
    Canoe classes and events
    Courses, equipment, and specifications
III. Yachting
  History
    Early English and American yachts
    Beginnings of organized yachting
    Development of design and rigging
    Rating rules
    Types of boats
    Yacht clubs and organizations
  The sport of sailing
    Courses
    How a boat sails
    The America's Cup
    Offshore sailing
    Small-boat racing
IV. Motorboating
  Speedboats
    History
    Speedboat design
    Speeds
    Racing
  Cabin cruisers
  Other motorboats
    Motor sailers
    Houseboats
    Utility boats
    Other variations

I. Rowing

Rowing, technically, is the manual propulsion of a boat using a single oar as a lever in a succession of strokes. The performer is called an oarsman; rowing is an activity for pairs of oarsmen. The use by one person of two oars, one in each hand, is termed sculling. An oar consists of a shaft of wood with a rounded handle at one end and a shaped blade at the other. The shaft may either be solid or consist of two halves partially hollowed out and glued together. The latter construction is used for racing oars in order to save weight and increase flexibility. The blade—a thin, broadened surface—is either flat overall or slightly curved at the sides and tip to produce a firmer grip of the water. The loom, or middle portion of the oar, rests in a notch or oarlock (rowlock) or between tholepins on the gunwale of the boat or on swivels on outriggers (called riggers), which serve as a fulcrum for the oar. The loom is protected against wear in this area of contact with a short sleeve of leather or plastic. Rowing boats used by fishermen often are equipped with modifications of these devices designed to keep the oar from slipping overboard if released; these may include, for example, a steel pin fitted to the gunwale or small outrigger, over which an eye, drilled through the loom of the oar and protected with a steel collar, is slipped. Racing oars have fixed leather or adjustable metal or plastic collars—called buttons—to prevent slippage outboard.

Description of rowing

The oarsman, seated with his feet resting on a brace, or stretcher, covers (dips) his blade square in the water (the entry). Then he uses his legs, shoulders, and body weight to exert pressure against the oar, which functions as a lever turning around its fulcrum to drive the craft past the point of entry. He completes the stroke by lifting the blade out of the water (the finish) and swings forward to begin the cycle again. No stroke is completely efficient, as some degree of blade slip occurs in the water. Fundamental factors in the mechanics of rowing are the position of the oar button, the height of the rigging, the setting of the thole or swivel pin, and outboard length of the oar, which together determine the amount of leverage available.

The characteristics of rowing apply equally to the use of an oar in each hand, or sculling, but a scull is much shorter than an oar. In both methods, the oarsman or sculler is seated on a thwart or sliding seat facing the stern of his craft. Among commercial watermen, sculling is the propulsion of a boat using a single, long oar with a flat blade worked to and fro from a notch in the stern transom. The blade is made to describe a figure eight beneath the surface, acting like the tail of a fish to give both propulsion and direction.

HISTORY

Rowing in ancient times was the principal method of propelling vessels of war and ships of state. As the size of such vessels increased, sails gradually displaced oars, although large galleys continued to be rowed in the Mediterranean until the 18th century. The oarsmen, generally prisoners of war or criminals, were chained to their benches, whence the term galley slaves. As the oars were extremely long overall (in order to reach the water below), it was not unusual for several men to handle each oar. Their combined weight was required to raise the oar out of the water and their combined strength to pull it. Galleys were rated according to the number of tiers of oars. The first recorded Roman fleet consisted of triremes (three-tier galleys). The complement of a galley was considerable; *e.g.*, Amenhotep II of ancient Egypt is recorded as the tireless stroke of a crew of 200 men. The first recorded amateur oarsmen were the islanders who entertained Odysseus on his return to Ithaca.

The earliest invasions of England were effected with the help of oars. In 54 BC Julius Caesar depended largely on oars when crossing the English Channel. Like the Romans, the Anglo-Saxons and later the Danes rowed and sailed across the North Sea to enter the estuaries along the east coast of England.

Early boat racing. Boat racing is first recorded at a Venetian regatta (Italian *regata*) that featured a gondola race in 1300. In 1529 the first recorded boat race for women took place, also in Venice.

The watermen of London

In England, racing in rowing boats dates from the time when the only bridges across the lower reaches of the Thames were London and Chelsea. Anyone wishing to cross the river elsewhere had to hail a ferry, typically a light sculling boat, or skiff, operated by a waterman. The occupation of waterman was defined by statute as "the trade of rowing," and in Queen Anne's time (reigned 1702–14) about 10,000 watermen were licensed to earn their living on the Thames above London Bridge. They all wore special livery, some being particularly magnificent. Wagering developed between the gentry as to the relative merits (speed and skill) of the watermen manning the ferries, and stakes sometimes ran high. The ferry craft became known as wager boats, and it was for these watermen that the race for "Doggett's Coat and Badge" was instituted in 1716. Thomas Doggett, its founder, was an actor who blandly decreed in his bequest that the race should be rowed "for ever," and so far it has been, annually each summer between those earliest bridges—London and Chelsea. The inaugural race commemorated the succession of the House of Hanover to the throne of England. The winner received an orange-coloured uniform; in the years since, the livery has changed and is now red, and the status of the competitor has altered from professional waterman to amateur rower.

The Henley Royal Regatta. The reaches of the Thames at Henley are the most beautiful along the river and, because of a straight stretch of more than a mile immediately below the town, offer an ideal course for rowing races. Thus, in 1839 the townsfolk included boat racing as a special attraction to bring visitors to their market town for the annual fete. In 1851 Prince Albert became patron and gave the regatta its royal prefix.

The regatta has attracted competitors from all parts of the world, and the Grand Challenge Cup for eight oars (established 1839) and the Diamond Challenge Sculls for single scullers (1844) have long been the most coveted trophies in rowing. Other events open to all amateurs are the Stewards' Challenge Cup for fours, the Silver Goblets for pair oars, and the Double Sculls (added 1939). The first trophy to be won by a crew from outside England was the Stewards' Cup (1847, for colleges and schools), won by Columbia University in 1878. The Princess Elizabeth Cup for schoolboys was established in 1947.

Other early races. The first recorded boat race in the United States was in 1811, when the ferrymen of Whitehall in New York City defeated their Long Island and Staten Island rivals on the Hudson. In 1824, in the first Anglo-American contest, they outrowed a crew of Thames watermen from the visiting British frigate "Hussar," in a four-mile race. Rowing as a sport is first recorded in Tasmania at a regatta in 1830, and in Australia in 1832; and the first rowing club in New Zealand was formed at Canterbury in 1861.

The sport was introduced in Russia in 1842 by the British colony in St. Petersburg, the rowers racing annually on the Neva River for a pair of silver sculls. In Austria, Belgium, Denmark, France, Finland, Germany, The Netherlands, Italy, Norway, Poland, Switzerland, and Yugoslavia the sport of rowing has also been engaged in for many years (see below *Major events*).

Development as an organized sport. The first form of rowing by amateurs in England was a simple imitation of the races between ferrymen. Sculling was coupled with boxing in the early days of rowing as a sport. The great oarsmen included some of the best boxing champions of their day, and some rowing clubs had boxing sections. Rowing began as a sport at Oxford in about the 1790s, when "the caps and tassels of the students formed a curious contrast with their employment at the oars." By 1805 the students were using six-oared boats borrowed from local people, and in the following year Eton College introduced the sport. Westminster School followed suit in 1813, when it started its *Water Ledger* (minutes book). Racing in eight-man boats is first recorded at Oxford in 1815, when Brasenose College led all others.

Rowing at English schools

Many clubs sprang to life but disappeared almost as quickly. There were no amateur regattas, so racing took the form of private challenges, often for a side stake. There were, however, annual regattas for professionals, the earliest being at Chester before 1814. Leander Club was founded in about 1815 and rapidly acquired the prestige it still enjoys. For nearly 150 years its crews were composed mostly of Oxford and Cambridge varsity oarsmen, but in the late 1960s its Henley-winning eights were pre-university students. Cambridge took up rowing later than Oxford, because its river, the Cam, was less suited to the sport than the Thames. Most of the oldest English clubs were founded between 1840 and 1870, and the Amateur Rowing Association was formed in 1882.

In Australia, rowing is organized in each of the six states. Most of the clubs are located in Melbourne and Sydney, and there has been competitive interstate rowing since 1863. In Canada, the 1870s were the heyday of the professional scullers. As the distinction between professional and amateur rowing became more sharply drawn, Canada, like England, developed numerous amateur rowing clubs. Among the rowing centres are Brockville, Toronto, Hamilton, St. Catherines, Vancouver, and Winnipeg.

In the United States the first organization of amateur clubs was the Castle Garden Boat Club Association of New York (1834). Light, keelless racing shells appeared in the U.S. soon after their introduction in England (1847), and in 1857 it was an American, J.C. Babcock of Nassau Boat Club, New York, who conceived the idea of the sliding seat, which did not replace the fixed seat in England until 1871. The boat clubs along the Schuylkill River at Philadelphia were organized as the Schuylkill Navy in 1858. The National Association of Amateur Oarsmen (NAAO) was founded in 1872.

Intercollegiate races

The oldest intercollegiate contest in the U.S. is the Yale–Harvard boat race (it antedates football by 17 years), first rowed in 1852 on Lake Winnipesaukee. In the 1870s rowing became popular at a number of Eastern colleges. In 1895 the Poughkeepsie Regatta was established. It attracted all the foremost collegiate and university crews to the Hudson. Rowing was established on the Pacific coast in 1899, and the Washington–California race soon became the rowing feature of the Far West. In 1902 the American Rowing Association was formed to increase intercollegiate competition with early-season short-distance sprints, concluding with an annual regatta over the Henley distance of 1 mi 500 yd (approximately). Thus, this regatta became popularly known as the American Henley.

The international authority and governing body for international competition is the Fédération Internationale des Sociétés d'Aviron (FISA). There are 40 member countries, including all those mentioned above, together with Argentina, Brazil, Bulgaria, Ceylon, Chile, Cuba, Czechoslovakia, Greece, Hungary, Republic of Ireland, Israel, Korea, Morocco, Peru, Poland, Romania, South Africa, Spain, Sweden, Turkey, United Arab Republic, and Uruguay.

COMPETITIVE ROWING

The governing bodies for rowing in each nation are responsible for establishing the qualifications necessary for anyone to engage in amateur rowing competition. It is the responsibility of each national governing body to hold a championship regatta each year and to select those who are to compete in international events organized by FISA.

FISA is responsible for organizing and conducting the rowing events in the Olympic Games. In addition, it conducts a European championship regatta in most years and a world-championship regatta every four years between the Olympics. A junior championship was introduced in 1970. Competition in all FISA regattas is open only to member nations. For the Olympic regatta, entries from any nation are acceptable, provided that the amateur status of competitors is in accordance with the Olympic rules.

In most rowing countries there are several regattas of national importance. In addition to open competition, rowing clubs throughout the world still hold private, annual matches with traditional rivals.

Crews race side-by-side at regattas, sometimes six or more abreast on inland courses and twice as many on the sea. Processional racing (racing in which competing crews or scullers start in procession and are timed individually at start and finish) has also become well established in a number of countries since 1926, when the Head of the River Race on the Thames was founded. It is rowed over 4¼ mi (7,200 m) between Mortlake and Putney. Over 3,000 competitors take part annually in more than 300 shells.

Major events. The oldest event of world renown in rowing is the Oxford versus Cambridge boat race (the University Boat Race), which was first rowed at Henley in 1829. Seven years later the second race took place from Westminster to Putney. There were four more races over this course before the present 4¼-mi Putney to Mortlake course was first used, in 1845. In 1856 it became an annual event that was interrupted only by World Wars I and II. Henley Royal Regatta (see above *The Henley Royal Regatta*) is the oldest major event in the rowing calendar, but it does not have international recognition because the course is not still water and is also too narrow to meet FISA requirements (see below *Course and equipment*).

The European Championships. The European Championships were founded in 1892 and included eights, coxed fours and sculls (see below *World Championships*). France was outstanding in the early days of the championships, doing equally well in eights, four-man boats with coxswain (fours with coxswain, or coxed fours), and single sculls. Belgium also held a dominant position in European rowing throughout the 20 years preceding World War I. The country's record of winning the European eights title seven years in succession (1897–1903) has never been equalled, and Belgian fours with coxswain were European champions 11 times (1897–1907). Swiss oarsmen were the first European fours with coxswain champions in 1893, and they have won titles in all seven classes. Germany, The Netherlands, and Italy have also been in the forefront of rowing in Europe since before 1930, and since then Austria, Denmark, Hungary, and Yugoslavia have also made their presence felt.

Leading European and Olympic champions

Olympic championships. The first modern Olympiad, held in Athens in 1896, did not include rowing, but the sport did find a place in the 1900 games held in Paris. In 1904 the games took place in St. Louis, Missouri, where the United States made a clean sweep in all four rowing events. Great Britain did likewise in the 1908 Olympic regatta at Henley. The United States, Germany, and Great Britain have been the most successful medal winners in the 16 Olympic Regattas held through 1972.

World Championships. The World Championships were founded in 1962, and they include the same events as the Olympics: eights, fours with and without coxswain, pairs with and without coxswain, and double and single sculls. Over two dozen countries compete, and the prestige of winning is prized among oarsmen as highly as an Olympic rowing title. The only other event of international importance beyond the confines of a single continent is the North American Championships, which were first held in Canada in 1967.

Course and equipment. All races under FISA rules must take place over a 2,000-m (6,600 ft) straight course on still water. Each crew or sculler races in a separate lane marked with buoys. There are six lanes, and it is customary for crews who fail to reach the finals to row for seventh to 12th place in a separate final.

Racing shells

Racing shells are used in all major events. They range in length overall from 58 ft (18 m) for an eight, down to 42 ft (13 m) for a four, 32 ft (10 m) for a pair, and 25 ft (8 m) for a single sculling boat. The weight varies according to the materials used for the skin, riggers, and framework; *e.g.*, a sculling boat without riggers can weigh as little as 20 lb (9 kg), but if it is to last several seasons it is more likely to be 5 lb (2 kg) heavier.

The weight of oars also depends on the method of construction, while the overall dimensions can vary appreciably depending upon the specification required by individual coaches or crews. Oars—with a wide variety of blade shapes and sizes—are mostly variations on an overall length of 12 ft (4 m) to 12 ft 6 in with a 24-in- (61-cm-) long blade that is seldom more than 6¼ in (15.9 cm) wide at the tip and 7½ in (19 cm) at its widest point of curvature. Two other key factors in racing shells are the type and setting of the outrigger and the length of the slide along which the seated oarsman travels each stroke. Here again, the choice is individual to coaches or crews. Light riggers save weight but are not so durable. Friction also has to be considered in deciding which kind of slide and seat runners to select. The length of a slide is usually between 26 and 29 in (66 and 74 cm).

Ewing Galloway

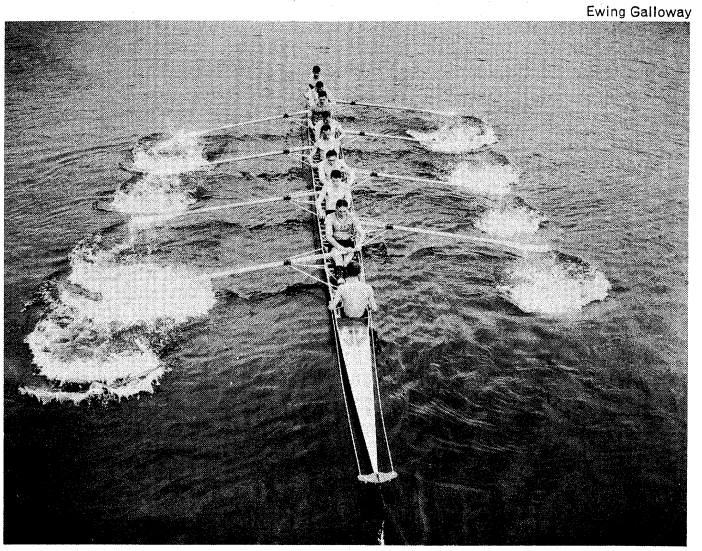


Figure 1: Crew stroking an eight-oared shell.

The race. The following synopsis of the rules governing competition is intended to aid the spectator in understanding the start, conduct, and finish of a race, including why a boat may be disqualified or a race re-rowed. In boat racing a competitor not at the start at the specified time may be disqualified. The race is started when the starter is satisfied that the competitors are ready. He holds up a flag, says "Are you ready?" and if there is no reply says "Go." There is a distinct pause between the two calls. If there is a false start (*i.e.*, a crew starts before "Go") the competitors are recalled. A competitor refusing to start again or persistently starting early may be disqualified. A competitor leaving his proper course does so at his peril and may be disqualified if he thereby interferes with an opponent or if there is a foul. A foul occurs if, after a race has been started, a competitor comes into contact by his oar, scull, boat, or person with the oar, scull, boat, or person of an opponent. It is considered interference if a competitor by his conduct impedes the progress of an opponent who is on his proper course. In the event of a foul or interference, the umpire has power to allow the race to continue, to restart the competitors not disqualified or to order competitors not disqualified to re-row the race.

Occasions for a disqualification or foul

A competitor must abide by his own accidents, but if he is interfered with by some outside agency the umpire can re-start the race according to his discretion or order a re-row. A competitor receiving advice, assistance, or steering aid during a race may be disqualified. The whole distance must be completed by the full crew before a competitor can be held to have won a race. The umpire's jurisdiction extends over the whole race, from the time it is specified to start until it ends. The judge decides the order in which boats pass the winning post. The decisions of both these officials are final, and a competitor failing to abide by the decisions of the umpire or failing to follow his directions may be disqualified.

Technique. Special terminology is used to describe the racing stroke. The entry of the blade into the water is known as the catch. A stroke consists of all of the motions of an oarsman until his oar is again in the position from which he started. The recovery is the part of the stroke when the blade travels forward out of the water as the oarsman swings toward the stern. Feathering is turning the blade horizontally by a motion of the wrist as the oar handle is depressed to raise the blade clear of the water at the beginning of the recovery. The finish is the extraction of the blade after driving the boat through the water. Squaring is the turning of the blade from horizontal to vertical toward the end of the recovery in preparation for the catch. Balance is the skill of applying gentle pressure with legs and hands against the stretcher (footrest) and swivel pin to keep the boat running level throughout the recovery. If an oarsman feathers too soon, too much, or too little or if the boat lurches (loses balance), and his blade is caught in the water (dives), he catches a crab. Rate of striking (rating) is the number of complete strokes rowed per minute. Swift calculations are usually based on three strokes on a special timer that shows the equivalent rate per minute; *e.g.*, three strokes in five seconds is a rate of 36 strokes per minute.

Racing-stroke terminology

Strokes and styles. The orthodox "English Stroke" was based on early fixed-seat rowing at Eton, Oxford, and Cambridge. Body swing was the main source of power, and the arms were used chiefly as connecting rods to transfer the weight of the body to the oar. After the

introduction of the sliding seat (in the United States 1857, in England 1871), leg drive was added. The emphasis, however, was still on body swing, with the body swinging well past the vertical position at the finish. This required strong stomach and back muscles, which took time to develop.

Challenge to the orthodox stroke

The first challenge to this orthodox stroke came from an Australian, Steve Fairbairn, who entered Jesus College, Cambridge, in 1881. A member and coach of his crew, he upset tradition by winning races with a stroke that emphasized leg drive and arm pull. He considered smooth blade work more important than exaggerated body work, and his stroke sacrificed form for speed.

His method was adapted by foreign crews and led to innovations that are now commonplace. Slides were lengthened, swivel rowlocks replaced fixed pins, and crews were seated in a straight line instead of the staggered seating used originally in eights and fours.

English methods did not take root in the United States, where the first college coaches were professional scullers who adapted the sculling stroke to sweep rowing. In 1873 Yale introduced a stroke based on English principles that brought them considerable success for the next 20 years. Hiram Conibear at the University of Washington developed the "American stroke" in 1907. His system dominated United States college rowing for the next 30 years. It was not until 1953 that another coach made a significant impact on world rowing. He was Karl Adam, of Ratzeburg, West Germany. Like Conibear, he had never rowed or sculled before he started coaching. He was a boxer who cut across previous training methods by transposing methods based on *Fahrtspiel* ("speed play") and interval training (short sprints alternating with long paddles) from athletics to rowing.

Women's rowing. There are some records of women's rowing before 1914, but they did not take up the sport seriously until after World War I. In the early 1920s several clubs were established in England, and they sent crews to Australia, Belgium, France, and Poland to race. Women in more than a dozen other countries also took to the sport, and by the mid-1950s there was sufficient strength in Europe to justify the introduction of a championship. The Women's European Championships have been held annually since the inaugural event in Amsterdam in 1954. Racing takes place over 1,000 m (3,300 ft) in eights, coxed fours, quadruple sculls, double sculls, and single sculls. The U.S.S.R. has an outstanding championship record with numerous wins in all classes. Next best have been Germany and Hungary. Other nations that have competed regularly but with only occasional success are Austria, Bulgaria, Czechoslovakia, Denmark, France, Great Britain, The Netherlands, Poland, and Romania. (K.L.O.)

II. Canoeing

Canoeing is the manual propulsion of a lightweight boat, usually pointed at both ends, by a crew of one or more, facing forward and using paddles (not oars) that are not attached to the boat; canoe sailing is sailing such a craft under dismountable sails and spars. Several types of boats are covered by this definition, but only three are employed in international sports competition: the open canoe evolved from the birchbark of the North American Indian, with which single-bladed paddles are used; the decked canoe derived from the kayak of the Eskimo, with double-bladed paddles; and the sailing canoe, which may be an open canoe fitted with sailing rig and leeboards or a very specialized racing craft derived from the decked canoe. In the United States the term canoe is usually reserved for the open canoe, and the term kayak is used for decked, paddling canoes. In Great Britain, until recently, the term canoe was used for both open and decked craft, the name Canadian canoe being used to distinguish the open from the decked form. The term kayak for all decked, paddling craft is gaining currency.

Origins of the canoe and the kayak

HISTORY

Early canoes and kayaks. The canoe was the first true boat (excepting rafts and floats) to be built by primitive man. It exists in every part of the world, and excavations have unearthed specimens from the Stone Age. The canoes built by primitive peoples belong to one of three basic types—bark canoes, skin boats, and dugouts. Bark canoes are associated with the North American Indians who inhabited the forests where the white birch tree grew; but bark canoes are also found in eastern Siberia, South America, Africa, and Australia. An example of a skin boat is the kayak of the Eskimo, used from the Bering Sea to Greenland, but this class also includes hide boats of the Plains Indians and the coracles of Britain and India. Dugouts are made by hollowing out and shaping a tree trunk, and they have a worldwide distribution. Dugout canoes range in size from small, personal canoes to the huge war canoes of the coastal Indians of British Columbia and the Maori of New Zealand.

Development of modern canoes and of the sport. Native canoes were used for transport and hunting, and it is likely that the young men raced against each other, but it is accepted that the modern sport of canoeing was introduced to the world by a Scottish sportsman, traveller, and philanthropist, John MacGregor, in the years following 1865. In that year he designed a decked travelling canoe, which he named "Rob Roy," and in it he made a journey on the waterways of Europe. On his return he wrote a book, *A Thousand Miles in the Rob Roy Canoe on Rivers and Lakes of Europe* (1866), which became a best-seller, and he also lectured on his experiences. Between 1866 and 1869 he made further journeys by canoe in Scandinavia, Egypt, and Palestine and published accounts of them. His books and lectures aroused great interest among sportsmen and many took up this new sport.

"Rob Roy" decked canoes

The first Rob Roy had a length of 15 ft (4½ m) and a beam of 30 in (76 cm), and, with paddles, mast, and sail, it weighed 90 lb (41 kg). MacGregor's second and subsequent canoes were 14 ft (4 m) long, 26 in (66 cm) in beam, and weighed 72 lb (33 kg). The hull was of oak and the deck of cedar, and the paddler sat in a well or cockpit, his back supported by a backrest and his feet braced against a footrest. He used a double-bladed paddle based on the kind MacGregor had seen used with the bark canoes of Kamchatka, in Siberia. In order to be able to make use of favourable winds, a small mast and sail were carried, which could be stowed inside the canoe.

Rob Roy canoes achieved great popularity, and soon racing models were developed called Single Streaks, from their construction from two thin strakes or planks. Although made to suit the user, these racing Rob Roys averaged about 20 ft (6 m) in length and 22 in (56 cm) in beam. Within a few years a Rob Roy model for four paddlers, the ancestor of modern kayak fours, had appeared. In North America the development of racing craft led to wider, deeper, and shorter craft than in England, the paddlers sitting higher and using longer paddles.

Explorers, trappers, and timbermen had adopted the Indian open birchbark for travel through the northern wilderness of forest and lake, but in the course of time the superior tools and skills of the white men were used to improve on the bark canoe, while retaining its best design features. About 1870 the all-wood canoe was perfected and came to be known as the Canadian canoe, and a few years later, in the United States, the canvas-covered canoe was developed. Together these modern versions of the indigenous canoe spread through North America and to several countries of western Europe. In America they supplanted the Rob Roy both for cruising and competition, and in Britain they won favour for pleasure boating and cruising on inland waters.

Canadian open canoes

During the latter part of the 19th and early part of the 20th centuries steady improvement took place in the design of racing craft, but between 1920 and 1930 came rapid developments. From Scandinavia came new designs for racing and touring kayaks, and in Germany the invention of the *Faltboot,* or folding boat, made boat ownership practicable for young people who did not live near water and who lacked facilities for housing and transporting rigid craft. By 1930 the folding canoe had reached Britain and had brought about a revival of in-

terest in canoe touring, and this led to a renewal of activity in every branch of the sport. In North America, where the open canoe had retained a strong following, the advantages offered by the folding canoe were slower in being appreciated, but finally it was accepted. Following 1945, new materials became available, and with the materials new methods of construction and new design possibilities. Waterproof glues made possible cold and hot molding processes leading to strong, light, and resilient hulls. Aluminum canoes were built combining lightness and strength with durability; and many designs appeared for home construction of kayaks by assembling a wooden framework and covering it with a skin of waterproofed canvas. The material being used increasingly for all types of canoe, by both amateurs and professionals, is glass reinforced plastic (GRP), commonly known as glass fibre. This material is cheap, strong and resilient, weatherproof, easily repaired, and requires a minimum of maintenance.

Prototype sailing canoes

In 1869 a pioneer British canoeist, W. Baden Powell, carried out a journey in Sweden and the Baltic, using a Rob Roy canoe, "Nautilus I," and on his return designed a new canoe, "Nautilus II," to make a more effective use of sail. In 1870 he went on to design "Nautilus III," in which he gave complete priority to sailing qualities. "Nautilus III" was the prototype of a long line of sailing canoes in Britain and America, all classed as "Nautilus Type." From this time the development of canoes for paddling and for sailing followed divergent paths. As the years passed, improved models and rigs were produced, and about 1886, a leading American canoe sailor, Paul Butler, added the sliding outrigger seat, which allowed a helmsman to slide out to windward as shifting ballast.

Following many years of separate development of sailing canoes in Britain and America, in 1933 a "Restricted Class" rule was drawn up and agreed between the Royal Canoe Club and the New York Canoe Club, and the "International Canoe" was born. This class of sailing canoe, with a maximum sail area of ten square metres (108 square feet), was in 1946 adopted by the International Canoe Federation as the official ICF International Class, and in 1970 it became a one-design class (see *One-Design* under *Yachting*, below).

Development of organized canoeing. Soon after the return of MacGregor from his European journey the young men who had taken up the sport suggested that he form a club to foster it, and in 1866 the Canoe Club was formed "To improve canoes, promote canoeing and unite canoeists." The first members included distinguished sportsmen, travellers, mountaineers, and athletes, and the heir to the British throne, H.R.H. Edward, Prince of Wales, who became the first commodore. In an appendix to the first edition of his book *The Rob Roy on the Jordan* (1869), MacGregor referred to the club as having "200 members and not one drowned." In 1873 Queen Victoria granted the Canoe Club a charter as the Royal Canoe Club (RCC) and it grew rapidly and extended its influence widely.

In 1887 the British Canoe Association (BCA) was formed, as a cruising organization. Though it was never formally dissolved, it became moribund early in the 20th century. In 1933 a new British Canoe Association came into being and fused with the Camping Club, and in 1936 the British Canoe Union became the governing body for the sport in all its aspects.

In 1871 the New York Canoe Club was founded, being the first canoe club in North America, and in 1880 came the American Canoe Association (ACA) as the governing body of the sport in the United States. By 1886, MacGregor was able to record that the clubs of the ACA had a total membership of at least 2,000 members. An international authority to legislate for the sport was founded in 1924 under the name Internationale Repräsentationsschaft des Kanusport (IRK), and in 1946 a new organization was set up with the name International Canoe Federation, taking over all the functions and responsibilities of the IRK.

Canoeing as a recreation. Canoeing began as a non-competitive recreation, and even now, in most countries, the majority of canoeists use their canoes for recreation rather than competition. Recreational canoeing covers a wide and varied field. It includes paddling on local streams and lakes for enjoyment and relaxation; touring on canals, rivers, lakes, and sea, with the canoe used primarily as a form of transport; the use of the canoe in fishing, hunting, camping, and other recreations; and wilderness expeditions, an extension of touring into more adventurous fields demanding greater resourcefulness. Other recreational uses include wild-water, or white-water, sport, running the rapids of mountain torrents, an activity demanding a high degree of skill and with a considerable element of danger; and surf canoeing, another form of wild-water sport in the ocean surf.

White-water and surf canoeing

Many of these activities are included in the programs of youth organizations because they encourage the development of observation, initiative, judgment, self-reliance, and endurance, as well as great technical skill. In organizations that stress the value of service to the community by young people, such as Britain's Corps of Canoe Lifeguards, the members are trained to apply their skills as canoeists to beach patrol and rescue service, flood reconnaissance, expedition leadership, and voluntary coaching of canoeing.

COMPETITIVE CANOEING

Canoe classes and events. Competitions are organized at world, continental, national, divisional, state, and club levels. World and continental championships (the latter at junior and senior levels) are organized by national federations under the authority and supervision of the ICF. In competitive canoeing, the one-man Canadian canoe is designated C-1 and the two-man canoe as C-2; the one-man kayak is designated K-1, the two-man kayak as K-2, and the four-man kayak as K-4. Championships are conducted in three classes: flat-water racing, wild-water racing and slalom, and canoe sailing. All international competition is at metric distances.

Flat-water racing. World flat-water racing championships are held annually and may be combined with continental championships. Events for men include 500-metre, 1,000-metre, and 10,000-metre races for K-1, K-2, and K-4, C-1 and C-2, and the 4x500-metre relay in which each member of each four-man team paddles 500 metres for a total of 2,000 metres. The events for women are the 500-metre for K-1 and K-2.

Wild-water racing and slalom. World wild-water racing and slalom championships are held biennially. Men's events are for K-1, C-1, and C-2; women's for K-1; and mixed pairs' for C-2. There are individual and team events in each class, each team consisting of three boats.

Canoe sailing. Canoe sailing world championships are held triennially. The championship is for men sailing the International Class 10-square-metre decked sailing canoe.

Olympic Games. Every fourth year the canoe regatta of the Olympic Games replaces the World Championships for that year. Canoe events for men have been included in the Olympic Games since 1936, and for women since 1948. The Olympic events are racing, for men, 1,000 metres for K-1, K-2, and K-4, and C-1 and C-2; and for women, 500 metres for K-1 and K-2; and slalom, for men, K-1, C-1, and C-2; and for women, K-1.

International challenge trophies. The Paddling Challenge Cup of the Royal Canoe Club (Great Britain) was instituted in 1874 and is for annual competition over 10,000 metres in single kayaks (K-1). The International Challenge Cup of the New York Canoe Club was instituted in 1885 as a perpetual prize for the decked, sailing canoe.

National competitions. In the United States, Canada, Australia, and Great Britain, annual competitions are organized at all levels for the recognized ICF classes and courses, and, in addition, each country includes a certain number of national classes and courses in flat-water regattas, long-distance, and marathon races. In the United States the national classes include four-man single blade, a cruising class of sailing canoes, and classes for open canoes with racing rigs. In Great Britain

there is an emphasis on long-distance events, of which a classic is the race from Devizes to Westminster, 125 miles (200 kilometres) with 77 portages and raced nonstop.

Courses, equipment, and specifications. Regulations controlling the equipment, courses, and rules for the several branches of competitive canoeing are laid down by the ICF.

Flat-water racing. Courses for 500 metre and 1,000 metre are straight with provision for nine lanes, each nine metres wide, and water of uniform depth, if possible, of 3 metres. If 10,000-metre races are held on looped courses the radius of the turns is not less than 50 metres and the last 1,000 metres must be straight. In 500 metre and 1,000 metre no "hanging" (riding the wash of another craft and gaining an advantage therefrom) is permitted, though it is allowed in the first 9,000 metres of the 10,000-metre events. When hanging is not allowed, a five-metre rule forbids any competitor to come within five metres (16 feet) of another competitor in any direction. In the approach to a turn, a competitor being overlapped by another on the inside must leave room for the overlapping boat if the bow of that boat has reached a point even with the body of the competitor being overtaken, or with the forward edge of the cockpit, in the case of a kayak.

Long-distance racing. Long-distance racing is not controlled by ICF rules. Regulations vary from country to country. The course may be on river, estuary, or sea, with natural hazards found in such waters. Weirs and falls may be run, or "shot," or portaged. Mass starts are the custom, the canoes being lined up either on the water or on the bank. Classic events are the Sella Race in Spain and the Liffey Descent in Ireland.

Canoe slalom. The slalom event consists of the negotiation of a course of up to 800 metres (2,600 feet) on fast, turbulent water with natural and artificial hazards. The competitor must pass through 25 to 30 gates (pairs of poles hung from overhead wires); to touch the poles, or miss a gate, is to incur the award of penalty seconds, which are added to the recorded time, to arrive at the final result. To increase the difficulty of the course, some gates must be taken stern first, some against adverse currents, and a "team gate" must be negotiated by all three members of a team within 15 seconds. Judges at each gate signal a clean passage or the award of penalties. The craft used in slalom must be highly manoeuvrable, and the competitor needs to be able to carry out recovery from a capsize by means of his paddle.

Racing on fast and turbulent waters

By courtesy of American Whitewater Affiliation



Figure 2: Two-man wild-water canoe in competition (U.S. team at the 1971 World Championships at Merano, Italy).

Wild-water (or white-water) racing. Wild-water is graded in six categories according to difficulty, and wild-water racing at international or championship level is held on a course of from three to eight kilometres (two to five miles) of the highest (*i.e.*, most difficult) grade, with every form of natural obstacle. This form of racing demands the utmost physical effort, mental alertness, skill in canoe handling, and high personal courage. The Canadian canoe used in slalom and wild-water racing is now decked in for additional strength, round holes being left in the deck in which the paddlers kneel on both knees, wearing spray covers to prevent water entering the canoe. Being partially held in by straps they are able to right the capsized canoe by a paddle stroke, similar to the Eskimo roll stroke. Competitors must wear crash helmets and life jackets or other personal buoyancy.

Canoe sailing. International races are held in the 10-square-metre (having a sail area of 10 square metres) decked, sailing canoe IC, a one-design craft controlled by the ICF. The rules for canoe sailing are as for all other forms of competitive sailing, and similar courses are used. In the United States there are also events for the cruising canoe with a 40-square-foot lateen rig, leeboards, and steered by the paddle. Sweden and Germany also have competitions for their own national classes of decked canoes. (J.W.D.)

III. Yachting

Yachting is considered here as the sport of racing and cruising in sailing craft. All boats designed primarily for pleasure, no matter what their size, can be called yachts. The yacht is traced to Holland in the 16th and 17th centuries. The *jacht*, short for *jachtschiff*, a fast, light, sloop-rigged sailing vessel (also fast pirate ship, from hunting ship: *jacht, jagen*, hunt or chase), was developed there on the many canals, lakes, and estuaries. "Mary," a boat of the type, was presented by the Dutch to Charles II of England in 1660. Yachting as a sport, however, dates from the early 19th century. Its beginnings are traced below.

HISTORY

From their royal introduction into Great Britain at the Restoration until the coronation of Victoria, racing and pleasure sailing were esoteric pastimes rather than national sports; and though there was a Cork Water Club (now the Royal Cork Yacht Club) in 1720 and a Cumberland Fleet (of which the Royal Thames Yacht Club is the descendant) in 1775, yachting and boating as they are now known did not appear until the three decades following the establishment of the Yacht Club (now the Royal Yacht Squadron) at Cowes, Isle of Wight, in 1812. The sport was encouraged by the royal and official patronage naturally extended to it in a seafaring nation. During these decades there was energetic founding of yacht clubs at points on the coast where the local conditions suggested the pleasures of sailing; and a national institution was born.

Early English and American yachts. Among the earliest English yachts of which there is any record were the "Pearl," 95 tons, built in 1820 at Wivenhoe, Essex, and the "Arrow," 84 tons, built in 1822, and for 58 years one of the most successful cutters afloat; one of the largest, the "Alarm," was built in 1830 at Lymington, Hampshire, from the lines of a famous smuggler captured off the Isle of Wight. Some yachtsmen at this time preferred still bigger vessels and owned square-topsail schooners and craft resembling the contemporary naval brigs. Of such vessels were the "Waterwitch," built at Cowes in 1832, while Lord Yarborough's second "Falcon" was a full-rig ship of 351 tons, pierced for ten guns a side and with a crew under naval discipline. It was a quaint, buccaneering phase of yachting.

Although sailing for pleasure was a popular diversion in the American Colonies at an early time, the first boat known to be built exclusively for that purpose was the "Fancy" of New York (1717). The second was the 22-ton sloop "Jefferson," built in 1801 for a wealthy shipmaster, of Salem, Mass., who, in 1815, also commissioned the schooner "Cleopatra's Barge" and sailed her to the Mediterranean. The first large U.S. yacht, "Cleopatra's Barge" was 83 feet (25 metres) long on the waterline, had a 23-foot (7-metre) beam, and was furnished with great luxury.

Living in Hoboken, N.J., on the banks of the Hudson

The Stevens Brothers

River across from New York City, was the Stevens family, after whom Stevens Institute of Technology was named. John C. Stevens and his brother Edwin A. were the first prominent yachtsmen in the New York area. In 1809 they built the 20-foot (6-metre) sailboat "Diver," in 1816 the 56-foot (17-metre) "Trouble," and in 1820, "Double Trouble," believed to have been the first sailboat with twin, side-by-side hulls, or catamaran (see below *Multihulled boats*), built in the U.S. Their other boats included, in 1832 the 65-foot schooner "Wave," in 1839 the 91-foot (28-metre) waterline schooner "Onkahye," and in 1844 the 49-foot (15-metre) waterline schooner "Gimcrack." The latter was the yacht aboard which the New York Yacht Club (NYYC) was organized on June 30, 1844. Another brother, Robert L. Stevens, in 1846 designed the 88-foot (27-metre) waterline sloop "Maria," which incorporated certain features to be found in modern yachts, such as a hollow boom and crosscut sails. By this time there were a number of yachts racing on New York Harbor, and the archives of the NYYC show eight yachts enrolled in the fleet in 1844.

The yachts of that date tended to follow the models that had been developed for commercial purposes, and most of them were schooner-rigged, following the lead of the fast New York pilot schooners, although the "Maria" resembled more closely the Hudson River sloops that carried cargo on the Hudson River between New York City and Albany.

As in the United States, the English yachts that evolved from the working craft in the last days of sail embodied in their shape the fish form—wide, bluff bows, their greatest beam at about one-third of their length from forward, and long, streamlined afterbodies. A trend toward finer lines forward was reinforced by the success of the "America" in 1851, and thereafter the form of yachts changed. Bluff bows with an angle of entrance of about 30° were replaced by bows with an angle of less than 20°.

Beginnings of organized yachting. Although the Knickerbocker Boat Club of New York was founded in 1811, it was disbanded the following year, and organized yachting can be regarded as dating from the founding of the NYYC in 1844. Some of America's more important early yacht clubs and their dates of organization follow: Southern Yacht Club, New Orleans, (1849); Detroit Yacht Club (1865); Boston Yacht Club (1866); San Francisco Yacht Club (1869); Eastern Yacht Club, Marblehead, Mass. (1870); Seawanhaka-Corinthian Yacht Club, Oyster Bay, N.Y. (1871); Chicago Yacht Club (1875); and Larchmont Yacht Club (1880). There were about 1,500 active yacht clubs in the United States in the 1970s. Many of them were located on freshwater.

The Yacht Racing Association (YRA) was founded in England in 1875 with the object of providing a set of rules governing regatta sailing, and by 1881 most of the important clubs were members of the association. The Prince of Wales (afterward Edward VII) was its president. He was also commodore of the Royal Thames Yacht Club and of the Royal Yacht Squadron, and during these years the influence of the YRA was established. The Royal Ocean Racing Club (RORC) was founded as the Ocean Racing Club in 1925 with the principal objects of encouraging long-distance yacht racing and the building and navigation of sailing vessels in which speed and seaworthiness were combined.

Much of Australia's yachting history is spanned by the life of the Royal Sydney Yacht Squadron, which held its centenary in 1962, and in that year also first challenged for the America's Cup. The Australian yachting scene falls naturally into the groups of the six principal coastal states. As increasingly is the case in other parts of the world, the sport consists largely of small-boat racing, and also as in other countries, Australians follow the pattern of going to the sea for their leisure, fishing, and cruising, as well as racing.

Development of design and rigging. Before 1870 yachts were usually turned out by rule of thumb by local builders, each builder having his favourite type. Usually the hulls were "modelled": *i.e.*, a half model was whittled out before the frames were set up for the full-sized boat.

Hull design. The type of hull most in favour along the U.S. Atlantic seaboard was a broad-of-beam, shallow-draft, centreboard craft with great initial stability, carrying a large spread of canvas, and with a relatively small amount of ballast. Boats of this type were usually fast in smooth water but were not particularly good sea boats. The centreboard was housed in a "trunk" located on the centreline of the yacht and was lowered through a slot in the keel in order to give the boat sufficient lateral plane to work into the wind. Even quite large yachts were built on this model. One, 94 feet (29 metres) overall, 26½-foot (8-metre) beam, had a draft of only five feet two inches (157 centimetres) with her centreboard up.

Design of the "America"

At this time the so-called fish, or cod's head and mackerel tail, form was in favour. It is generally believed that George Steers of New York, designer of the yacht "America," introduced the narrow, pointed bow in designing the pilot schooners for which he was famous. The victory of 100-foot (30-metre), 170-ton "America" in a 53-mile (85-kilometre) race around the Isle of Wight against a large fleet of British yachts in 1851 was probably the greatest single stimulant to the early development of American yachting. Steers' boats caused a revolution in yacht design, but it was not until 1870 that the lines of the cutter "Vindex" were drawn on paper. It was the first American yacht to be built from a drawing rather than a model.

English yacht designers of high technical ability and rare artistic gifts designed yachts as a distinct branch of naval architecture. The evidence of contemporary scientific researches in ship design was applied to the shaping of yachts and, after 1874, led to a second fundamental change in design—the reduction of wetted surface. Incidental to this evolution was a return toward the fuller bows of the pre-"America" period.

In 1876 the yacht "Mohawk," a 140-foot (43-metre) schooner, capsized at anchor in a squall, drowning her owner and several guests. The disaster caused the advocates of beamy, shallow-draft centreboarders to take stock of type, and led to the design of the so-called compromise sloop "Mischief," a 67-foot (20-metre) iron centreboarder, with a slightly deeper hull than the old skimming dishes. She relied not only on beam but also on the placement of her ballast lower in the hull for stability. "Mischief" successfully defended the America's Cup in 1881 against the Canadian challenger "Atalanta," also a centreboarder.

Decline of the shallow-draft skimming-dish design

In that year a 46-foot (14-metre) cutter "Madge" arrived in the U.S. from Scotland. During her first summer she won most of her races boat-for-boat against the best centreboarders. Her phenomenal success greatly influenced the design of the compromise cutter "Puritan," which defended the America's Cup against the "Genesta" in 1885.

Influence of the America's Cup. The America's Cup contests exerted great influence in the matter of design. Initially, they brought together two opposed ideas: the beamy, shallow-draft, centreboard American type and the deep-keeled vessel of the British tradition. In 1885 the opposing types were well exemplified in the "Puritan" and the "Genesta"; but the four matches of the next decade witnessed a gradual merging of the two conceptions, until in 1895 the British challenger was actually the beamier of the two boats. Subsequently, the enormous sums spent on Cup yachts raised the science and art of design to a level of exceptional refinement.

The success of "Puritan" made her designer Edward Burgess the most popular U.S. designer, and the Burgess hull soon became the accepted U.S. standard. He subsequently designed two more successful America's Cup defenders, "Mayflower" and "Volunteer."

In 1891 Nathanael G. Herreshoff, of Bristol, R.I., turned out a yacht that was destined to revolutionize yacht design once more. This yacht was "Gloriana," with the yacht's forefoot boldy cut away so that her profile showed an easy sweep from the stemhead to the bottom of the keel.

When the next challenge for the America's Cup (1893) was received from the Royal Yacht Squadron, Herreshoff

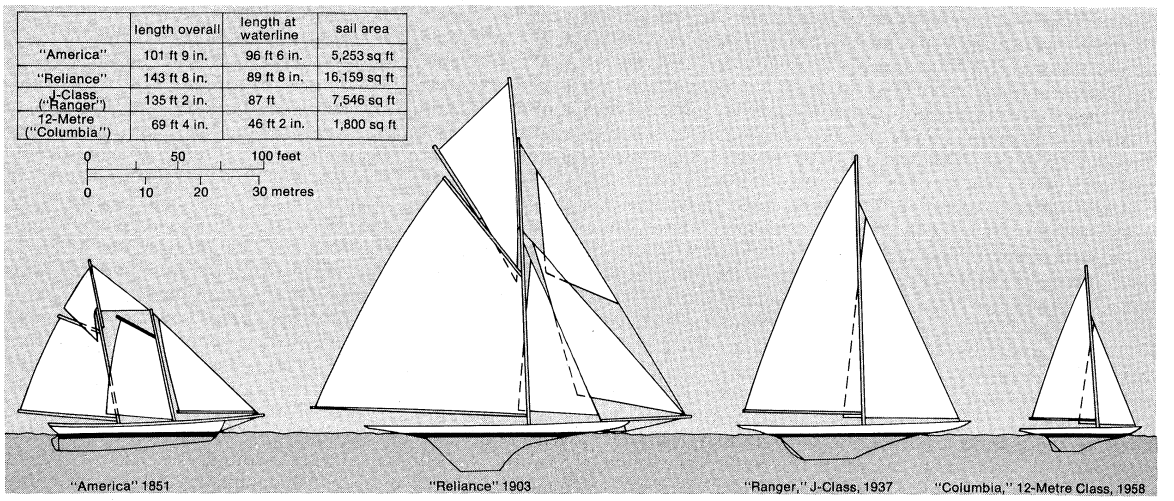


Figure 3: Evolution of America's Cup racing yacht design (from left): "America," original winner of the cup; "Reliance," largest defender; "Ranger," typical J-Class defender; "Columbia," typical 12-Metre Class yacht, the first 12-Metre defender.

From W.H. De Fontaine and E. Ratsey, *Yacht Sails*, copyright 1948, and © 1957 by Ernest A. Ratsey and W.H. De Fontaine

The Herreshoff designs

designed the successful defender, the sloop "Vigilant." While a keelboat, she also carried a centreboard that worked through a slot in the lead keel. After the "Vigilant," Herreshoff designed the successive defenders up to and including the 1920 contest, when "Resolute" successfully defended, and Herreshoff was recognized as the foremost yacht designer in America.

By 1930 when Sir Thomas Lipton made his last challenge with "Shamrock V," Herreshoff had retired and "Enterprise," the defender of that year, was designed by Burgess' son W. Starling Burgess.

The "Rainbow," defender in the 1934 race, was also from Burgess' drafting board. In 1937, he collaborated with a young designer, Olin Stephens, in the design of "Ranger." Stephens himself was to produce three more Cup defenders: "Columbia" (1958), "Constellation" (1964), and "Intrepid" (1967).

"Intrepid" brought to Cup racing the cutaway, underwater profile that had come to general popularity among ocean racers in the latter part of the 1960s, following the tremendous success enjoyed over an extended period of time by a number of stock fibreglass so-called Cal-40 class yachts built in California. The modern underbody configurations saved wetted surface and increased downwind steering ability by separating the rudder from the keel and placing it away aft, either as a free-standing balanced spade or with a skeg (a projection aft of the keel to support the rudder) at its leading edge. Small auxiliary rudders, popularly known as trim tabs, were sometimes retained on the trailing edge of fin keels in order to impart greater lift. "Intrepid" was fitted with a double-rudder configuration of this nature.

Rig design. Probably no one factor contributed more to the speed of yachts than the development of modern rigs. After 1920 the science of aerodynamics was applied by naval architects to the design of sail plans, resulting in a revolutionary change in yacht rigs. Tall, narrow sail plans, with a long, efficient leading edge, replaced the older gaff rig of roughly trapezoid-shaped mainsails with the base longer than the perpendicular.

The new mainsails were triangular (called Marconi, Bermudian, or jib-headed) and were set on tall masts with a ratio of hoist to foot (aspect ratio) of 2:1 up to 3½:1. In the case of the J-Class yachts (America's Cup class, 1930–37), masts were as high as 165 feet (50 metres) on a waterline length of 80 feet (24 metres). Large masthead fore-triangles and overlapping headsails followed as a natural consequence and tremendous jibsails, overlapping the mainsail by as much as two-thirds the length of the main boom and with a luff (forward edge) as tall as the mast, came into use. These were called Genoa jibs because the first one was used on a yacht racing off Genoa, Italy. So efficient were the new rigs that yachts carrying from 25 to 40 percent less sail area than formerly had as much as or more speed than the old, larger sail spread. The new rigs were also easier to handle and a smaller crew sufficed, while the elimination of the gaff and the topmast, to say nothing of the use of aluminum that became the dominant spar material in the 1960s, saved weight aloft.

Large, full spinnakers, cut much like a parachute, also replaced the old, smaller, flatter spinnakers and had greater pulling power.

Rating rules. Prior to the emergence of the one-design classes (see below), nearly all sailboats were individually designed and custom-built. To enable such boats to race each other, either on equal terms or on a handicap basis, rating classes were established. These were based on measurement formulas applied to length, sail area, and other factors affecting design and speed, the theory being that all boats with the same rating could compete fairly against each other. Each owner and designer, of course, attempted to create the fastest boat possible within limits set by the rules. Before 1905, for example, American boats were rated by adding the square root of the sail area to the waterline length and dividing by two. Under this rule, the same tendency became apparent as in England— to develop flat-bodied hulls with long overhangs and light displacement, producing an extreme skimming-dish type. This reached its limit in the America's Cup defender "Reliance" (1903), which had overhangs totalling more than 50 feet (15 metres) on a waterline length of 89 feet 8 inches (27 metres). Following this match a conference of yachting organizations in 1905 adopted the Universal Rule that, while retaining length and sail area as the chief factors, also used displacement in the formula and controlled overhangs, draft, freeboard, and other elements, by imposing penalties. It produced a more seaworthy vessel and, while amended several times, the America's Cup matches were sailed under it through 1937. The formula was 18 percent of the product of length times the square root of the sail area, divided by the cube root of displacement, equals the rating. The International Rule (adopted 1906) was considerably more complex but retained many of the same factors as the Universal Rule.

The Universal Rule

Ocean racing was conducted primarily under the Cruising Club of America (CCA) and Royal Ocean Racing Club measurement rules after the decade of the 1930s. This type of racing is on a handicap basis with time allowance derived from rating. The offshore racing yacht evolved during 1929–39 was the derivative of the International Rule craft, whose fundamental seaworthiness was proved in offshore racing. Encouraging the development were such American yachts as "Dorade," winner of the Fastnet Race (see below) in 1931 and 1933, and "Stormy

Weather," winner in 1935. The British yachts "Maid of Malham" and "Ortac" of 1936 epitomized the new type of seagoing racing yacht in which the speed of the International Rule boats was blended with the robustness of the earlier cruisers to produce a new and exceptionally versatile type of yacht.

In 1933 a formula was adopted by the RORC to approximate the immersed bulk of the boat by means of simple internal measurements capable of being made with the yacht afloat. It was considered essential that it should be possible to measure the boat without access to scale drawings or the necessity of hauling it out of the water. Apart from the basic formula, the rule comprised clauses by which bonuses or penalties in rating were derived in accordance with the yacht's characteristics.

As a result of the interest in the Bermuda Race, U.S. naval architects commenced to design to the measurement rule then in use, which was a simple one. But it soon became apparent that a more restrictive measurement rule was required to equalize the changes of divergent types and promote sound design. Accordingly, a scientific rule called the Cruising Club of America Rule, drawn up in 1940, became the basis for measurement for nearly all important long-distance races in the United States.

In the late 1960s, with ever-growing interest in international ocean racing, a demand arose for the RORC and CCA to merge their rules so that a yacht could be measured in home waters and race anywhere in the world under this rating. The most enduring features of the RORC and CCA rules were thus retained in the International Offshore Rating (IOR), which quickly became a world standard. (B.D.B.)

Types of boats. *Small boats and one-design classes.* One of the fastest growing areas in the field of sailing is that of one-design class boats. All boats in a one-design class are built to the same specifications in length, beam, sail area, and other elements. Racing between such boats can be held on an even basis with no handicapping necessary.

True one-design and development classes

In practice there are variations within the class-boat concept. These range from the true one-design to boats that are called development classes. Development classes race on a boat-for-boat basis but the class rules controlling them encourage individuals to experiment in specified fields within established parameters.

An example of the true one-design class would be the 13-foot 10-inch (4-metre) Sunfish, popular throughout the world and strictly controlled in all areas. Hulls, sails, and rig specifications are set, and even minor changes are not permitted without an official ruling by the governing body of the class. The International 14, on the other hand, is a development class that by its very nature allows continuous experimentation in hull design, construction techniques, sails, and rigging within the class rules. The 11-foot (3-metre) Moth and 32-foot 5.5-Metre are other examples of open or development classes.

In the middle area of these two extremes lies the majority of the one-design classes—boats that may vary to some extent in such areas as construction materials and cockpit layout (for example, many classes permit hull construction in both fibreglass and wood, and spars in aluminum or wood). The Snipe, Lightning, and Star are examples of this middle area.

The one-design concept is an old one, one that appears to have started in Ireland with a proposal for a 13-foot (4-metre) centreboard boat. The designer, in a circular issued in 1886, stated that such a boat would result in

> a race where every boat will have the same chance, and will call for throughout continued attention in order to gain every little advantage to be got in order to win, and not a mere procession of boats, and a race that will be a contest of the crew and not one of designers and sailmakers.

The boats were known as Water Wags. At just about the same time (1887), records indicate that one-design racing started in the United States in North Haven, Maine, when a class called North Haven dinghies became active.

Shortly after World War I, there was a boom in yacht building—particularly in the smaller classes. During that period numerous one-design classes, all small by comparison with prewar yachts, came to the fore. This was partly attributable to the great increase in postwar costs but also because interest in yachting was spreading and many owners preferred sailing their own boats to employing professionals. The first popular one-design class, the Star, was developed from the 17-foot (5-metre) Bug Class. In 1910 the design was modified, increasing the overall length to 22 feet 8½ inches (7 metres) with a sail area of 285 square feet (26 square metres), and the larger boat was christened "Star." The Star Class Association of America was formed in 1915, being replaced in 1923 by the International Star Class Yacht Racing Association. By the second half of the 20th century, the class was worldwide in scope and numbered more than 5,000 boats. Competition in the class is keen; it is one of the Olympic class boats, and international as well as national and regional contests are held annually. In the late 1920s, several International Rule classes also became popular, including the 6-, 8-, and 12-metre, the "sixes" being most numerous. A number of international races were held in this class in which British, Scottish, Swedish, Italian, and other yachtsmen participated. Formerly an Olympic class yacht, its expense resulted in its decline after the 1952 Games.

The popular Star and other one-design classes

Some Popular One-Design Classes

	length (feet and inches)
Aqua-Cat catamaran	12 2
Bluejay	13 6
Cal 20	20
El Toro	7 11
Ensign	22 6
Enterprise	13 3
Fireball	16 2
5-0-5	16 6
Flying Junior	13 3
420	13 9
Hobie Cat	13 11
Lightning	19
Mirror	10 10
OK Dinghy	13 1¼
Optimist pram	8
Penguin	11 5
Snipe	15 6
Sunfish	13 10
Thistle	17
Tornado catamaran	20

One-design classes range from those geared to the junior sailor (such as the 8-foot [2-metre] Optimist pram) to such high-performance boats as the two-man, trapeze-rigged 19-foot 10-inch (6-metre) Flying Dutchman, designed by Uffa van Essen (see below *Olympic boats*). Class organizations plan and schedule championships, maintain the class rules, and through a technical committee, maintain a close watch on new developments.

Scows. On certain waters of Canada, Barnegat Bay, N.J., and other sheltered coastal waters, and especially in Wisconsin, Illinois, Iowa, Missouri, Minnesota, and Indiana in the United States, the scow type, also known as the Inland Lake scow, has been highly developed. These boats are extremely flat, draw only a few inches, have two leeboards (bilge boards), instead of one centreboard, double rudders, and are sailed without fixed ballast. On a reach (at a right angle to the wind) they are exceedingly fast, having been clocked at speeds of more than 25 miles per hour (40 kilometres per hour). Contests for the Seawanhaka Cup, first offered in 1895 by the Seawanhaka-Corinthian Yacht Club of Oyster Bay, N.Y., for international competition among small yachts, did much toward the development of the scow. The Inland Lake Yachting Association (1897) governs an extensive scow racing and regatta schedule. There are recognized classes, ranging in size from the M, 16 feet (5 metres) overall, to the A, 38 feet (12 metres) overall. Class X, a conventional one-design, the only nonscow sponsored by the Inland Lake Yachting Association, is used as a junior trainer.

Multihulled boats. In the field of multihulled sailboats, the catamaran (two hulls) has made the greatest

Catamarans

inroads into the one-design and class fields. The concept of the multihull design is, of course, an old one, dating back about 2,500 years. Its acceptance for sports sailing on a large scale is of fairly recent origin, however. Numerous classes exist, with great appeal to those in search of high speed and shallow draft. The 12-foot 2-inch (4-metre) Aqua-Cat, 18-foot 9-inch (6-metre) Cougar, 14-foot and 16-foot (4- and 5-metre) Hobie Cat, 18-foot 9-inch (6-metre) Pacific Cat, 18-foot (5-metre) Phoenix, and 20-foot (6-metre) Shark are all popular classes, and the International Yacht Racing Union has discussed the 20-foot (6-metre) Tornado catamaran for potential Olympic status in 1976. Larger catamarans have engaged in ocean sailing. A similar type of boat employing three hulls is known as a trimaran.

Day sailers and cruisers. Day sailing and cruising purely for pleasure and recreation attract a far greater following than any other aspect of the sport. Day sailers take many forms. They are roomy, steady, practical, and enjoyable boats on which to spend a day or part of a day or evening sailing. Generally the term applies to any sailboat, not primarily a racing boat, that lacks accommodations for overnight living aboard.

The cruising sailboat, although small enough to be handled by a crew of one or two, or perhaps half a dozen, is equipped for living aboard, at least overnight. Many are capable of making long ocean voyages and even of circling the globe. Instead of the tall masts of the racer, the masts of the cruising sailboat are lower and stronger and sail area is less, to put less strain on hull and rigging. Hard-to-handle spinnakers are almost nonexistent. Rigging and construction are heavier. The boat is built to weather storms and withstand the weight of crushing waves. Rigs are "balanced" so that with the wheel lashed the boat can sail herself hour after hour while the crew sleeps, eats, or mends.

Yacht clubs and organizations. The controlling body of the sport of yachting is the International Yacht Racing Union (IYRU), founded in 1907. The organization is made up of national authorities from the various member nations, such as the Australian Yachting Federation, Fédération Royale Belge de Yachting, Federación Mexicana de Vela, the North American Yacht Racing Union, and others. The object of IYRU is to promote the sport of amateur yachting. Committees are the permanent committee, keelboat technical committee, centreboard boat technical committee, constitution committee, class policy and organization committee, and youth committee. The latter has established guidelines for an annual IYRU Youth Championship with entries from 15 to 19 years of age, held on different waters throughout the world. The permanent committee nominates the classes of boats that are sailed in Olympic Games competition.

The International Yacht Racing Union

Sailing and yacht clubs. With few exceptions, sailboat racing and all organized boating activity are conducted under the auspices of a vast network of chartered yacht clubs found clustered around boating centres all over the world. The oldest is the Royal Cork Yacht Club established in Ireland in 1720 as the Cork Water Club. Most of these clubs are affiliated with their own national or regional authority; *e.g.*, the Royal Yachting Association, for Great Britain, the North American Yacht Racing Union, for North America. While the primary purpose of most clubs is social, they also make an important contribution to boating as a whole by establishing competitive standards, promoting competitive events, and attending to the endless details of executing these.

Some clubs may have specialized objects, such as the Cruising Club of America or its British counterpart, the Royal Ocean Racing Club, both organizing offshore racing events and regulating the handicapping of ocean racers. There are also many clubs that cater exclusively to powerboats.

In addition to races for sailboats and predicted log races for powerboats, most yacht clubs hold regattas in which events are not limited to boats built for racing, and many conduct overnight, or longer, cruises.

The North American Yacht Racing Union. Regulations affecting yacht racing and measurement rules, up to about 1900, had been in the hands of individual yacht clubs or local yacht-racing associations composed of clubs in the same locality. There was thus lack of uniformity in different sections of the country. At the time of the agitation for a new measurement rule to supersede the length-and-sail-area rule, the New York Yacht Club called a conference of yachting organizations of the Atlantic coast and the Great Lakes to bring about uniformity by persuading the other sections to adopt the Universal Rule then being formulated. This was a step forward, but after the conference there still was no real governing body for yachting affairs in the United States until local yacht-racing associations and yacht clubs formed the North American Yacht Racing Union (NAYRU) in 1925. This was the first permanent legislative and governing body for sailing of national scope in the United States, and through it both the racing and measurement rules were standardized and an appeals board for racing members was formed.

In 1927 delegates from this union met with delegates of the International Yacht Racing Union in London to bring about closer international cooperation. Realizing the value to the sport of international racing, the union recognized (in addition to the Universal Rule) the International Rule used in European countries, with the result that several international classes were built and became popular in the United States. Meetings of the International Union's permanent committee are held annually, with NAYRU representatives participating.

NAYRU championship events

The NAYRU annually administers five North American sailing championships that bring top U.S. and Canadian sailors together after regional qualifying events are held. The five are the Mallory Cup (men), Adams Cup (women), Sears Cup (junior), O'Day Trophy (single-handed), and Prince of Wales Trophy (match racing).

Beginning in 1967, NAYRU became active in administering ocean racing. This rapidly growing part of the sport required national organization for numbering, equipment requirements, and the choosing of international teams.

The Royal Yachting Association. Immediately after World War II, the need for a national representative body for yachting in Great Britain, competent to deal with port, harbour, and other authorities, and with the government, became evident. The Yacht Racing Association, which was financially supported by the yacht clubs, was a basis for such an organization, and steps were taken to broaden its activities. A general purposes committee was founded to deal with yachting matters as a whole. These included such diverse questions as that of a projected purchase tax on yachts, the municipal rating of yachts' moorings, the siting of oil refineries and bombing ranges in unspoiled coastal areas, and the pollution of coastal and estuary water. To express the enlarged functions of the association its name was changed early in 1952 to that of the Yachting Association. Later in the same year the prefix Royal was granted. (E.M.Ho.)

THE SPORT OF SAILING

Courses. Sailboat races are held over two general types of courses: point-to-point and closed courses. Most ocean racing is point-to-point, as, for example, across the Atlantic, the Bermuda Race from Newport to Bermuda, and the Transpacific from California to Hawaii. Some ocean races are sailed to or around some objective (an island, a buoy, or other marker) and back; and some competitions combine both types of races.

Small-boat races usually are sailed on closed courses, most commonly triangular.

How a boat sails. The sails of a boat are airfoils that by their angle of incidence to the airstream (the wind) generate the drive that propels the boat. The trim of the sails is governed mainly by the angle of the desired course to the wind. When sailing to windward (Figures 4F and 4G), the boat's course is at the smallest practicable angle to the wind. The sails are then trimmed to an angle with the wind that produces their greatest driving force—an angle that is usually about 10° and varies with the efficiency of the sails. The resultant force produced by the sails may be further resolved into a component in the di-

rection of the desired course (drive) and another normal to it (leeway force). Under the conditions of Figures 4F and 4G, the drive force will be about one-third of that producing leeway. The latter has to be resisted by the boat, whose ability in this respect is one determinant of the smallest practicable angle of the course to the wind.

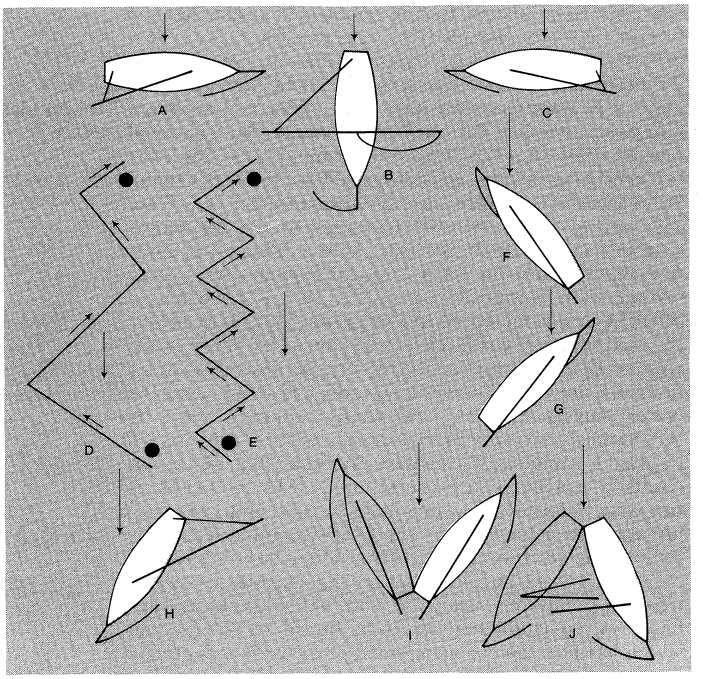


Figure 4: *Points of sail.*
(A) Reaching on port tack. (B) Running. (C) Reaching on starboard tack. (D, E) Course tacking to windward. (F) Close-hauled on starboard tack. (G) Close-hauled on port tack. (H) Sailing on a broad reach with a quartering wind. (I) Tacking. (J) Jibing. Arrows indicate direction of the wind.

Sailing into the wind

Very efficient racing craft may sail at about 35° to the wind; undistinguished performers may not be able to sail closer than 55° without suffering from excessive drift to leeward.

When the boat's course lies at right angles to the wind (Figures 4A and 4C) the sails are still trimmed at approximately the same angle of incidence to the airstream, but their angle to the course is bigger and accordingly the driving force is increased and that to leeward reduced. On this point of sailing the drive and leeway forces are approximately equal, and a boat is faster under the condition in Figures 4A and 4C, than those of Figures 4F and 4G. When the course is such as to bring the wind still further aft, the sails reach the aerodynamic condition known as stalled, and also the direction of the drive force swings with the sails closer to the direction of motion (Figures 4H and 4J) until the two coincide when the wind is astern as in Figure 4B. When the direction of the wind is abaft the beam and it is no longer possible to trim the sails to a uniform angle of incidence of about 10°, the most efficient sails are those with a full curvature—such as the spinnaker (Figure 4B)—which deflects the maximum volume of wind through the biggest possible angle. It will be evident that under the conditions of Figures 4F and 4G, when a boat is said to be close-hauled or on a tack, it is necessary to follow a zigzag course (Figures 4D and 4E) in order to reach a point dead to windward. When the bow of a boat is steered through the eye of the wind she is said to be tacking (Figure 4I); when the same evolution, of bringing the wind from one side to the other, is performed by steering so that the wind passes round the stern, she is said to be jibing or gybing (Figure 4J; see also SAILS AND SAILING VESSELS).

The America's Cup. This 100-guinea cup was offered in 1851 by the Royal Yacht Squadron for a race around the Isle of Wight, and was won by the "America." In 1857 the syndicate that had built "America" gave the cup (thereafter known as the America's Cup) to the New York Yacht Club as a perpetual challenge trophy to be raced for by yachts of foreign countries. The first challenge for the Cup was made in 1870 with "Magic" (U.S.) successfully defending against "Cambria" (U.K.) and starting an unparalleled winning streak.

The original deed of gift imposed disadvantages on challenging yachts, but a mutual consent clause ironed out several difficulties, and subsequent modifications of the original deed during the lifetime of the donors further improved the challengers' chances. Finally, in 1956, at the request of the New York Yacht Club, the New York State Supreme Court eliminated a clause requiring the challenger to sail on her own bottom to the scene of the contest. At the same time the permissible waterline length for single-masted vessels was reduced from 65 to 44 feet (20 to 13 metres), in order to include 12-Metre class boats, which are approximately ⅔ the length and 1/6 the displacement of J-class yachts previously used. These changes permitted a challenger to tune up in her home waters and be shipped for final prerace trials.

Changes in America's Cup rules

The record of America's Cup races is given in SPORTING RECORD in the *Ready Reference and Index*.

The America's Cup races enjoy the greatest public fame of all international yachting contests. They have on occasion, however, become battles of wits and complicated legislation as much as sailing, and a source of discords that at times were inflated into minor international incidents. Among such incidents were the defending club's early refusal to permit the challenging yacht to race against only one defender; and Lord Dunraven's allegations in 1895 that his "Valkyrie III" had been fouled by the U.S. yacht "Defender" and that the spectator fleet's crowding had endangered him. In the late 1890s, ruffled feelings were smothered by the genial tea magnate Sir Thomas Lipton, who was thereafter so much in the public eye with his "Shamrocks" and five challenges between 1899 and 1930 that the famous Cup was often thought of erroneously as the Lipton Cup. Subsequent contretemps have arisen from the New York Yacht Club's decision in 1934 that a protest could not be entertained because T.O.M. Sopwith's "Endeavour" had not promptly shown a protest flag; and "Gretel II's" disqualification in 1970 (second race of series) for a starting-line infraction after she had crossed the finish line first.

Many causes can be found for the unvarying failure of the challengers. So long as defenders were lightly built, the challengers were handicapped by the requirement of having to sail the ocean and the rugged construction the voyage necessitated. But after 1920 rules governing the construction of the yachts ensured that challenger and defender alike should be of comparable strength, and the most persistent cause of the challenger's failure was probably inferior organization and handling. The defenders were tuned to a high pitch racing against one or several defense candidates while challengers often never had a serious race until the Cup match started. For the first time in 1970, challenger eliminations (won 4–0 by Australia's "Gretel II" over "France") were permitted—further enhancing the possibility the sporting world's most remarkable winning streak might someday be broken.

Offshore sailing. *Transatlantic racing.* The first transatlantic yacht race (Sandy Hook, N.J., to England) took place among three American schooners in December 1866. Contestants were the "Henrietta," "Fleetwing," and "Vesta." The "Henrietta" won; her time, 13 days 21 hours 45 minutes. Transatlantic match races were held between "Cambria" and "Dauntless" from Daunt Rock, Ire., to Sandy Hook in 1870, and between "Coronet" and "Dauntless" from Sandy Hook to the Lizard, a prominent headland in the English Channel, in 1887. In 1905, 11 yachts raced from Sandy Hook to the Lizard. The winner, "Atlantic," a three-masted schooner, covered the 3,031 miles (4,877 kilometres) in 12 days 4 hours 1 minute. Her best day's run of 341 miles (549 kilometres) remained a record for yachts into the second half of the 20th century.

Again in 1928, under new ownership, "Atlantic" made a try for transatlantic honours, but was defeated by the

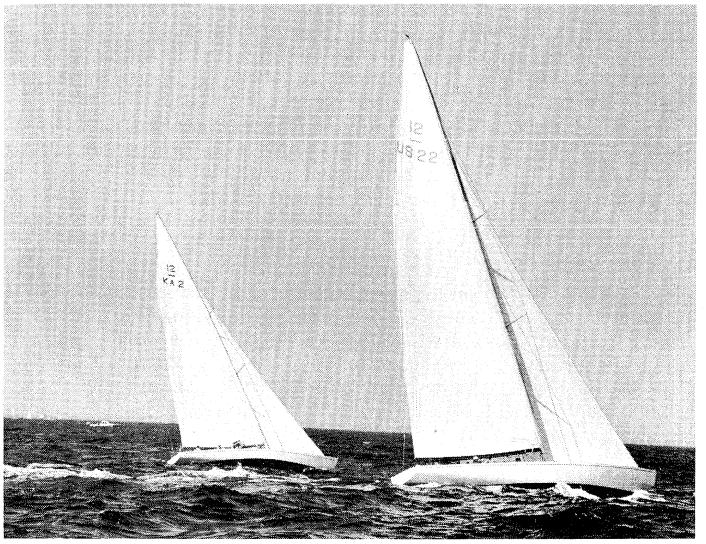


Figure 5: America's Cup competitors (left) "Dame Pattie" (Australia) and "Intrepid" (U.S.) close-hauled on starboard tack in the first race of the 1967 challenge series. "Intrepid" was the winner.
B. Devereux Barker III

Small craft in trans-atlantic racing

schooner "Elena," in a race to Santander, Spain, for the King of Spain's Cup. In a concurrent race for the Queen's Cup, which marked the debut of small craft in transatlantic racing, the 59-foot (18-metre) schooner "Niña" won from three competitors—one of which, the schooner "Rofa," sank without loss of life. "Niña" cruised from Spain to England where she became the first U.S. yacht to win the Fastnet Race. Other transatlantic races for small craft were held at irregular intervals after that time.

In 1931, "Dorade," a 52-foot (16-metre) yawl, built the previous year for her designer, was a winner by a remarkable 66 hours in a 10-boat fleet that raced from Newport, R.I., to Plymouth, Eng. Her elapsed time of 17 days 1 hour (32 years later the 57-foot "Ondine" sailed an identical course in 18 days 8 hours) proved for good that small yachts could be raced in safety across the Atlantic and started a trend away from the schooner rig, which had been most popular among ocean racers. In 1936, a 3,400-mile (5,500 kilometre) race from Bermuda to Cuxhaven, W. Ger., attracted one Dutch and seven German entries, but no American or British boats. The 59-foot (18-metre) "Roland von Bremen" was overall winner.

Rafael Posso, life commodore of the Havana Yacht Club, in 1951 promoted a 4,200-mile (6,800 kilometre) transatlantic race from Havana to San Sebastián, Spain. Sailed in June and July, the American ketch "Malabar XIII" won from three competitors. In 1955 the race from Havana to San Sebastián was again held and was won by the Spanish yawl "Mare Nostrum," from American, Cuban, and Argentine entries.

In 1960 the 47½-foot (14.5-metre) yawl "Figaro" won a 3,370-mile (5,240-kilometre) race from Bermuda to the Skaw light vessel in the Skagerrak, from an entry list of 17 yachts. The U.S., England, Sweden, and Germany were represented. Three years later the 57-foot (17-metre) aluminum yawl "Ondine" won from 13 other U.S. yachts a race from Newport to Plymouth. In 1966 a record transatlantic entry list of 42 yachts raced from Bermuda to the Skaw. "Ondine" won the top prize again.

In 1968 there was a race for 33 yachts from Bermuda to Travemünde, W. Germany, that was plagued by light air on the last leg across the North Sea to the finish. The winner was the 47-foot (14-metre) yawl "Indigo." In 1969 a race from Newport to Cork, Ire., in celebration (a year early) of the Royal Cork Yacht Club's 250th anniversary, attracted 23 entries over which the 73-foot (22-metre) yawl "Kialoa II" prevailed. Transatlantic races were scheduled to be sailed from Bermuda to Bayona, Spain, in 1972, from Newport to Plymouth in 1975.

The Bermuda and other ocean races. In 1906 Thomas Fleming Day, to demonstrate that small yachts, if properly designed, built, and handled, could go to sea with safety, promoted a race from New York to Bermuda in which there were three starters. The winner was "Tamerlane," a 38-foot (12-metre) yawl. After five years, interest in the Bermuda Race died out.

The Bermuda Race

In 1922 a group of cruising yachtsmen organized the Cruising Club of America. Among this group were several who felt that ocean racing would be beneficial to the development of both yachts and yachtsmen. Accordingly, a committee was formed (not under CCA auspices, however) to revive the race to Bermuda (660 nautical miles). There were 22 entries; the winner was "Malabar IV," designed, owned, and skippered by John G. Alden. Subsequent races were run thereafter in even years, except during World War II. The start of the 1932 race was shifted to Montauk Point, Long Island (distance 628 miles), and the 1936 and subsequent races were started from Newport (635 miles). In 1970 the length of the race was increased to 685 miles (1,100 kilometres) by making the Argus Bank Texas Tower, 25 miles southwest of Bermuda, a mark of the course. A record of 71 hours 35 minutes 43 seconds, established in 1932 by the cutter "Highland Light," was finally broken in 1956 by the Swedish-owned yawl "Bolero," with a time of 70 hours 11 minutes 37 seconds. Winner that year was Carleton Mitchell's keel-centreboard 38½-foot (12-metre) yawl "Finisterre," winner also in 1958 and 1960—an unparalleled achievement. The 1966 Bermuda Race was remarkable for the number of participants—167 yachts, of which 11 did not finish, primarily because of severe weather conditions near the end. All Bermuda Race finishes were handled by the Royal Bermuda Yacht Club, which has been a cosponsor since 1923. (For a complete list of winners see SPORTING RECORD in the *Ready Reference and Index*.)

Immediately prior to the forming of the RORC in 1925 a race had been sailed from Ryde, Isle of Wight, around the Fastnet rock (off the southwest coast of Ireland), and back to Plymouth, Devon, a distance of about 605 miles (975 kilometres). A race around the Fastnet was held annually between 1925 and 1931 and thenceforward every other year in alternation with the Bermuda Race.

Beken of Cowes

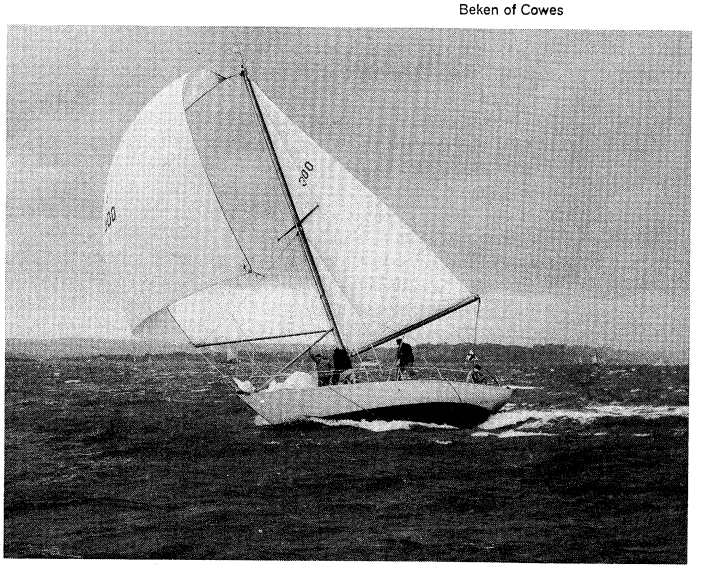


Figure 6: Racing sailboat running before the wind with spinnaker ("Clarion of Wight," 1963 winner of the Fastnet Cup, during the race from Plymouth, England, to Fastnet rock and return).

The Transpacific Race

The 2,225-mile (3,580-kilometre) Transpacific Race from San Pedro, Calif., to Diamond Head, Oahu, Hawaii (formerly sailed also from the California harbours of Santa Barbara, Balboa, Santa Monica, and San Francisco), ranks with the Bermuda Race in age, importance, and total number of races sailed. First held in 1906 and

repeated at irregular intervals, it was in 1928 embraced by the newly organized Transpacific Yacht Club and in 1939 was made a biennial to alternate with the Bermuda Race. There being no limitation on overall length (the largest yacht to sail the course was the 161-foot [49-metre] schooner "Goodwill") and the course being predominantly downwind, the race is notable for the good speeds sustained. Thus in 1949 the 98-foot (30-metre) schooner "Morning Star" covered the distance in 10 days 10 hours 13 minutes, and in 1955, when she had been rerigged as a ketch, reduced her previous time to 9 days 15 hours 5 minutes. In 1965 the 72-foot (22-metre) ketch "Ticonderoga" bettered the record by 1 hour 14 minutes. This record was bettered by the 73-foot ketch "Blackfin" in 1969. "Blackfin's" record is 9 days 10 hours 21 minutes. (For a complete list of Fastnet and Honolulu winners, see SPORTING RECORD in the *Ready Reference and Index*.)

Of major importance in the Southern Hemisphere is the 635-mile race (rated distance, 680 miles) from Sydney to Hobart, Tasmania, annually scheduled for December 26. Inaugurated in 1945 by the Cruising Yacht Club of Australia and the Royal Yacht Club of Tasmania, it was first won by John H. Illingworth in his cutter "Rani." By 1967 participation had increased to 67 entries. A three-time winner (1963–64–65) was "Freya."

Solo races across the Atlantic

A phase of ocean racing that grew at a remarkable pace in the 1960s was that of single-handing across the Atlantic Ocean. Three such races, which were to be called the world's loneliest and toughest sporting contests, were sailed westward the distance of nearly 3,000 miles (5,000 kilometres) between Plymouth, England and the United States. In 1960 there were five entrants, and Francis Chichester, who was later to be knighted for his single-handed passage around the world, won with "Gipsy Moth III" in 40 days. In 1964 Eric Tabarly of France sailed the course in 27 days 2 hours in "Pen Duick II," and in doing so beat 14 other entries, including Chichester, who shaved 10 days off his 1960 winning time in finishing second in the race. In 1968, 35 boats started the race and Geoffrey William of England won in the 57-foot (17-metre) ketch "Sir Thomas Lipton" in 25 days 18 hours 33 minutes.

The Admiral's Cup. In 1957 the Royal Ocean Racing Club established the Admiral's Cup as a perpetual award for biennial international team racing in yachts of 30–60-foot waterline length, later modified to 29–60-foot International Offshore Rating. Two short races and two ocean races, culminating in the Fastnet and totalling about 900 miles (1,450 kilometres), constitute a series. Great Britain retained the trophy in its first year, but lost to The Netherlands in 1959 and in 1961 to the United States. In 1963 the British team regained the increasingly important Cup from the United States and again won in 1965. Australia claimed the Cup in 1967; the United States in 1969.

The One Ton Cup. From 1907 to 1955 a premier international trophy, the One Ton Cup was put up in 1965 for challenge by the Cercle de la Voile de Paris for boats of up to 22-foot Royal Ocean Racing Club rating. Competing nations are allowed to enter up to three yachts each. The first three challenges in the 1960s consisted of three races, the first and third of 30 miles "round the buoys" and the second, counting double, of up to 300 miles (500 kilometres). Beginning in 1965 the format was changed so that one long race was sandwiched between four day races, a total of five. The first contest was won by "Diana III," of Denmark, from over 13 starters. In 1966 in Denmark, the 37-foot (11-metre) "Tina" won over 24 boats from nine countries. In 1967 another and quite similar design, "Optimist," from Germany, won at Le Havre, France, where the 1965 series also had been staged. "Optimist" repeated her victory in 1968 off Helgoland Island, West Germany. A New Zealand crew sailing "Rainbow II" won at Helgoland in 1969, and in early 1971 Australia's "Stormy Petrel" took top honours off Auckland. The concept of ocean racing without handicap appeared to have caught on permanently with deep-water sailors the world over. The most competitive of the One Ton Cup boats are designed especially for the series, but they also compete in other offshore racing events. Maximum rating in 1971 was 27.5 under the International Offshore Rating rule. (B.D.B.)

Small-boat racing. *Olympic Games.* The first yachting events in the history of the Olympic Games were scheduled for 1896 in Greece but were cancelled because of bad weather. It was 1900, therefore, at the next Games in Paris, when competitive sailing became an actual part of the Olympic program.

Olympic classes of boats

The classes of boats used in the Olympics have ranged from fleets of 8-, 6-, and 10-Metres and an over-10-metre category in the first games to the 5.5-Metre/Dragon/Star/Flying Dutchman/Finn slate prevalent from 1960 through 1968, the longest period without a change. The all-time high in the number of classes competing was at Antwerp, Belgium, in 1920 when there were 14. As of 1972, there were officially six classes competing.

Ellen Horan



Figure 7: Small, one-design boats, International Finn Class, just before the start of a race. Finns are an Olympic class.

Reasons behind the selection and/or dropping of a specific class have varied. At the 1912 Games in Stockholm, for example, the committee suggested that the 15-metre boats that raced in the 1908 Games be dropped because of the difficulty of having such a large vessel manned exclusively by amateurs. Today, classes for the Olympics are officially named by the International Olympic Committee (IOC) based on the recommendations and advice of the International Yacht Racing Union. The IYRU makes its selection of Olympic classes from what it has designated as Group A classes, a categorization that was established in 1965 and is defined as "modern, high-performance classes incorporating the latest design features and techniques." Included in Group A are the Dragon, Soling, Star, Tempest, 470, Flying Dutchman, 5-0-5, Finn, Fireball, Contender, Tornado catamaran, Australis catamaran, 5.5-Meter, and C-Class catamaran.

This list is added to from time to time. The Group B designation is given to "classes which are not considered suitable for Group A and classes which have been granted international status because of popular support, or because they provide a good training for yacht racing." These include the 12-Metre, 6-Metre, 8-Metre cruiser/racer, International 14, 12-square metre Sharpie, Lightning, Snipe, Vaurien, Enterprise, and Cadet. At the IYRU meeting in 1967, it was decided that at least one class in each Olympiad should be replaced. The committee stated, however, that this would not necessarily be the class that had been in the Games the longest, nor would it mean that the dropped class could not be reinstated later to Olympic status.

At the Olympic Games in 1972, when the sailing events were held at Kiel, Germany, the classes were the 29-foot 2 inch (9-metre) Dragon (an Olympic class since 1948), the 26-foot 9 inch (8-metre) Soling (a new Olympic class replacing the 5.5-Metre, which had been in the Games since 1960), the 22-foot 7½ inch (7-metre) Star (an Olympic class since 1932), the 19-foot 10 inch (6-metre)

Flying Dutchman (an Olympic class since 1960), the 14-foot 9 inch (4.5-metre) Finn (an Olympic class since 1952), and the 21-foot 11¾ inch (6.7-metre) Tempest (in the Olympics for the first time in 1972 as a result of the IYRU's request to the IOC for a sixth class, granted in 1969). For the Games in 1976, the IYRU issued a "Declaration of Intent" stating that the Olympic classes should be modern, high-performance boats and should consist of two keelboats, three centreboard boats, and the Tornado catamaran.

To qualify for an Olympic berth, the sailor must expose himself continuously to competition on an international level. Generally speaking, to select the sailing team for the Olympics every four years, the larger countries hold trials in each of the Olympic classes (some countries send partial teams if they are not particularly strong in a specific class). Actual selection practices vary. The German Yachting Association in determining its team for the 1968 Games at Acapulco, Mexico, for example, specified that those wishing to participate in the final trials would have to have placed in the top 25 percent of at least three of some six major European regattas during the previous year. In other countries, where fleets in the final trials might prove unwieldy in size, an area or district qualifying regatta is held to establish who may participate in the final trials. Most countries try to hold their final trials in an area that resembles the actual Olympic site in wind strength and sailing conditions. At the Olympics themselves, the IYRU is responsible for all the technical details of organizing the regatta.

Olympic scoring

Under the rules for an Olympic regatta, there are seven races held for each class and scoring is based on the best six of these. The Olympic scoring system is 0 points for first place, 3 for second, 5.7 for third, 8 for fourth, 10 for fifth, 11.7 for sixth, and the finishing place plus 6 points for seventh place and thereafter. The lowest total score wins. (The complete rules on the Olympic Scoring System are printed in the *International Yacht Racing Union 1971 Year Book*.)

The influence of the Olympics is widespread. With its emphasis on high performance, it has spawned a type of regatta in many parts of the world that adheres closely to the Olympic format. One such regatta is the Canadian Olympic Training Regatta at Kingston, Ontario, which was first held in 1969. Open to the Olympic classes as well as a number of Olympic-type classes, it was started with the intention of exposing Canadian sailors to the best competition possible with the result of fielding a strong Olympic team. (For Olympic Games winners, see ATHLETIC GAMES AND CONTESTS: *Olympic record*.)

Yachting events are included in the quadrenniel Pan-American Games, which started in 1951. Classes involved are the Olympic classes and the Snipe and Lightning, though not all of these classes have participated each time. The class organizations within each country are responsible for holding elimination trials to determine the individual representatives.

One-design class championships

Other small-boat championships. Depending on their size and popularity, one-design classes set up annual international and world championships in addition to national and district events. Fleets are the backbone of every class organization and, where the class involved is a large one, certain qualifying requirements at either the fleet or district level are necessary to establish which skippers may sail in the major championships. The Snipe class, for example, works on both the fleet and district level for entry to the U.S. National Championship, going on the basis of one boat eligible for every five in the fleet. The top three boats from the district championships may also attend the nationals, as well as the previous year's champion. Entries must have participated in at least five season races at the fleet level. The Snipe class holds its World Championship every two years, open to one sailor from each member country (usually the national champion) and to the previous world titleholder.

Smaller classes may not have any prerequisites for a class member to participate in national title events but often, if the fleet at a championship is too large, it will be divided up into sections or flights that then proceed to race against one another to determine the champion. A number of classes also hold junior championships in conjunction with their national events. Classes vary the locale of their championship events in a number of ways, usually alternating among the various districts it has set up within a country. Other classes permit the national champion to set the site for the following year's national championship. All of these decisions are regulated by the individual class bylaws.

In the United States, a number of the scow classes hold what they refer to as Blue Chip regattas in addition to their major national title events. Participation is limited to approximately 25 top boats, selected by a committee on the basis of their standings throughout the year.

In areas where a particular one-design class may not be well enough established to race as a fleet, or in specific regattas that pit one class against another in order to measure relative performance, methods of handicapping are set up. One method that is used throughout the world is the Portsmouth Yardstick System, devised in 1949 in England. The purpose of it is to rate boats so that they may race on an even basis, and it revolves around establishing a Portsmouth Number for each class. This number represents the time it takes for specific classes to sail over a common distance, and the numbers, administered by the Royal Yachting Association, are studied each year and revised as necessary. Handicapping is also done by using such systems as the *Yachting* One-of-a-Kind formula, a measurement system of rating one-design classes based on length and sail area.

In addition to championships held within the format of the class organizations, yacht clubs also hold regattas that may be open to numerous classes. The "race week" format is an established one with such large events as the Marblehead (Mass.) Race Week, annually attracting local sailors for the competition. Fixtures such as the Bermuda Invitational International Race Week and Germany's Kiel Week appeal not only to the local sailors but continue to attract more and more participants from other countries as they become synonymous with top competition.

Multihull boat races

One of the most interesting areas of competition in multihulls is the match-race series for the International Catamaran Challenge Trophy, known unofficially as the Little America's Cup. The event was started in 1961, and the trophy was donated by Sea Cliff (N.Y.) Yacht Club. Boats used are the C-Class catamaran, a highly refined racing machine that, because of the simplicity of the class rules, offers an exceptional scientific and technological challenge to the designer, as well as to the crew of two that races the boats. The maximum length of the C-Class catamaran is 25 feet (8 metres), maximum extreme beam is 14 feet (4 metres), and maximum sail area including the spar is 300 square feet (28 square metres).

The competition is truly international in flavour, with the governing rules stating that the event should move around the world irrespective of winner. The title is awarded after a best-of-seven series, with the first boat to win four races declared the winner. Though there are relatively few C-Class catamarans, the class commands a considerable amount of interest and also holds a World Championship. C-Class catamarans have been clocked at 30 miles (48 kilometres) per hour. (E.M.Ho.)

IV. Motorboating

Motorboating is the operation of a relatively small vessel driven by an internal-combustion engine or engines. As a sport it is generally practiced in competition in the form of racing or piloting and navigation contests; for travel by water (cruising) or the pleasure and freedom derived from being afloat; or for the enjoyment of related sports such as fishing, hunting, swimming, skin diving, and water skiing.

SPEEDBOATS

History. The internal-combustion engine was applied to boat propulsion in the late 1880s and the 1890s, but it was the internal-combustion engine as developed for the automobile that started motorboating as a sport

Beginning of organized races

when, about 1900, some French automobile enthusiasts adapted the engines to small boats and the autoboat was created. The idea quickly spread to England and the United States and organized races began to be held. In 1903, Sir Alfred Harmsworth (later Viscount Northcliffe) established the British International Trophy for motorboat competition (later known as the Harmsworth Trophy) and the first international, perpetual trophy in the sport was won that year by S.F. Edge's "Napier I," with a speed of 19.53 miles per hour (31.42 kilometres per hour) on a course at Queenstown (Cobh), Ireland. The same year, the American Power Boat Association was formed and in 1904 the Columbia Yacht Club of New York presented the challenge cup (Gold Cup) to the association "for the purpose of promoting speed contests and improving . . . engines for, and the lines of, power boats." The cup was won that year by C.C. Riotte's "Standard" at 23.6 miles per hour (38.0 kilometres per hour). Also in 1904, the magazine *Motor Boat* was introduced by the Temple Press, London, and a magazine of the same name became the first United States periodical to be devoted to the new sport. In 1905, the first U.S. annual National Motor Boat Show was held in New York City and became an international marketplace for motorboats that still flourished in the 1970s.

Speedboat design. The hulls of boats are classed as planing types, which skim across the surface, and displacement types, which push through the water. All speedboats of the early 1900s were displacement types and attempts to improve performance involved making the shape better suited to move through the water or installing more horsepower. Because the displacement hull, with its V-shaped or rounded bottom, relatively deep draft and narrow width for its length, sharp bow, and narrow stern, can never plane on the surface no matter how much power is applied, the efforts of designers were directed toward the development of hulls with flat surfaces that at speed would rise to the surface and skim across the water, thus reducing the wetted surface and thereby the friction and resistance between hull and water. William Froude of England originally advanced the idea of the three-point hydroplane in the 1870s and a plan for a planing-type hull was submitted to the British Admiralty in the same decade, but no engine with low enough weight per horsepower was available, so that the idea of the hydroplane languished until the autoboat had been developed. Just who applied the idea first is unknown, but simple, rectangular flat-bottomed skimmers were being raced in Europe in 1907. Early hydroplanes were stepped to reduce wetted surface, some with a single step and some with multiple steps, but they were merely a series of single, flat planes. In 1908, William H. Fauber of Chicago and Nanterre, France, saw the possibility of improving the box-section, stepped skimmers by combining the step principle with the V-section bottom and he patented an idea for an eight-step, V-bottom hull in England. About the same time, single-screw, round-bottom displacement hulls were giving way to twin-screw, V-bottom types. In the 1910 Harmsworth Trophy race, a multiple-step Fauber-Saunders British hydroplane named "Pioneer" sped past the speedy "Dixie III," an American V-bottom displacement boat. The fact that "Pioneer" broke down and "Dixie III" won the race at 36.04 miles per hour (57.99 kilometres per hour) did not change the picture for observers. In 1911, an Italian named Enrico Forlanini built the first hydrofoil boat, and in 1912, at least five builders were offering stock boats of the planing type.

Speeds. From the early competitive speeds of 20 miles per hour (32 kilometres per hour) in 1904, race boats developed to the point where unlimited propeller-driven hydroplanes were running heats at well over 100 miles per hour (160 kilometres per hour) in 1955. The huge craft would do over 150 miles per hour (240 kilometres per hour) on the straightaways, but had to slow down for the turns at each end of the course. Sir Malcolm Campbell held the one-mile water speed record of 141.74 miles per hour (228.6 kilometres per hour) from 1939 to 1950, when the revolutionary three-point hydroplane "Slo-Mo-Shun IV" took the record with an average of 160.323 miles per hour (257.960 kilometres per hour) on Lake Washington at Seattle, Washington, in June 1950. Two years later, the same boat increased its speed to 178.497 miles per hour (287.202 kilometres per hour). Donald Malcolm Campbell, son of Sir Malcolm, on July 23, 1955, became the first man successfully to pilot a jet-propelled boat over an official time course. His mark was 202.32 miles per hour (325.53 kilometres per hour). At least two men, John Cobb of England and Mario Verga of Italy, were killed in the 1950s trying to reach 200 miles per hour (320 kilometres per hour). The younger Campbell later boosted his record to 239.07 miles per hour (384.66 kilometres per hour) on Coniston Water in England on November 7, 1957, and on the same water on May 14, 1959, he was officially timed at 260.35 miles per hour (418.90 kilometres per hour). On June 30, 1967, "Hustler," with a Westinghouse jet engine, raised the speed to 285.213 miles per hour (458.901 kilometres per hour) at Guntersville, Alabama.

Water speed records

While inboard speeds were approaching 200 miles per hour (320 kilometres per hour) the world's outboard pilots were making a concerted effort to pass 100 miles per hour (160 kilometres per hour) on a mile straightaway run. Finally, Massimo Leto di Priolo of Italy hit 100.36 statute miles per hour (161.48 kilometres per hour) in a Molinari Class X hydroplane with a special Lesco engine in December 1954. Bert Ross, Jr., owner and driver of a Class X Mercury-powered Jones hull, raised that to 115.547 miles per hour (185.915 kilometres per hour) at Seattle, Washington, on May 4, 1960. In inboards, "Miss U.S. I" was clocked at 200.419 miles per hour (322.53 kilometres per hour), at Guntersville, Alabama, on April 17, 1962. It was the first non-jet-propelled craft to surpass 200 miles per hour.

Racing. The major divisions in motorboat competition are stock inboard, stock outboard, inboard hydroplane, and outboard hydroplane. Each division has a number of classes, depending mostly on engine displacement. Some of the smaller craft are barely big enough to hold the driver; the largest, the unlimited hydroplanes, weigh close to four tons. Many hundreds of races and regattas are held each year in the United States under the sanction of the American Power Boat Association, governing body for the sport in that country. Most races are held over closed courses. Some, like the Mississippi Marathon or the Six Heures de Paris, are endurance events. In a separate division, pleasure boats compete in marathons of 50 to more than 250 miles. In 1922, the Union of International Motorboating (UIM) was organized, largely through the work of John Ward of Ireland, who was its first secretary. The idea was to provide a clearinghouse for world records for the various classes racing in Europe, but it spread to encompass the whole world. The voting members grew to more than 40 nations, each represented by a national authority, usually the largest federation of power-boating clubs in the nation. Expansion of UIM activities has produced a commission for technical matters, one for sports matters, one for offshore racing, and a tourist commission that governs pleasure boating in Europe and has greatly simplified passage between countries. The UIM awards a world championship by allowing points for the first six finishers in selected offshore powerboat races in different countries. The 1970 winner was Vincenzo Balestrieri of Rome, who won for the second time in three years by taking four world events: the Wills International in England, the Miami-Nassau Race from the U.S. to the Bahamas, and two Italian events, the Naples Trophy and the Viareggio. Previous winners were James R. Wynne (1964 and 1966), Richard Bertram (1965), and Donald Aronow (1967 and 1969).

The Union of International Motorboating

CABIN CRUISERS

The earliest cruisers developed from long, narrow rowing or sailing vessels of 16 to 20 feet (5 to 6 metres) that had been popular in the 1890s. These were adapted to take engines and became launches, and the first step toward the cruiser was the addition of a removable awning.



Figure 8: Outboard hydroplanes skimming over the water during a race at Southbury, Connecticut, 1970.
H.W. Magnuson

Next came a light, wooden roof supported by stanchions screwed to the coaming. Side curtains followed, then glass windows. Eventually, cabin launches from 30 to 80 feet (9 to 24 metres) were being built. The hulls were graceful but the windowed, trolly-like superstructures were ungainly. The first departure from this type came with the substitution of a low, forward cabin similar to those common on sailboats. Fitted with small windows or deadlights, it was called a hunting cabin to appeal to duck hunters. The low cabin, with narrow beam and sitting or minimum standing headroom, persisted right up to World War II, and the hulls remained displacement types. After the war, small outboard cruisers began to appear, with planing hulls basically the same as the outboard runabouts that had become popular. Small inboard cruisers followed, also with planing hulls, and a boating boom in small craft that peaked at the end of the 1950s lent further impetus to cabin-cruiser design, for many people became ready to move up from the runabout hulls they had purchased. Beam, depth, and headroom increased in small cruisers and maximum accommodation in minimum length became the byword of builders. There has been an upward trend in the length of cruisers, for family boating has become a common source of recreation and more room afloat is required. Keeping pace, boatyards have become marinas that provide dockage, fuel, water, electricity, and such shoreside facilities as lounges, restaurants, entertainment, laundry service, and even swimming pools.

Changes in cruiser design

The U.S. Power Squadrons, for owners and others interested in cabin cruisers and powerboats, was organized in 1914 as something of a small-boat naval reserve. With about 80,000 members in about 400 units in 30 districts in the United States as well as Yokohama, Japan; Okinawa; the Panama Canal Zone; and Puerto Rico, it is the most highly organized boating force in the world. A notable feature of its activities is the provision of instructional classes covering all aspects of boating, including pilotage and navigation. There are many clubs in Great Britain that offer instruction, notably the Little Ship Club and the Cruising Association of London, and the Clyde Cruising Club with headquarters in Glasgow, although none of these confines its membership to either powerboat or sailing men.

Many yacht clubs organize competitive powerboat events. Called predicted log races, navigational skill rather than speed is the basis for scoring. In these, the skipper of a boat predicts the exact time he will pass specified points around a predetermined course. Not allowed a watch, he must adjust the speed of his boat to such variables as wind and tide and current so that the time requirement is met. The skipper coming closest to his prediction wins.

OTHER MOTORBOATS

Motor sailers. The motor sailer is what the name implies—a boat that can be driven under power or sail or both. The engine is larger than that found in an auxiliary sailboat (where the power is used primarily for getting in and out of harbour) and the sails are less efficient than they are in a cruising or racing sailboat. However, the lesser sail efficiency is a result of bluff ends and heavier construction with roomier accommodations below. The motor sailer still has a smooth design that enables it to make a steady eight to ten knots under power, with the sails providing a steadying effect, and, if necessary it can always make port under sail alone.

Houseboats. The houseboat in its simplest form is a floating home, permanently moored in a sheltered location, with moving possible only by towing. Such a craft has a barge-type hull, broad-beamed and flat-bottomed, with maximum living area and minimum seaworthiness. In the 1950s, small houseboats with barge-type hulls and outboard motor brackets began to achieve some popularity. Other houseboats with pontoon-type hulls bridged by a flat deck were less satisfactory from the standpoint of load-carrying capacity and stability. As motor horsepower increased and controls improved, hull refinements to improve performance included curving the bow up and shaping it more to a point and changing the flat bottom to a shallow V. The advent of the stern drive, inboard-outboard motor in 1959, followed by models of ever-increasing horsepower in the 1960s, lent the final impetus. Within the decade, 40-foot (12-metre) houseboats were travelling 30 miles per hour (50 kilometres per hour) in coastal waters. High-speed houseboats, which have achieved the most popularity, range from 22-foot (7-metre) camper types to complete luxury homes of 60 feet (18 metres) or more. All have the ability to operate in relatively shallow water and most can travel at speed in three-foot (one-metre) seas. However, they are not considered to be offshore cruisers, for their shallow draft, high sides, and large window expanse make them highly vulnerable in heavy-weather conditions.

Utility boats. The utility boat is most often thought of as an unpretentious motorboat from 15 to 23 feet (5 to 7 metres) with a forward deck, a windshield, forward helmsman and companion seats, and a large, open cockpit that can serve a multitude of purposes. Power may be outboard, stern drive, or inboard, and the design is such that care and maintenance are simplified. These craft appeal to fishermen and to families, different members of which may enjoy fishing, water skiing, day cruising, picnicking afloat, swimming, or skin diving, either separately or in groups. The utility may have a removable shelter top or a hardtop over the forward area of the cockpit. In either case, side and after curtains may fully

enclose the boat and air mattresses may be used in the cockpit for roughing it on overnight or weekend trips.

Other variations. Other variations of the motorboat include sport fishermen, which are utilities or offshore cruisers with special emphasis on cockpits rigged for fishing; pontoon deckboats, which have hulls consisting of pontoons or cylinders bridged by a flat deck surrounded by a railing and topped by a pipe-supported canopy and are used for swimming, picnicking or day cruising; inflatable craft, which serve as outboard-powered dinghies with the addition of a bracket or as high-speed sport boats with the addition of rigid flooring and a strong transom to take an outboard of higher horsepower. (J.Sm.)

BIBLIOGRAPHY

Rowing: G.C. BOURNE, *A Text-Book of Oarsmanship* (1925), describes the theory and art of rowing, oar and boat design, coaching, and muscular action. DESMOND HILL, *Instructions in Rowing* (1963), is a comprehensive guide for coaching beginners upward. J.A.N. RAILTON, *International Rowing* (1969), analyzes the factors behind the international success of several countries. HYLTON R. CLEAVER, *A History of Rowing* (1957), is an historical review of rowing in England from the earliest days of the professional watermen on the Thames. R.F. KELLEY, *American Rowing* (1932), traces the development of rowing in the United States. R.D. BURNELL, *Sculling for Rowing* (1968), examines the use of sculling to train for rowing; his *Oxford and Cambridge Boat Race 1829–1953* (1954), tells the story of the first 99 races, and the *Henley Regatta* (1957), is a comprehensive record of the development of the event since its inception in 1839.

Canoeing: PETER DWIGHT WHITNEY, *White-Water Sport* (1960), on the art and science of canoeing on white water (well illustrated with diagrams and photographs), the book gives valuable guidance to all taking up this sport both as a recreation and in the field of slalom and white-water racing. ALAN BYDE, *Living Canoeing* (1969), is a book on the sport of canoeing with special emphasis on advanced techniques for white-water sport on river and surf; includes clear instructions well supported by diagrams. AMERICAN NATIONAL RED CROSS, *Canoeing* (1956), is an interesting, well-presented textbook for all involved in canoeing instruction (exceptionally sound on the handling of the Canadian canoe). CHARLES SUTHERLAND, *Modern Canoeing* (1964), describes in detail the many types of canoes and how to handle them; it examines canoeing possibilities in Britain, including general touring, white-water sport, and sea canoeing, and gives guidance on all aspects of competitive canoeing. PIERRE PULLING, *Principles of Canoeing* (1954), written by a professional woodsman and guide, provides sensible advice for both students and teachers. ISTVAN GRANEK, *Paddling Kayaks and Canoes*, trans. from the Hungarian (n.d.), is the authoritative coaching textbook on the subject of canoe racing, by the leading European coach. JAMES HORNELL, *Water Transport* (1946), is an erudite and comprehensive account of the origins and early development of the craft used by men on river, lake, and sea. BRITISH CANOE UNION, *Guide to the Waterways of the British Isles*, rev. ed. (1970), is the only complete guide to the inland and coastal canoeing waters of the British Isles.

Yachting: There are a number of books geared to the interests of both the beginning and advanced small-boat sailor. The following represent some of those currently available: BILL ROBINSON, *Better Sailing for Boys and Girls* (1968), is aimed at the youngest would-be sailors and explains the basics of sailing technique. PHILIPPE HARLE, *The Glenans Sailing Manual*, rev. ed. (1967), is directed more toward the adult who is starting off in the sport and includes details about boat handling, sails, the characteristics of different types of boats, and related subjects. FESSENDEN BLANCHARD and THEODORE JONES, *Sailboat Classes of North America*, rev. ed. (1968), includes pictures and basic details about popular sailboat types. Books on tactics are numerous; among those of continuing interest are: R.N. BAVIER, *Sailing To Win* (1969), based on the author's experience in dinghies and class boats up to ocean racers and Twelve Metres; PAUL ELVSTROM, *Elvström Speaks on Yacht Racing* (1969), probably the finest small-boat sailor in the world and winner of four Olympic Gold Medals; JOHN D.A. OAKELEY, *Winning* (1970), one of England's top Flying Dutchman sailors, who delves into valuable detail on all parts of the small one-design boat and discusses racing techniques, and STUART WALKER (ed.), *Performance Advances in Small Boats* (1969), in which 22 of America's top small-boat sailors discuss the latest developments in design and racing techniques in their classes. (*History and other specific topics*): H.L. STONE, W.H. TAYLOR, and W.R. ROBINSON, *The America's Cup Races*, rev. ed. (1970), is the best and most accurate of many histories of the event. F.S. KINNEY (ed.), *Skene's Elements of Yacht Design*, rev. ed. (1962), is the classic work on the subject. R.N. BAVIER, *A View from the Cockpit* (1966), is the skipper's own story of the successful 1964 America's Cup defense. G.W. MIXTER, *Primer of Navigation*, 5th ed. (1967), is a classic reference work. Included among the better general and special histories are: H.I. CHAPELLE, *American Small Sailing Craft* (1951); BILL ROBINSON, *The World of Yachting* (1966); ALFRED F. LOOMIS, *Ocean Racing: 1866–1935* (1936, reprinted 1967); and C. SHERMAN HOYT, *Memoirs* (1950). GERSHOM BRADFORD, *Glossary of Sea Terms* (1954), is a dictionary of nautical language. W.H. DE FONTAINE and B.D. BARKER III, *1001 Questions Answered About Boats and Boating* (1966), is a general reference work that covers a wide variety of topics. J.H. ILLINGWORTH, *Further Offshore*, 6th ed. rev. (1969), is the best known guide among beginning offshore sailors.

Motorboating: JOHN TEALE, *High Speed Motor Boats* (1969) contains information on hull forms employed, model testing, speed predicting, spray rails and hull trimming devices, with an analysis of stern gear, engine bearers, and fuel systems; UFFA FOX, *Seamanlike Sense in Powercraft* (1968), on propulsion systems from the oar to the turbine, hulls from rowboats to cruiser-type warships, air cushion craft, the screw propeller, underwater gear, hull efficiency, and modern ocean power-boat racing; JACK WEST, *Modern Powerboats* (1970), on hull types, propulsion, control, instrument, fuel, wiring, water, refrigeration, and sanitation systems, deck and safety equipment, navigation and communication, and competitive powerboating; PETER DU CANE, *High-Speed Small Craft*, 2nd rev. ed. (1957) on the naval architecture of powerboats up to 130 feet in length; C.F. CHAPMAN, *Piloting, Seamanship and Small Boat Handling* (1971), the overall operation of small craft, including terminology, regulations, equipment requirements, rules of the road, navigation, weather, electronics, signaling, and marlinspike seamanship; D. PHILLIPS-BIRT, *Motor Yacht and Boat Design* (1953), an introduction to the naval architecture of displacement and planning powerboats, with sections on aesthetics, behaviour at sea, and examples in design.

(K.L.O./J.W.D./B.D.B./E.M.Ho./J.Sm.)

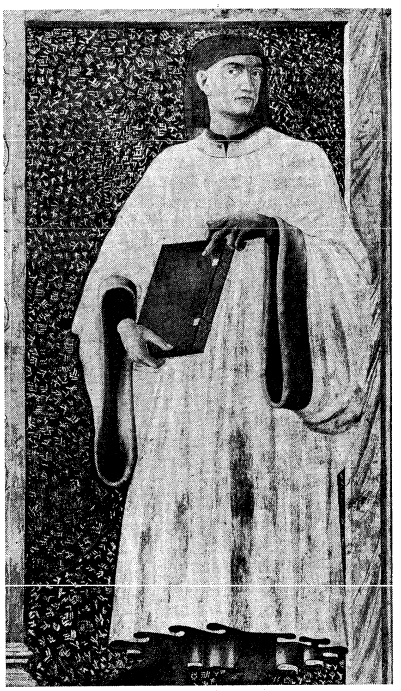
Boccaccio, Giovanni

Giovanni Boccaccio, generally thought of only as the author of the earthy tales in the *Decameron*, was in fact one of the greatest figures in the history of European literature. The early scholarship and writing on the texts of ancient Greece and Rome that became the driving force behind the Humanism of the Renaissance can be traced to Boccaccio and his older contemporary Petrarch and their friends. With Petrarch, he not only marked out the paths along which this Humanism was to develop but also participated in the raising of literature in modern languages to the level and status of the classics of antiquity. It was Boccaccio, too, who raised to literary dignity ottava rima, the verse metre of the popular minstrels, which was eventually to become the characteristic vehicle for Italian verse.

Boccaccio was born in Paris in 1313. His father, Boccaccio di Chellino, called Boccaccino, was a merchant whose family lived originally in Certaldo in Tuscany. Boccaccino's brother Giovanni, however, had moved to Florence not later than 1297, and Boccaccino had followed him. Boccaccio could thus describe himself as a citizen of Certaldo or of Florence, despite the accident of his Parisian birth. His mother was probably French, but his contradictory statements about her in his *Filocolo* and in his *Ninfale d'Ameto*—that she was a young girl of royal birth and that she was a widowed gentlewoman—may both be discounted as examples of a sort of autobiographical embellishment to which he was much inclined, particularly when writing of his early years, in conformity with recognized literary patterns of the time. The year of his birth is given by Petrarch, and Boccaccio's own statement that he was born in Paris of a French mother is supported by documents showing that his father was in Paris in 1310, 1313, and 1314 for business transactions in which the great banking house of the Bardi was engaged.

Family

Youth. Boccaccio passed his early childhood rather unhappily in Florence. His father had no sympathy for Boccaccio's literary inclinations and sent him, not later than 1328, to Naples to learn business, probably in an of-



Boccaccio, detail of a fresco by Andrea del Castagno (c. 1421–57). In the Cenacolo di Sant' Apollonia, Florence.
Alinari—Mansell

fice of the Bardi. Boccaccio claims to have spent six years in this situation and seems also to have spent an equal period studying canon law—likewise on his father's insistence because of the prospect of emoluments but against his own inclination. His claim to have been determined on a literary career from birth is probably to be regarded partly as another autobiographical embellishment of his past.

The Bardi, of whose business his father had become an associate, dominated the court of Naples by means of their loans. Moreover, the king, Robert of Anjou, took pleasure in surrounding himself with men of letters. All doors, then, were open to Boccaccio in Naples, and the years that he spent there were decisive for his education and his tastes. His commercial activity brought him into contact with everyday life in its varied aspects, and his experience of the aristocracy of the business world was to give his writing the special quality that recommended it to that milieu, above all. Contact with the court, on the other hand, showed him all that survived of the splendours of chivalry and feudalism and aroused in him a taste for high and decorous endeavour and for the old nobility's courage, self-control, and courtesy, which he was to pass on to his own world, that of the rising bourgeoisie. He mixed with the learned men of the court and the friends and admirers of Petrarch, through whom he came to know the work of Petrarch himself.

These years in Naples, moreover, were the years of Boccaccio's love for Fiammetta, whose person dominates all his literary activity up to the *Decameron*, in which there also appears a Fiammetta whose character somewhat resembles that of the Fiammetta of his earlier works. Boccaccio's accounts of the love affair are certainly romanticized and contradict one another occasionally: in the earliest works Fiammetta is unfaithful, whereas in the *Elegia di Madonna Fiammetta* it is she who is betrayed; and in *Il filocolo* she is a king's daughter, whereas in *Ameto* and in *Fiammetta* she is only the daughter of a noblewoman married to a rich bourgeois and courted by the king. Attempts to use passages from Boccaccio's writings to identify Fiammetta with a supposedly historical Maria, natural daughter of King Robert and wife of a count of Aquino, are therefore untrustworthy—the more so since there is no documentary proof that this Maria ever existed.

It was probably in 1340 that Boccaccio was recalled to Florence by his father, involved in the bankruptcy of the Bardi. The sheltered period of his life thus came to an end, and thenceforward there were to be only difficulties and occasional periods of poverty. From Naples, however, the young Boccaccio brought with him a store of literary work already completed. *La caccia di Diana* ("Diana's Hunt"), his earliest work, is a short poem, in terza rima (an iambic verse consisting of stanzas of three lines), of no great merit, in which a number of beautiful women of Naples are reviewed within the loose framework of a fable. Much more important are two works with themes derived from medieval romances: *Il filocolo* (c. 1336), a prose work in five books on the loves and adventures of Florio and Biancofiore (Floire and Blanchefleur), and *Il filostrato* (c. 1338), a short poem in ottava rima, telling the story of Troilus and the faithless Criseida. The *Teseida* (probably begun in Naples and finished in Florence, 1340–41) is an ambitious epic of 12 cantos in ottava rima in which the wars of Theseus serve as a background for the love of two friends, Arcita and Palemone, for the same woman, Emilia: Arcita finally wins her in a tournament but dies immediately.

Early works

While all these themes of chivalry and love had long been familiar in courtly circles and were already in favour also with the bourgeoisie and even with the lower classes, Boccaccio not only enriched them with the fruits of his own acute observation of real life and insight into the human heart but also sought to present them nobly and illustriously by a display of learning and rhetorical ornament, so as to make his Italian worthy of comparison with the monuments of Latin literature. These early works had immediate effect outside Italy: Chaucer drew inspiration from *Il filostrato* for his own *Troilus and Criseyde* (as Shakespeare was later to do for *Troilus and Cressida*) and from *Teseida* for his "Knight's Tale" in *The Canterbury Tales*.

Florentine years. The 10 or 12 years following Boccaccio's return to Florence, about 1340, are the period of his full maturity, culminating in the *Decameron*. From 1341 to 1342 he was working on *Il ninfale d'Ameto* ("Ameto's Story of the Nymphs"), in prose and terza rima, which describes how a rude shepherd is raised to spiritual refinement and worldly honour through love. Next came *L'amorosa visione* ("The Amorous Vision"; 1342–43), a mediocre allegorical poem of 50 short cantos in terza rima, in obvious imitation of Dante but revealing also some affinity with Petrarch's *Trionfi* ("Triumphs"), so that some critics have argued that Petrarch's poem was in fact inspired by Boccaccio's. The prose *Elegia di Madonna Fiammetta* (1343–44) shows remarkable psychological penetration, though the style and narrative method are heavy.

The slightly later *Ninfale fiesolano* (perhaps 1344–45), in ottava rima, on the love of the shepherd Africo for the nymph Mensola, is written with a graceful and studied simplicity foreshadowing 15th-century poetry. To complete the survey of Boccaccio's Italian writings before the *Decameron*, mention must be made of the *Rime* ("Poems"), begun in his youth but continued throughout his life, for the most part in the style established for the love lyric by the Sicilian school, by the exponents of the *dolce stil nuovo* ("sweet new style") school, which emphasized delicacy of expression, or by Petrarch.

Boccaccio, meanwhile, was trying continually to put his financial affairs in order, though he never succeeded in doing so. Little is known, however, of the detail of his life in the period following his return to Florence. He was at Ravenna between 1345 and 1346, at Forlì in 1347, in Florence during the ravages of the Black Death in 1348, and in Florence again in 1349.

The "Decameron." It was probably in the years 1348–53 that Boccaccio composed the *Decameron* in the form in which it is read today. In the broad sweep of its range and its alternately tragic and comic views of life, it is rightly regarded as his masterpiece. Stylistically, it is the most perfect example of Italian classical prose, and its influence on Renaissance literature throughout Europe was enormous.

The *Decameron* begins with the flight of ten young people (seven women and three men) from plague-stricken Florence in 1348. They retire to a rich, well-watered countryside, where, in the course of a fortnight, each member of the party has a turn as king or queen over the others, deciding in detail how their day shall be spent and directing their leisurely walks, their outdoor conversations, their dances and songs, and, above all, their alternate storytelling. This storytelling occupies ten days of the fortnight (the rest being set aside for personal adornment or for religious devotions); hence the title of the book itself, *Decameron*, or "Ten Days' Work." The stories thus amount to 100 in all. Each of the days, moreover, ends with a canzone (song) for dancing sung by one of the storytellers, and these canzoni include some of Boccaccio's finest lyric poetry.

Structure of the *Decameron*

In choosing a framework of this sort for his stories, Boccaccio was following a tradition familiar in Oriental and medieval literature. To this tradition, however, he brought a new element. In addition to the 100 stories, he has a master theme, namely, the way of life of the refined bourgeoisie, who combined respect for conventions with an open-minded attitude to personal behaviour.

The sombre tones of the opening passages of the book, in which the plague and the moral and social chaos that accompanies it are described in the grand manner, are in sharp contrast to the scintillating liveliness of Day I, which is spent almost entirely in witty disputation, and to the playful atmosphere of intrigue that characterizes the tales of adventure or deception related on Days II and III. With Day IV and its stories of unhappy love, the gloomy note returns; but Day V brings some relief, though it does not entirely dissipate the echo of solemnity, by giving happy endings to stories of love that does not at first run smoothly. Day VI reintroduces the gaiety of Day I and constitutes the overture to the great comic score, Days VII, VIII, and IX, which are given over to laughter, trickery, and license. Finally, in Day X, all the themes of the preceding days are brought to their most exalted pitch, the impure made pure and the common made heroic. Even if it is not always artistically convincing, this conclusion forms a noble testament, ending with the glorification of fidelity, constancy, and womanly obedience in the story of Griselda.

The prefaces to the days and to the individual stories and certain passages of especial magnificence based on classical models (on Livy at his most florid or on Apuleius at his richest, rather than on Cicero), with their select vocabulary and their elaborate periods, have long held the attention of critics. But there is also another Boccaccio: the master of the spoken word and of the swift, vivid, tense narrative free from the proliferation of ornament. These two aspects of the *Decameron* make it the fountainhead of Italian literary prose for the next centuries.

The romantic view of the *Decameron*, propounded by Francesco De Sanctis, who regarded it as a "Human Comedy" in succession to Dante's *Divine Comedy* and Boccaccio as the pioneer of a new moral order superseding that of the Middle Ages, is no longer tenable, since the Middle Ages can no longer be presented as having been wholly ascetic or wholly turned toward God in contrast with a Renaissance concerned only with the human. Medieval literature in general does not ignore man and the flesh and even has passages that exalt the individual; and the men of the Renaissance, far from ignoring God, strove in many ways to reconcile religious truth and Christian doctrine with both ancient philosophy and the new science.

Also, in particular, the whole corpus of Boccaccio's work is basically medieval in subject matter, form, and taste, at least in its point of departure. It is the spirit in which Boccaccio treats his subjects and his forms that is new. Boccaccio in the *Decameron* for the first time deliberately shows man striving with fortune and learning to overcome it and even, when possible, to exploit it. This marked dualism of virtue and fortune lies at the root of the feeling and thought of the Renaissance. The *Decameron* exalts essentially what Machiavelli was to call the virtue of man: his intelligence, his eagerness, his sense of proportion, his tireless self-control, and his power to bend events to his own designs. To be truly noble, according to the *Decameron*, man must accept life as it is, without bitterness, must accept, above all, the consequences of his own action, however contrary to his expectation or even tragic they may be. To realize his own earthly happiness, he must confine his desire to what is humanly possible and renounce the absolute without regret. Thus Boccaccio insists both on man's powers and on their inescapable limitations. A sense of spiritual realities and an affirmation of moral values underlying the frivolity even in the most licentious passages of the *Decameron* are features of Boccaccio's work that modern criticism has brought to light and that make it no longer possible to regard him only as an obscene mocker or sensual cynic.

Attitudes toward man

During the years in which Boccaccio is believed to have written the *Decameron*, the Florentines appointed him ambassador to the lords of Romagna in 1350; municipal councillor and also ambassador to Louis, duke of Bavaria, in the Tirol in 1351; and ambassador to Pope Innocent VI in 1354.

But of far more lasting importance than official honours was Boccaccio's first meeting with Petrarch, in Florence in 1350, which began a friendship doing honour to both. Boccaccio, who had already written a life of Petrarch in Latin (*De vita et moribus Francisci Petrarchae*), revered the older man as his master, though respect did not preclude frank criticism. Petrarch proved himself a serene and ready counsellor and a reliable helper. Together, through the exchange of books, news, and ideas and through the stimulus that they afforded one another, they laid the foundations of the humanist reconquest of classical antiquity.

Friendship with Petrarch

His meeting with Petrarch, which occurred when he was already working on the *Decameron*, helped to bring about decisive change in Boccaccio's literary activity. After the *Decameron*, of which Petrarch remained in ignorance until the very last years of his life, Boccaccio wrote nothing in Italian except *Corbaccio* (a satire on a widow who had jilted him), his late writings on Dante, and perhaps an occasional lyric. Turning instead to Latin, he devoted himself to humanist scholarship rather than to imaginative or poetic creation. His encyclopaedic *De genealogia deorum gentilium* ("On the Genealogy of the Gods of the Gentiles"), medieval in structure but Humanist in spirit, was probably begun in the very year of his meeting with Petrarch but was continuously corrected and revised until his death. His *Bucolicum carmen* (1351–66), a series of allegorical eclogues (short pastoral poems) on contemporary events, follows Classical models on lines already indicated by Dante and Petrarch. His *De claris mulieribus* (1360–74), a collection of biographies of famous women, is the complement of Petrarch's *De viris illustribus* ("On Famous Men"), just as his *De casibus virorum illustrium* ("On the Fates of Famous Men"; 1355–74), on the inevitable catastrophe awaiting all who are too fortunate, reflects Petrarch's *De remediis utriusque fortunae* ("On the Relief of Each Man's Lot"). Finally, there is his compilation of classical geographic names, *De montibus, silvis, fontibus, lacubus, fluminibus, stagnis seu paludibus, et de nominibus maris* ("On Mountains, Forests, Springs, Lakes, Rivers, Swamps or Marshes, and on the Names of the Sea"; 1355–74).

The meeting with Petrarch, however, was not the only cause of the change in Boccaccio's writing. A premature weakening of his physical powers and disappointments in love may also have contributed to it. Some such occurrence would explain how Boccaccio, having previously written always in praise of women and love, came suddenly to write the bitterly misogynistic *Corbaccio* and then turn his genius elsewhere. Furthermore, there are signs that he may have begun to feel religious scruples. Petrarch describes how the Carthusian monk Pietro Petrone, on his deathbed in 1362, sent another Carthusian, Gioacchino Ciani, to exhort Boccaccio to renounce his worldly studies; and it was Petrarch who then dissuaded Boccaccio from burning his own works and selling his library. As early as 1360, moreover, Boccaccio's way of life was regarded as austere enough to justify his being

entrusted with a pastoral cure of souls in a cathedral. He had taken minor orders many years earlier, perhaps at first only in the hope of being given benefices.

Boccaccio's circle in Florence was of vital importance as a nucleus of early Humanism. Leonzio Pilato, whom Boccaccio housed from 1360 to 1362 and whose nomination as reader in Greek at the Studio (the old University of Florence) he procured, made the rough Latin translation through which Petrarch and Boccaccio became acquainted with Homer's poems—the starting point of Greek studies by the Humanists. The recovery of Latin Classical texts—Varro, Martial, Apuleius, Seneca, Ovid, and, above all, Tacitus—likewise occupied Boccaccio's admiring attention. Even so, he did not neglect Italian poetry, his enthusiasm for his immediate predecessors, especially Dante, being one of the characteristics that distinguish him from Petrarch. His *Vita di Dante Alighieri* or *Trattatello in laude di Dante* ("Little Tractate in Praise of Dante") and the two abridged editions of it that he made show his devotion to Dante's memory.

Last years. All these studies were pursued in poverty, sometimes almost in destitution, and Boccaccio had to earn most of his income by transcribing his own works or those of others. In 1362, on the invitation of Niccolò Acciaiuoli, a Florentine banker and statesman who served as chief adviser to Joanna, queen of Naples, Boccaccio went to Naples in some hope but was disappointed and returned at once to Florence. Then, in 1363, he retired to Certaldo. He was twice sent as ambassador to Pope Urban V, at Avignon in 1365 and in Rome in 1367. In 1370–71 he paid another visit to Naples with no better success than in 1362. In October 1373 he began public readings of Dante's *Divina commedia* in the church of S. Stefano di Badia in Florence. A revised text of the commentary that he gave with these readings is still extant but breaks off in the 17th canto of the *Inferno*, at the point that he had reached when, early in 1374, ill health and the criticisms of those who disapproved of his explaining Dante to the multitude made him lose heart. Petrarch's death in July 1374 was another grief to him, and he retired again to Certaldo.

Commentaries on Dante

There Boccaccio died, on Dec. 21, 1375, and was buried in the church of SS. Michele e Jacopo. Franco Sacchetti expressed the general dismay of men of letters at the death within 18 months of the two great writers when he said that all poetry was now extinct.

Boccaccio and the Renaissance. Boccaccio was a man of the Renaissance in almost every sense. His Humanism comprised not only Classical studies and the attempt to rediscover and reinterpret ancient texts but also the "humanism of the vernacular," which he and Petrarch consciously initiated. This Humanism implied the raising of literature in the modern languages to the level of the classical by setting standards for it and then conforming to those standards. Such an undertaking, however, requires the writer to master his own caprices and impulses, to submit to the restraints of an illustrious and unalterable tradition, to tame his passions by the exercise of his intellect in the interests of formal perfection: poetic originality is to be sought only within the confines of tradition. In the second half of the 15th century and in the 16th century, Humanism did produce a vernacular literature in conformity with these requirements: indeed, the Humanist principle that the Classical ideal should permeate the whole of contemporary life had as its natural consequence that the language and literature in which modern life was expressed should receive the stamp of classicism. Boccaccio advanced farther than Petrarch in this direction not only because he sought to dignify prose as well as poetry but also because, in his *Ninfale fiesolano*, in his *Elegia di Madonna Fiammetta*, and in the *Decameron*, he ennobled everyday experience, tragic and comic alike. Although his *Teseida* and his *Ninfale d' Ameto* invite comparison with classical genres (as do Petrarch's *Trionfi*), his *Filocolo* and his *Filostrato* raise to the level of learned art the literature of chivalry and love that had fallen to the level of the populace. The same attention to popular and medieval themes characterized Italian culture in the second half of the 15th century. Without Boccaccio, the culmination of the Renaissance would be historically incomprehensible.

MAJOR WORKS

IN ITALIAN: *La caccia di Diana* (earliest known work), short poem in terza rima; *Il filocolo*, 5 bks. (*c.* 1336; *A Pleasaunt Disport of Diuers Noble Personages* by Henry Grantham, 1567; with introduction by Edward Hutton, 1927; modernized with introduction by Thomas Bell, 1931), prose; *Il filostrato* (*c.* 1338; Eng. trans. by W.M. Rossetti, 1873; by N.E. Griffin and A.B. Myrick, 1929; by R.K. Gordon, 1934), short poem in ottava rima; *Teseida* (*c.* 1340–41), epic of 12 cantos in ottava rima; *Il ninfale d'Ameto*, or *Commedia delle Ninfe fiorentine* (1341–42), prose and terza rima; *L'amorosa visione* (1342–43), allegorical poem of 50 short cantos in terza rima; *Elegia di Madonna Fiammetta*, or *Fiammetta amorosa* (1343–44; *Amorous Fiammetta* by Bartholomew Young, 1587; rev. by Edward Hutton, 1926; new ed. by K.H. Josling, 1929), prose; *Il ninfale fiesolano* (perhaps 1344–45; in *Two Tracts*, Cyril H. Wilkinson's edition of John Goubourne's Elizabethan version in prose of A. Guercin du Donno, 1946), poem in ottava rima; *Decameron* (probably 1348–53; Eng. trans. 1620; by J.M. Rigg, 1903 and 1960; by Richard Aldington, 1930 and 1958), poem and 100 prose stories, each ending with a canzone of poetry; *Il Corbaccio*, or *Laberinto d'amore* (1354–55), prose; *Vita di Dante Alighieri*, or *Trattatello in laude di Dante*, 1354–55; *Life of Dante* by P.H. Wicksteed, 1898), prose; *Esposizioni sopra la Commedia di Dante* (1373, incomplete at death), prose; *Rime* (composed throughout life, critical and purified text pub. 1914), poems.

IN LATIN: *De vita et moribus Francisci Petrarchae* (1343–45), prose; *De genealogia deorum gentilium* (probably begun 1350, but continuously corrected and rev. until death; *Boccaccio on Poetry: Being the Preface and the Fourteenth and Fifteenth Books of Boccaccio's "Genealogia . . ."* by Charles G. Osgood, 1930), prose, general proem, and 15 books; *Bucolicum carmen* (1351–66; *Boccaccio's Olympia* by Israel Gollancz, 1913), allegorical ecologues; *De casibus virorum illustrium*, 9 bks. (1355–74; *Fall of Princes* by John Lydgate from a French version, 1431–38; ed. by Henry Bergen, 1923–27), prose; *De claris mulieribus* (*c.* 1360–74; *Forty-six Lives Translated from Boccaccio's De claris mulieribus*, by Henry Parker, Lord Morley, and ed. by Herbert G. Wright, 1943; *Concerning Famous Women* by Guido A. Guarino, 1964), prose; *De montibus, silvis, fontibus, lacubus, fluminibus, stagnis seu paludibus, et de nominibus maris* (1355–74), prose.

LETTERS: *Epistola consolatoria a messer Pino de' Rossi; Eristulae rerum familiarum; Le Lettere edite e inedite di Messer Giovanni Boccaccio* (1877).

BIBLIOGRAPHY

Editions: Modern editions of Boccaccio's minor works in Italian are as follows: *Il filocolo*, ed. by S. BATTAGLIA (1938); *Il filostrato* and *Il ninfale fiesolano*, ed. by V. PERNICONE (1937); *Teseida*, ed. by S. BATTAGLIA (1938) and by A. RONCAGLIA (1941); *L'Ameto, Lettere, il Corbaccio*, ed. by N. BRUSCOLI (1940); *Il Corbaccio*, ed. by TAUNO NURMEELA (1968); *L'amorosa visione*, ed. by V. BRANCA (1944); *Elegia di Madonna Fiammetta, con le chiose inedite*, ed. by V. PERNICONE (1939) and by F. AGENO (1954); *Le rime, l'amorosa visione, la caccia di Diana*, ed. by V. BRANCA (1939; new ed. of *Le rime* and *La caccia di Diana*, 1958); *Il commento alla Divina Commedia e altri scritti intorno a Dante*, ed. by D. GUERRI, 4 vol. (1918–26). Complete works, ed. by V. BRANCA: II. *Filostrato*, ed. by V. BRANCA, *Teseida delle nozze di Emilia*, ed. by A. LIMENTANI, *Commedia delle Ninfe fiorentine*, ed. by A.E. QUAGLIO (1964); VI. *Esposizione sopra la Commedia di Dante*, ed. by G. PADOAN (1965); I. *La caccia di Diana*, ed. by V. BRANCA, *Il Filocolo*, ed. by A.E. QUAGLIO. Selections from *Il filocolo*, the *Ameto* and the *Elegia di Madonna Fiammetta* are included in the selections from the *Decameron* by E. BIANCHI, C. SALINARI, and N. SAPEGNO, 6th ed. (1952). Of Boccaccio's Latin works *Bucolicum carmen* and some poems and letters, as well as some shorter poems, are printed in *Opere latine minori*, ed. by A.F. MASSERA (1928); there is an edition of the *Genealogia deorum gentilium libri* by V. ROMANO, 2 vol. (1951). There are editions of the *Decameron* by U. BOSCO (1946–51); by G. PETRONIO, with full commentary (1950); by V. BRANCA, with full commentary (1951–52); by C.S. SINGLETON (1955); by N. SAPEGNO (1956); and by M. MARTI (1958). On the text see M. BARBI, *La nuova filologia e l'edizione dei nostri scrittori* (1938); M. SAMPOLI SIMONELLI, "Il Decameron: problemi e discussioni di critica testuale," *Ann. Sc. norm. sup. Pisa*, vol. 18 (1949); V. BRANCA, "Per il testo del Decameron," *Studi di filologia italiana*, vol. 8 and 11 (1950–53); V. BRANCA, *Tradizione delle opere di Giovanni Boccaccio* (1958); *Concordanze del Decameron*, by A. BARBINA (1969).

Life: A. DELLA TORRE, *La giovinezza di G. Boccaccio (1313–1341)* (1905); E. HUTTON, *G. Boccaccio: A Biographical Study*

(1910); F. TORRACA, *Per la biografia di G. Boccaccio* (1912); H. HAUVETTE, *Boccace, étude biographique et littéraire* (1914); S. BATTAGLIA, "Elementi autobiografici nell'arte del Boccaccio," *La Cultura*, vol. 9 (1930); G. BILLANOVICH, *Restauri boccacceschi* (1945); F. MACMANUS, *Boccaccio* (1947); V. BRANCA, *Boccaccio medievale* (1956).

Studies: For general surveys of Boccaccio's work, see N. SAPEGNO, *Il trecento*, new ed. (1952), with full bibliography; C. GRABHER, *Giovanni Boccaccio* (1941); J. LUCHAIRE, *Boccace* (1951). On the *Decameron* see U. FOSCOLO, *Saggi e discorsi critici*, ed. by C. FOLIGNO (1953); F. DE SANCTIS, *Storia della letteratura italiana*, ed. by B. CROCE (1939); U. BOSCO, *Il Decameron* (1929); B. CROCE, *Poesia popolare e poesia d'arte* (1933); G. PETRONIO, *Il Decamerone: saggio critico* (1935); F. NERI, *Storia e poesia* (1944); G. GETTO, *Vita di forme e forme di vita nel Decameron* (1958). See also E.H. WILKENS, "An Introductory Boccaccio Bibliography," *Philological Quarterly* (1927). On Boccaccio's language and style, see G.R. SILBER, *The Influence of Dante and Petrarch on Certain of Boccaccio's Lyrics* (1940); A. SCHIAFFINI, *Tradizione e poesia nella prosa d'arte italiana della latinità medievale a Giovanni Boccaccio*, new ed. (1943); E.G. PARODI, *Lingua e letteratura*, ed. by G. FOLENA (1957). On the development and history of criticism, see G. PETRONIO, "Giovanni Boccaccio," in *I Classici italiani nella storia della critica*, ed. by W. BINNI (1954).

(U.Bo.)

Boccherini, Luigi

An excellent cellist and a prolific writer of instrumental music, Luigi Boccherini helped to develop a comparatively new musical form, the string quartet (two violins, viola, cello), and he created the quintet for strings only (two violins, viola, two cellos; or two violins, two violas, cello), as well as the quintet with strings and piano. He wrote about 500 works, including more than 100 quartets and 50 trios. He also wrote sacred music, symphonies, and concerti, particularly for his own instrument. A typical musician-traveller of the 18th century, Boccherini's style developed through his artistic contacts in Lucca, Rome, Vienna, Milan, Paris, and Madrid. Both forceful and gentle, his power of musical expression dissolved the previous generation's musical formalism into beauty and gracefulness. The Spanish theoretician Antonio Eximeno called him "the delight of Europe." Although Boccherini's work has frequently been compared to that of Joseph Haydn, his music often lacks his contemporary's characteristic forward drive and virility. Gentle warmth, often with a hint of melancholy, and superlative elegance are perhaps the qualities that best distinguish Boccherini's music.

The third child of a double-bass player, Leopoldo Boccherini, and Maria Santa Prosperi, Luigi Rodolfo was

By courtesy of the Istituto Musicale Boccherini, Lucca



Boccherini, portrait by an unknown artist, 1790.

born in Lucca, Italy, on February 19, 1743. At a very early age he was put under the care of the musical director of the local cathedral. When he reached the age of 13, there was nothing more he could learn there; consequently, he was sent to Rome to study with the renowned cellist Giovanni Battista Costanzi, musical director at St. Peter's. In Rome he was influenced by the polyphonic tradition —*i.e.*, music with two or more interweaving melodic parts—stemming from the works of Giovanni da Palestrina and the instrumental music of Arcangelo Corelli. At the vespers in Rome's churches, he was attracted by those symphonies that ended with a minuet—the dance that he later loved to idealize rather than stylize.

Study in Rome

In 1757 Boccherini and his father were invited to play in the Imperial Theatre orchestra in Vienna; where young Boccherini had the chance to absorb the musical classicism, dominated by the personality of Christoph Willibald Gluck, which was rising in the Austrian capital. On his second journey to Vienna (1760), Boccherini, at 17, made his debut as a composer with his *Six Trios for Two Violins and Cello*, G. 77–82 (Opus 1 in Boccherini's autograph catalog), which were appreciated by the renowned Gluck. During his third stay in that city (1764), a public concert by Boccherini was enthusiastically received.

In spite of his success, Boccherini grew homesick for Lucca, to which he returned (August 1764), having obtained a permanent position with the local church and theatre orchestras there. As a composer, he took part in the religious celebrations in the Church of Sta. Croce and other festivities. Around 1765 he composed various oratorios at Lucca. Boccherini was also in Lombardy in 1765, in the orchestra of Giovanni Battista Sammartini. Through his association with this Milanese composer, the 22-year-old Boccherini strengthened the new "conversational" style of the quartet: the cello's line was now as important as the counterpoint (*i.e.*, the intertwining of independent melodic lines) of the violin and viola. Boccherini had a chance to put into practice this conquest with an extraordinary string quartet made up of outstanding Tuscan virtuosi.

After his father's death (1766), he decided to leave Lucca for good. His destination, agreed upon with the violinist Manfredi, one of the Tuscan virtuosi, was Paris—a happy choice since France welcomed Italian musicians. The French publishers Grangé, Venier, and Chevardière published Boccherini's compositions of the previous years (*Six String Quartets*, G. 159–164 [Boccherini's Opus 2], and *Six Duets for Two Violins*, G. 56–61 [Boccherini's Opus 3], of 1761) and the new ones (*Six Trios for Two Violins and Cello*, G. 83–88 [Boccherini's Opus 4], and *Symphony in D Major*, G. 500, of 1766 and *c.* 1766?), and musical Paris competed for the young man from Lucca. From his contact with Mme Brillon de Jouy, the harpsichord player, were born the wonderful *Six Sonatas for Harpsichord and Violin*, G. 25–30 (Boccherini's Opus 5). Boccherini's style spread throughout Europe, and his *Cello Concerto No. 6 in D Major*, G. 479 (*c.* 1768?), became the model for Mozart's *Violin Concerto in D Major*, K. 218 (1775). Such vital contact and enthusiasm were interrupted when the Spanish ambassador to Paris persuaded Boccherini to move to Madrid. Attracted by this hopeful and flattering offer, he began his long segregation at the intrigue-ridden court of Charles III. The King's brother, Infante Don Luis, conferred on him a yearly endowment of 30,000 reals as a cellist and composer and wanted to keep Boccherini with him even when, after a marriage that met with opposition, he was forced to renounce the throne and retire to the Las Arenas castle, several miles from Madrid. Far away from the intrigues of the scornful Prince of the Asturias (afterward Charles IV) and from jealous court musicians, Boccherini worked in tranquillity and also performed in a quartet formed by members of the aristocracy. During this period he wrote his well-known *Six String Quartets*, G. 177–182 (Boccherini's Opus 15, Opus 11 in some editions, composed 1772).

Move to Madrid

Madrid became Boccherini's second home. There he married Clementina Pelicho, who bore him five children. At the Infante's death (1785), the King granted him a pension of 12,000 reals. He received another pension from Frederick William II of Prussia, who was an ama-

teur cellist. Boccherini had a long, friendly relationship with Frederick and dedicated to him a symphony and dozens of trios, quartets, and quintets. Lastly, the Duchess of Osuna appointed him conductor of her private orchestra at the Puerta de la Vega Palace in Madrid. To his prodigious instrumental production, Boccherini added vocal compositions: the *Stabat Mater*, G. 532 (1781), the zarzuela (a form of comic opera) *La Clementina*, G. 540 (composed 1786), with libretto by Ramon de la Cruz, and the Christmas *Villancicos*, G. 539 (1783). Though faithful to traditional classical musical forms, Boccherini was also receptive to Spanish popular music. He became, in fact, the musical representative of the "Goya epoch," the age of the great Spanish painter Francisco Goya (1746–1828), who was one of Boccherini's friends.

Having lost his first wife, Boccherini married Joaquina Porreti (1787). From 1787 to 1797, he was probably in Berlin, at a post provided by Frederick William II, but historical evidence for his stay there is lacking. In 1798, the new king of Prussia withdrew Boccherini's pension, the Duchess of Osuna moved to Paris, and Boccherini's financial stress was aggravated by poor health. His trusting nature was shaken by the greed of the Parisian publisher Pleyel, and his life was saddened by the death of his second wife and two daughters during an epidemic. After living for some years in straitened circumstances, his situation was improved briefly by the patronage of Lucien Bonaparte, who had been appointed in 1800 as ambassador of the French Republic in Madrid. After 1801 he fell into greater poverty, and in 1804 he was forced to live in one room with his three surviving children. His last complete work, the *String Quartet No. 90 in F Major*, G. 248, was composed that year. He died on May 28, 1805, in Madrid and was buried in the Church of San Justo. In 1927 Boccherini's remains were moved to the Church of S. Francesco in Lucca.

In 1969 the French scholar Yves Gérard published his *Thematic, Bibliographical, and Critical Catalogue of the Works of Luigi Boccherini*. Numbers preceded by "G." are the numbers assigned by Gérard according to type of composition and are not in chronological order.

MAJOR WORKS

OPERA: *La Clementina*, G. 540 (composed 1786).

ORCHESTRAL MUSIC: 20 symphonies; four cello concerti.

CHAMBER MUSIC: 125 string quintets; 12 quintets for piano and strings; 18 quintets for flute or oboe and strings; 102 string quartets; 60 string trios; 27 violin sonatas; six cello sonatas.

CHURCH MUSIC: *Stabat Mater*, G. 532, for three voices and strings (first version completed 1781, second version 1800); mass sections for four voices and instruments; *Christmas Cantata*, G. 535 (Boccherini's Opus 63, composed 1802); various motets.

BIBLIOGRAPHY. G. DE ROTHSCHILD, *Luigi Boccherini* (1962; Eng. trans., 1965), a biography incorporating much new information: L. PICQUOT, *Notice sur la vie et les ouvrages de Luigi Boccherini* (1851; new ed., 1930); A. BONAVENTURA, *Boccherini* (1931), a standard biography in Italian; G. BARBLAN, "Boccheriniana," in *La Rassegna Musicale*, vol. 30, pp. 33–44 (1960); A. BONACCORSI, "Boccherini," in *La Musica* (1966), two articles in Italian reference works: Y. GERARD, *Thematic, Bibliographical, and Critical Catalogue of the Works of Luigi Boccherini* (1969), with 580 compositions.

(Gu.B.)

Body Cavities and Membranes, Human

The entire body has many cavities—the brain cavity within the head, for example, discussed in the article SKELETAL SYSTEM, HUMAN; the sinuses, described in the article SINUS; the space within the stomach and intestines (see DIGESTIVE SYSTEM, HUMAN) and that within the urinary bladder (see EXCRETORY SYSTEM, HUMAN). The present article is focussed on the major cavities within the trunk: the chest, or thoracic cavity, and the abdominal cavity, together with the membranes that line these cavities and cover the organs and structures within them. The chest and the abdominal cavity contain most of the important organs concerned with respiration, circulation, and digestion. The article also includes a discussion of the principal diseases and disorders that affect these two cavities.

The chest and the pleura

Thoracic cavity. The thoracic cavity, second only in size to the abdominal cavity, contains the lungs, the middle and lower airways—the tracheobronchial tree—the heart, the vessels transporting blood between the heart and the lungs, the great arteries bringing blood from the heart out into the general circulation, and the major veins into which the blood is collected for transport back to the heart. The chest also contains the esophagus, the channel for food from the throat to the stomach (see Figure 1).

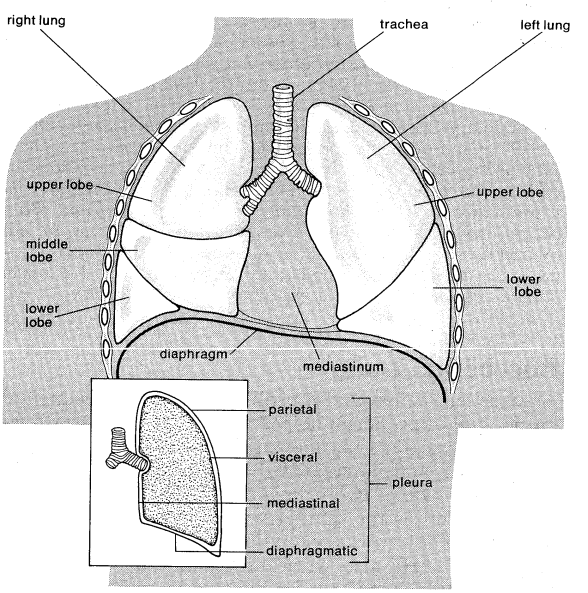


Figure 1: Chest cavity and pleura.

The chest cavity, like the abdominal cavity, is lined with a serous membrane, so called because it exudes a thin fluid, or serum. This portion of the chest membrane is called the parietal pleura. The membrane continues over the lung, where it is called the visceral pleura, and over part of the esophagus, the heart, and the great vessels: the mediastinal pleura, the mediastinum being the space and the tissues and structures between the two lungs. Because the atmospheric pressure between the parietal pleura and the visceral pleura is less than that of the outer atmosphere, the two surfaces tend to touch, friction between the two during the respiratory movements of the lung being eliminated by the lubricating action of the serous fluid. The pleural cavity is the space, when it occurs, between the parietal and the visceral pleura.

Inside the thoracic cavity is another smaller cavity, the pericardial sac, which contains the heart and the beginnings of the great vessels that enter and leave the heart (see CARDIOVASCULAR SYSTEM, HUMAN).

The structure and vessels of the pleura

The pleura is a continuous sheet of endothelial, or lining, cells supported by a thin base of loose connective tissue. The membrane is well supplied with blood vessels, nerves, and lymph channels. The vessels of the visceral part of the pleura are intimately related with those of the lungs and bronchi: its arteries are branches of the bronchial arteries, and its veins mingle with the pulmonary network of capillaries. Beneath its inner side is a network of tiny lymph channels, or capillaries, that penetrate the lung substance, or parenchyma, and drain to the lymph nodes at the hilus of each lung, the point of entrance and departure for bronchi, blood vessels, and nerves.

Diseases affecting the pleura and pleural cavity, other than primary tumours, are brought by the blood vessels or may spread from contiguous structures. The pleural cavity may be contaminated by rupture of either the visceral pleura or the parietal pleura.

Accumulation of fluid in the pleural cavity is called hydrothorax. If the fluid is bloody, it is termed hemothorax; and if it contains pus, pyothorax. The accumulation of fluid may or may not be accompanied by air. When air is present, the prefix pneumo is attached to each of the names mentioned—*e.g.*, hydropneumothorax, etc.

Hydrothorax; pneumothorax; pleurisy

The penetration of air into the pleural cavity from outside, as from a penetrating wound of the chest, or from within, by rupture of dilated alveoli (air sacs of the lung) or of a cyst, will produce a pneumothorax, converting this cavity into a positive pressure chamber and collapsing the lung, which in turn will lead to lessened oxygenation of the venous blood. The collapse may also have a deleterious effect upon the heart.

Pleurisy is a term applied to the inflammation of the pleura, usually diffuse, affecting one or both sides. Two forms are distinguished: (1) simple, dry, or fibrinous pleurisy; and (2) exudative pleurisy, in which the pleural membrane gives off excessive fluid. Since the pleura is well supplied by nerves, pleurisy can be extremely painful, especially as the lung moves on respiration.

If the pleurisy is severe and extensive enough, there may be an effusion, outpouring of fluid, at times infected. Pleurisy is commonly caused by infection in the underlying lung. More rarely, effusion results from diffuse inflammatory conditions such as rheumatoid arthritis. Rupture of the thoracic duct, the main channel for lymph, gives rise to chylothorax, characterized by escape of lymph into the pleural space.

The most common symptoms of persons who experience pleural effusion are pain and shortness of breath. Auscultation (listening to sounds within the organ) and percussion, diagnostic techniques utilizing the stethoscope and tapping, reveal replacement of the tympanic (drumlike) sound by a dull and solid one. The breath sounds are barely heard because of the interposition of fluid between the lung and the chest wall. Fever is usually present. The diagnosis of pleurisy is suggested by the affected person's account of his symptoms and is confirmed by clinical examination; a more precise evaluation of the amount and location of the fluid and the effects upon lungs and heart is made by X-ray examination.

The treatment of pleurisy is directed toward the evacuation of the fluid and the treatment of the underlying condition.

Epidemic pleurodynia, or Bornholm disease, an acute infection of the various tissues of the pleural cavity by the group B Coxsackie virus, is characterized by a general feeling of ill health and by pleuritis-like pain in the chest muscles and the upper part of the abdomen. This pain is usually increased by respiration and cough, and pain in other muscles is often present. The condition usually subsides in two to five days but sometimes may take weeks to disappear.

The peritoneum and the abdominal cavities

The abdominal cavity. The abdominal cavity is separated from the chest cavity by the diaphragm, a sheet of muscle and connective tissue. The cavity, lined by the peritoneum, a membrane similar to the pleura (see Figure 2), covers not only the inside wall of the abdominal cavity (parietal peritoneum) but also almost every organ or structure contained in it (visceral peritoneum). The space between the visceral and the parietal peritoneum, the peritoneal cavity, is a potential space only, normally containing a small amount of serous fluid that permits free movement of the viscera inside the peritoneal cavity. This motion is particularly true of the gastrointestinal tract. The peritoneum, by connecting the visceral with the parietal portions, assists in the support and fixation of the abdominal organs. The diverse attachments of the peritoneum divide the abdominal cavity into several compartments, of which the compartment called the lesser sac is of particular importance; it is located behind the stomach and is connected with the rest of the peritoneal cavity, the greater sac, by the foramen (opening) of Winslow.

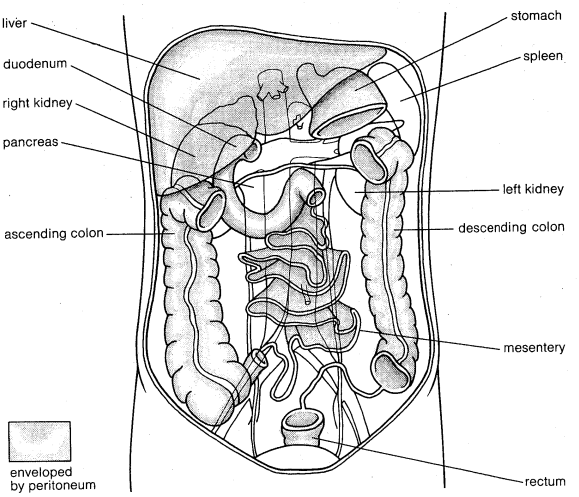


Figure 2: Abdominal cavity, showing relationship of the peritoneum and the abdominal viscera.

Some of the viscera are attached to the abdominal walls by broad areas of peritoneum, as is the pancreas. Others —*e.g.*, the liver—are attached by folds of peritoneum and ligaments, usually poorly supplied by blood vessels.

The peritoneal ligaments are actually rather strong peritoneal folds, usually connecting viscera to viscera or viscera to the abdominal wall; their name usually derives from the structures connected by them (*e.g.*, gastrocolic ligament, connecting the stomach and the colon; the splenocolic ligament, connecting the spleen and the colon), or from their shape (*e.g.*, round ligament, triangular ligament).

The mesenteries are folds of peritoneum that are attached to the wall of the abdomen and enclosing viscera. They are richly supplied with vessels that carry blood to or from the organs that they enfold. The three most important mesenteries are the mesentery for the small intestines; the transverse mesocolon, which attaches the transverse portion of the colon to the back wall of the abdomen; and the mesosigmoid, which enfolds the sigmoid portion of the colon.

The omenta are folds of peritoneum enclosing nerves, blood vessels, lymph channels, and fatty and connective tissue. There are two omenta: the greater omentum hangs down in front of the small intestine like an apron; the lesser omentum is much smaller and extends between the stomach and the liver.

Diseases of the peritoneal membranes. Ascites is the accumulation of fluid in the peritoneal cavity. This fluid contains relatively small numbers of cells, differing in this respect from the fluid associated with peritonitis. The most common causes of ascites in the following order are cirrhosis of the liver with portal hypertension (liver disease involving destruction of liver cells, formation of scar tissue, hardening and contraction of the liver, and elevation of blood pressure in the portal vein); heart failure; tumour invasion of the peritoneum; and escape of chyle (lymph laden with emulsified fats).

Fluid in cavity

The presence of fluid in the peritoneal cavity is not noticeable until such a volume is attained as to distend the abdomen. The accumulation of fluid will produce pressure against the abdominal viscera and veins and also against the thoracic cavity by pressing upon the diaphragm. Treatment is directed toward removal of the cause. Decrease in portal vein pressure in many cases relieves the ascites that accompanies cirrhosis. Chylous ascites is best treated by closure of the leaking lymphatic vessel. Adequate treatment of the heart failure will usually produce regression of the ascites that the heart failure has caused.

Peritonitis, like pleurisy, is usually secondary to an inflammatory process elsewhere, which may come from an adjacent structure or organ; may be introduced from the outside by surgery or by injury; may come from organs in the abdomen, or may be borne by the bloodstream or the lymphatics. The most common origin of peritonitis is the gastro-intestinal tract. Peritonitis may be acute or chronic, generalized or localized, and it may be due to one agent or to a number of them. It is secondary to perforation of the intestines, for example. The severity of the reactions is related, at least in part, to the extent of the peritoneal contamination. In localized peritonitis, the surrounding structures, mainly the greater omentum, will enclose the infected area and will temporarily control the infection. If no treatment is started, the infection may progress throughout the entire abdominal cavity. Often the period in which the peritonitis is localized is short,

and the peritoneal inflammation becomes generalized with great rapidity.

Control of the source of inflammation, either by surgical or by medical means, is followed either by remission of all evidence of peritoneal inflammation or infection or by formation of localized abscesses inside of the peritoneal cavity. Antibiotic therapy has considerably decreased the incidence of the latter complication. When an abscess does develop, antibiotic therapy and adequate external drainage are necessary. The most frequent sites for development of these abscesses are the spaces beneath the diaphragm and the pelvic cavity.

Adhesions between layers of peritoneum are commonly due to abdominal operations but may result from infections in the abdomen or the pelvis. Loops of small bowel or sigmoid colon may become caught between adhesions or between these and the abdominal wall or organs or may even become twisted, with resultant partial or total obstruction of the intestine. If the obstruction interferes with the blood supply, it may cause death of a section of the bowel. Obstruction of the bowel causes cramping abdominal pain and often causes nausea and vomiting. The obstruction is generally eliminated by surgery; in rare cases it disappears spontaneously.

Use of peritoneum in removal of waste products from the blood

Absorptive properties of the peritoneum. The peritoneum has a wide surface and is able to absorb water and substances in solution. This property has proved to be a liability in certain situations—for example, the membrane is known to absorb toxins arising from obstruction and strangulation (cutting off the circulation) of the intestine—but has been useful in the treatment of kidney disease. In the latter situation, the peritoneum has been used, like the semipermeable membrane of the artificial kidney, to remove waste products from the blood.

Torsion and infarction of the omentum. The omentum can become twisted either without known cause or as a result of adhesions, tumours, inflammation, or hernia. The twisting of the omentum may press on the blood vessels and cause mild interference with the blood supply or, in its severe form, may cause infarction (tissue death) in a part of the omentum. The first and predominant symptom is pain in the lower part of the abdomen. When the amount of infarcted omentum is rather large, the mass can be felt. Signs of acute inflammation are present. In treatment, the dead tissue is removed by surgery, and the underlying cause is corrected.

BIBLIOGRAPHY. Additional information may be found in the following anatomical texts: E.D. GARDNER, D.J. GRAY, and R. O'RAHILLY, *Anatomy: A Regional Study of Human Structure*, 3rd ed. (1969); J.C.B. GRANT, *An Atlas of Anatomy, by Regions*, 5th ed. (1962); and H. GRAY, *Anatomy of the Human Body*, 28th ed. by C.M. GOSS (1966).

(J.D.H./C.M.C.)

Boethius

Anicius Manlius Severinus Boethius, Roman scholar, philosopher, theologian, and statesman, is best known for his *De consolatione philosophiae* (Eng. trans., *The Consolation of Philosophy*, 1963). When the barbarians were invading the Western Roman Empire in the late 5th and early 6th centuries and cultural life was declining and the future of Rome was unclear, Boethius was one of a few in the upper classes who set about transmitting the texts of ancient thought to posterity.

The most succinct biography of Boethius and the oldest was written by Cassiodorus, his senatorial colleague, who described him as the accomplished orator in both Greek and Latin who delivered a fine eulogy of King Theodoric, who also wrote on theology, composed a pastoral poem, and was most famous as a translator of works of Greek logic and mathematics.

Other ancient sources, including Boethius' own *De consolatione philosophiae*, give more details. He belonged to the ancient Roman family of the Anicii, which had been Christian for about a century and of which Emperor Olybrius had been a member. Boethius' father had been consul in 487 but died soon afterward, and Boethius was raised by Quintus Aurelius Memmius Symmachus, whose daughter Rusticiana he married. He became consul in



Boethius, detail of a miniature from a Boethius manuscript, 12th century. In the Cambridge University Library, England. (MS Ii.3.12(D)).
By courtesy of Cambridge University Library, England

510 under the Ostrogothic king Theodoric, who had usurped the imperial rule of Italy. Although little of Boethius' education is known, he was evidently well trained in Greek. His early works on arithmetic and music are extant, both based on Greek handbooks by Nicomachus of Gerasa, a 1st-century-AD Palestinian mathematician. There is little that survives of Boethius' geometry, which was based on Euclid, and there is nothing of his astronomy.

Aristotelian translations and Christian theological works

It was his scholarly aim to translate the works of Aristotle with commentary into Latin, and all the works of Plato "perhaps with commentary," to be followed by a "restoration of their ideas into a single harmony." Boethius' dedicated Hellenism, modelled on Cicero's, supported his long labour of translating Aristotle's *Organon* (six treatises on logic) and the Greek glosses on the work.

Boethius had begun before 510 to translate Porphyry's *Eisagogē*, a 3rd-century Greek introduction to Aristotle's logic, and elaborated it in a double commentary. He then translated the *Katēgoriai*, wrote a commentary in 511, the year of his consulship, and also translated and wrote two commentaries on the second of Aristotle's six treatises, the *Peri hermeneias* ("On Interpretation"). A brief ancient commentary on Aristotle's *Analytika Protera* ("Prior Analytics") may be his too; he also wrote two short works on the syllogism. Instead of elucidating his translation of Aristotle's *Topika*, Boethius wrote a commentary on the *Topica* of Cicero, which was already familiar to Latin readers.

About 520 Boethius put his close study of Aristotle to use in four short treatises in letter form on the ecclesiastical doctrines of the Trinity and the nature of Christ; these are basically an attempt to solve disputes that had resulted from the Arian heresy, which denied the divinity of Christ. Using the terminology of the Aristotelian categories, Boethius described the unity of God in terms of substance and the three divine persons in terms of relation. He also tried to solve dilemmas arising from the traditional description of Christ as both human and divine, by deploying precise definitions of "substance," "nature," and "person." There is also a titleless treatise, a brief confession of the Catholic faith, not quite so Aristotelian in its idiom as the other four, the authenticity of which has been disputed. Indeed, doubt has at times been cast on all of Boethius' theological writings because in his logical works and in the later *Consolation*, Christian idiom is nowhere apparent. The 19th-century discovery of Cassiodorus' biography, however, confirmed

Boethius as a Christian writer, even if his philosophic sources were non-Christian. The idiom of his theological writing suggests that there, too, as in the logic and the arts, he kept close to Greek models; for it was in the Greek part of the empire that heretical disputes and new terminology had arisen.

About 520, Boethius became *magister officiorum* (head of all the government and court services) under Theodoric. His two sons were consuls together in 522. Eventually, Boethius fell out of favour with Theodoric. The *Consolation* contains the main extant evidence of his fall but does not clearly describe the terms of the accusation against him. After the healing of a schism between Rome and the church of Constantinople in 520, Boethius and some other senators may have been suspected of communicating with the Byzantine Emperor, who was orthodox in faith whereas Theodoric was Arian. Boethius openly defended the senator Albinus, who was accused of treason "for having written to the Emperor Justin against the rule of Theodoric." The charge of treason brought against Boethius was aggravated by a further accusation of the practice of magic, or of sacrilege, which the accused was at great pains to reject. Sentence was passed and was ratified by the Senate, probably under duress. In prison, awaiting execution, Boethius wrote his *De consolatione philosophiae*.

Content of the *Consolation*

The *Consolation* is the most personal of Boethius' writings, the crown of his philosophic endeavours. Its style, a welcome change from the Aristotelian idiom that provided the basis for the jargon of medieval Scholasticism, seemed to the 18th-century English historian Gibbon "not unworthy of the leisure of Plato or Tully." The argument of the *Consolation* is basically Platonic. Philosophy, personified as a woman, converts the prisoner Boethius to the Platonic notion of Good and so nurses him back to the recollection that, despite the apparent injustice of his enforced exile, there does exist a *summum bonum* ("highest good"), which "strongly and sweetly" controls and orders the universe. Fortune and misfortune must be subordinate to that central Providence, and the real existence of evil is excluded. Man has free will, but it is no obstacle to divine order and foreknowledge. Virtue, whatever the appearances, never goes unrewarded. The prisoner is finally consoled by the hope of reparation and reward beyond death. Through the five books of this argument, in which poetry alternates with prose, there is no specifically Christian tenet, no certain biblical quotation. It is the creed of a Platonist, though nowhere glaringly incongruous with Christian faith. The most widely read book in medieval times, after the Vulgate Bible, it transmitted the main doctrines of Platonism to the Middle Ages. The modern reader may not be so readily consoled by its ancient modes of argument, but he may be impressed by Boethius' emphasis on the possibility of other grades of Being beyond the one humanly known, and other dimensions to the human experience of time.

After his detention, probably at Pavia, he was executed in 524. His remains were later placed in the church of S. Pietro in Ciel d'Oro in Pavia, where, possibly through a confusion with his namesake, St. Severinus of Noricum, they came to receive the veneration due to a martyr, and a memorable salute from Dante.

Influence of Boethius

When Cassiodorus founded a monastery at Vivarium, in Campania, he installed there his Roman library, and included Boethius' works on the liberal arts in the annotated reading list (*Institutiones*) that he composed for the education of his monks. Thus, some of the literary habits of the ancient aristocracy entered the monastic tradition. Boethian logic dominated the training of the medieval clergy and the work of the cloister and court schools. His translations and commentaries, particularly those of the *Katēgoriai* and *Peri hermeneias*, became basic texts in medieval Scholasticism. The great controversy over Nominalism (denial of the existence of universals) and Realism (belief in the existence of universals) was incited by a passage in his commentary on Porphyry. Translations of the *Consolation* appeared early in the great vernacular literatures, with King Alfred (9th century) and Chaucer (14th century) in English, Jean de Meung (a 13th-century poet) in French, and Notker Labeo (a monk of around the turn of the 11th century) in German. There was a Byzantine version in the 13th century by Planudes and a 16th-century English one by Elizabeth I.

Thus the resolute intellectual activity of Boethius in an age of change and catastrophe affected later, very different ages; and the subtle and precise terminology of Greek antiquity survived in Latin when Greek itself was little known

BIBLIOGRAPHY. General works include M. CAPPUYNS, "Boèce," in *Dictionnaire d'histoire et de géographie ecclésiastique*, vol. 9 (1936), with bibliography; EDWARD K. RAND, *Founders of the Middle Ages* (1928); PIERRE COURCELLE, *Les Lettres grecques en occident de Macrobe à Cassiodore* (1943; Eng. trans., *Late Latin Writers and Their Greek Sources*, 1969); and HOWARD R. PATCH, *The Tradition of Boethius* (1935). For editions of the texts and a fine general survey and bibliography, see LORENZO MINIO-PALUELLO's article "Boethius" in the *Dictionary of Scientific Biography*, vol. 2, pp. 228–236 (1970). For the abundant literature on the *De consolatione philosophiae*, see the edition, with bibliography, by LUDWIG BIELER, in the "Corpus Christianorum Series" (1957). On the Aristotelian translations, see LORENZO MINIO-PALUELLO in the *Aristoteles Latinus* editions (1955); on the commentaries, JAMES SHIEL in *Mediaeval and Renaissance Studies*, vol. 4 (1958); and LORENZO MINIO-PALUELLO in *Journal of Hellenic Studies*, vol. 77 (1957); and on the theological treatises, VICTOR SCHURR, *Die Trinitätslehre des Boethius im Lichte der "skythischen Kontroversen"* (1935).

(J.Shi.)

Boğazköy

Boğazköy is the site of the ancient capital of the Hittites, a people who established a powerful empire in Anatolia and northern Syria in the 2nd millennium BC. The site lies about 125 miles (200 kilometres) by road east of Ankara, Turkey. The official modern name of the village is Boğazkale; it is incorporated as a township (Turkish *belediye*) and is the administrative centre of a county (*bucak*) in the district (*ilçe*) of Sungurlu, province (*il, vilâyet*) of Çorum. Before its recent renaming the place was called Boğazköy (Gorge Village) after its most conspicuous natural feature; this name, under which it became known, is still widely used.

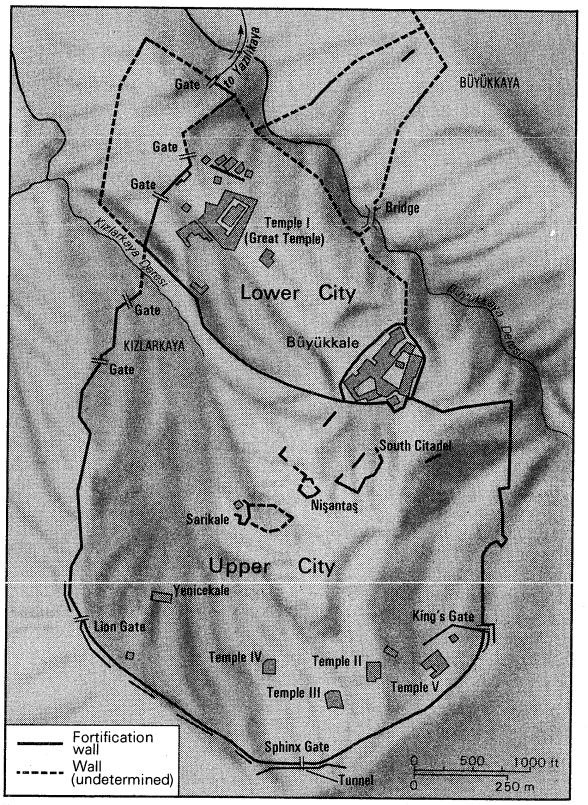
Topography

The ancient city. The ancient city occupies a section of a mountain slope at the southern end of a small fertile plain. It lies between two deeply cut stream beds, filling the angle between their converging courses. Their confluence at the level of the plain (elevation about 3,100 feet) marks the northernmost point of the city area, which rises toward the south by about 1,000 feet (300 metres) on a length of 1¼ miles (two kilometres). The eastern valley narrows in some places to form a real gorge.

The earliest settlement in the city area dates to the 3rd millennium BC, in the so-called Early Bronze Age. There are no written documents that would reveal the identity of the first settlers. Remains of this period were found on top and at the northwest foot of the high hill dominating the east side of the city, known as Büyükkale (Great Fortress), which later became the acropolis of the Hittite kings.

The earliest written sources found at Boğazköy are clay tablets inscribed in cuneiform writing in the Old Assyrian language. They attest the presence of Assyrian merchants on the site, which at that time was called Hattus. The largest Assyrian trade colony was at Kanesh (Kültepe, near Kayseri). Whereas the latter flourished from around 1950 to 1850 BC and, after a destruction, re-emerged sometime around 1820 and lasted through another two generations, that of Boğazköy is contemporary with only this later period.

Hattus was the name of the city also in the language of the early inhabitants of the "Land of Hatti," a language still little understood and not belonging to any known family. Scholars call it Hattic to distinguish it from Hittite, the name of the Indo-European official language of the Hittite kingdom. Just as in other parts of the world, the Indo-European speakers must have been invaders



Plan of Boğazköy.

After *Hattusha: The Capital of the Hittites* by Kurt Bittel, copyright © 1970 by Oxford University Press, Inc.; reprinted by permission

who conquered the older population, but the exact date and the details are unknown, except that individuals bearing Indo-European Hittite names are attested before 1850 in Kanesh in documents of the older Assyrian colony. In Hattus the population or its rulers may still have been "Hattians" even during the later colony period, but there is no proof of it. The houses of the merchants were in the lower city, north of the site of the later Great Temple. The town extended up to Büyükkale, probably culminating in the palace of the local king. Both this town and the merchants' houses were destroyed, probably by King Anittas of Kussara (after 1800). A Hittite text ascribed to Anittas tells of his conquests in Anatolia and how he defeated King Piyusti of Hattus, destroyed the city, and put a curse on the site.

The Indo-European speakers added the vowel *a* to the city name and declined it according to their language; thus the nominative case became Hattusas. The first mention of the name in the form of Hattusa is on a tablet from Mari, on the middle Euphrates, datable to the time of Hammurabi (or Hammurapi; 1792–50); possibly this refers to the city before its destruction by Anittas.

The Old Hittite city

About the middle of the 17th century BC, another king of Kussara, disregarding Anittas' curse, made Hattusa his capital; while his own name was Labarnas, he became known as Hattusilis I, "The one from Hattusa." He is the first ruler of whom there are authentic texts in the Hittite language and one of the founders of the Old Hittite kingdom. One of his successors, Hantilis, is said to have fortified the city. The line of the Old Hittite city wall can be followed today: it surrounds the northern terraces and follows the edge of the eastern valley up to Büyükkale, which must then have been the acropolis; in the west it descends along a side valley to the lower terrace.

Hattusa remained the capital throughout Hittite history with few interruptions. The sources tell of one destruction of the city by enemies around 1380 BC. Soon afterward, King Suppiluliumas I not only restored the city but, by his conquest of most of Anatolia and of Syria to the Lebanon, made it the capital of a real empire. Scholars have wondered how this empire could be ruled from such a remote place. Apart from tradition, it must have been the natural advantages of the site—plentiful water and the protection provided by the terrain—that kept the dynasty at Hattusa. King Muwatallis (around 1300 BC) is said to have transferred the capital to the south "upon the command of the gods." The real reasons behind this move are not known, and his successor brought the government back to Hattusa where it remained to the end.

In the New Kingdom, or Empire, Period (*c.* 1400–*c.* 1190 BC) the city was extended toward the south: the wide arc of the city wall, running from Büyükkale up to the top and down along the western valley, dates from the New Kingdom. The inclusion of these southern hills strengthened the fortification system considerably, especially by the sophisticated construction of the highest section. Here a deep fosse, or moat, and a high earth rampart carrying the wall protected the city against the hills to the south, while a corbelled tunnel, called Yerkapu (Ground Gate), a paved glacis (or defensive slope), and flights of steps aided the defenders. The gates with sculptural decoration also belong to this period: the King's Gate in the southeast (whose relief of a warrior god is in the Ankara museum) and the Lion Gate in the southwest.

Ara Guler



Lion Gate, southwest section of Boğazköy, New Kingdom, or Empire, Period (*c.* 1370–1290 BC).

Although temples are mentioned in the Old Kingdom, the actual temple ruins all date from the New Kingdom. Four temples in the upper city belong to the extension just mentioned. Also the Great Temple in the lower city, in the shape revealed by excavation with its large complex of subsidiary buildings, is a work of the empire. The acropolis, Büyükkale, was completely rebuilt in a monumental manner during the same period.

A further extension of the city wall across the gorge so as to enclose the plateau of Büyükkaya, the "Great Rock" east of the valley, belongs to the 13th century, as do the reliefs and structures at the rock sanctuary, Yazılıkaya, about a mile from the city on the slope of the eastern mountains.

Residential quarters have been excavated only in small parts. Large sections of the city area are taken up by temples and other official buildings, while other parts are too precipitous for building. It is therefore not possible to estimate the population of the city.

Destruction

With the downfall of the Hittite Empire (*c.* 1190 BC) the city was destroyed; traces of burning are found in all parts. The site, it seems, remained vacant for a long time. The next settlement, mainly on Büyükkale and in the lower city, was in size and layout much more modest than the Hittite capital. Through pottery and other finds, this settlement is linked to the Phrygians, to whose kingdom the region belonged in the 8th century BC. These post-Hittite settlements saw several rebuildings and lasted into the Hellenistic Age (3rd to 1st centuries BC), when a tribe of the Galatians, Celtic invaders from Europe, settled in the region. There are only scattered remains of Roman and early Christian times, after which the place

was again uninhabited until the foundation of the village of Boğazköy in the 18th century AD.

Excavations. Boğazköy was discovered in 1834 by the French explorer Charles Texier, who saw Yazılıkaya and those remains of the ancient city that were above ground. After visits by British and German travellers, it was another Frenchman, Ernest Chantre, who in 1892–93 made the first soundings and found the first cuneiform tablets there. The language in which these texts were written was not known at the time, but its identity with that of the so-called Arzawa letters found in Tell el-Amarna in Egypt was soon recognized. This led the Berlin Assyriologist Hugo Winckler to undertake excavations in 1906 together with Theodore Makridi (Bey) of the Istanbul Museum. This first season yielded 2,500 fragments of tablets from the west side of Büyükkale, including some in Akkadian; these showed that Boğazköy was the capital of the kings of Hatti. Winckler and Makridi returned in 1907, 1911, and 1912. In 1907 another German expedition under Otto Puchstein excavated and surveyed the fortifications and temples. After World War I new excavations were started by the German Archaeological Institute and the German Orient Society, with Kurt Bittel as field director. They continued from 1931 to 1939 and again after World War II. These excavations established the stratigraphy and, thereby, the history of the site, besides yielding many more tablets from several locations on Büyükkale and the area of the Great Temple. The excavators had to remove the post-Hittite structures in order to reach the Hittite levels, and they covered all levels earlier than the Hittite Empire with earth again in order to present and preserve as much as possible of the remains of the city of the 13th century BC (see also ANATOLIA, ANCIENT).

BIBLIOGRAPHY. KURT BITTEL, *Hattusha: The Capital of the Hittites* (1970), is a comprehensive description of the city and its history and remains, with an extensive bibliography. See also E. AKURGAL, *The Art of the Hittites* (Eng. trans. 1962).

(H.G.G.)

Bogotá

Bogotá (Santa Fé de Bogotá) is the capital of the Republic of Colombia and of the Department of Cundinamarca. The mountain-rimmed city is built on the eastern margin of a fertile plateau 8,660 feet (2,640 metres) above sea level in the Eastern Cordillera of the Andes mountains. In the city, skyscrapers of the 1970s stand side by side with churches of the 1600s. High-powered automobiles share modern expressways with mule carts. These and other unlikely contrasts are evidence of the city's successful transition from the 17th to the 20th century in one giant leap. Beautiful artisan handicrafts, the world's finest emeralds, excellent international hotels and restaurants, and the scenic grandeur of its mountain landscape (including the 475-foot falls of the Río Bogotá [Funza], at Tequendama 20 miles away) have made Bogotá a prime tourist attraction in South America. The area of the city is 46 square miles (120 square kilometres), and its population in 1971 was 2,500,000. The Special District of Bogotá, consisting of the city of Bogotá and six other cities or boroughs (Bosa, Engativá, Fontibón, Suba, Usaquén, and Usme) has an area of 600 square miles.

History. In April 1536, Gonzalo Jiménez de Quesada, an Andalusian lawyer and writer, set out from Santa Marta on the Caribbean Sea with more than 900 men to explore the sources of the Río Magdalena. In two years his company, drastically reduced, reached the cooler slopes of the central cordillera. Once upon the mountain plateau, they approached Bacatá, the main seat of the Chibcha Indians. Tribal resistance proved no match for Quesada's men, who soon enslaved the Chibchas, burned their temples, and carried off their gold.

In 1538 a civil government was organized with two mayors and a city council of seven members who served under a lieutenant governor. The settlement was christened Santa Fé de Bacatá: Santa Fé after Quesada's birthplace in Spain, and Bacatá for the original Indian name, which was soon corrupted to Bogotá. Bogotá's fate was closely allied to that of the viceroyalty of Nueva Granada, of which it was made the capital city, soon becoming a centre of Spanish colonial power in South America. Once the territory achieved independence from Spain in 1819, Bogotá was made capital of Gran Colombia, a confederation that included the present republics of Venezuela, Ecuador, Panama, and Colombia. During the troubled history of the confederation and down to the present time when only the Republic of Colombia remains of the original territory, Bogotá has remained the capital city.

Origin of name

Turbulent struggles for political power in the capital city, as well as its geographical isolation, stunted Bogotá's growth and prosperity in the 19th century. Development proceeded at a modest pace prior to World War II. In April 1948, following the assassination of the leader of the Liberal Party and likely presidential candidate Jorge Eliécer Gaitán, the central part of the city was severely damaged by riots. A wave of chaotic violence subsequently known as the *bogotazo* swept the city immediately and spread to the Colombian countryside. After the invocation of martial law in the city, the rioting subsided, and rural dwellers still threatened by uncontrolled violence fled to the city for security. Undercurrents of unrest existed in Bogotá until 1958 when the Liberal and Conservative parties reached a settlement and most of the violence in the country was quelled.

The contemporary city. *The city site and plan.* The city lies on a high sloping plain at the western base of two mountains, Guadalupe and Monserrate, upon the crests of which stand two imposing religious shrines. From a broad avenue at the base of the mountains, streets slope downward to the west, carrying streams of cold water from the mountains. North–south streets cross these at right angles, the blocks forming large terraces. Aside from a few avenues (which have proper names), all of the streets are laid out in this gridiron pattern—those running east to west being known as *calles* and those going from north to south as *carreras*. The city today extends nine miles from the southern suburbs of Bello Horizonte and San Isidro to the extreme northern ones of Santa Bárbara and Usaquén. The old municipal area is still the centre of the city and the site of the federal government. This central section merges naturally with the poorer residential suburbs to the south and the commercial and business districts to the north, beyond which are the richer suburbs, becoming progressively more expensive the further they are from the city's centre. Traditional suburbs such as Teusaquillo, Chapinero, and Chico were originally separate towns that the city engulfed. Today, there are no clear boundaries between suburbs. Many businesses and most of the finer shops have located in the northern suburbs. Until the mid-1960s, Carrera 56, three miles from the eastern edge of the city, was considered to be the western extremity of Bogotá, but suburbs have continued to sprout up further west as far as the new Eldorado Airport, about seven miles from the eastern edge. These areas include low and middle income housing developments such as Minuto de Diós and Bonanza. Several plazas, often dominated by churches, are located throughout the city.

Suburbs

Climate. Although Colombians consider Bogotá to be in the country's cold zone, its mean annual temperature is about 58° F (14° C) and averages less than a 2° F (1° C) difference between the coldest and warmest months. Rainy seasons occur in April and May and from September to December. During these periods the mountain plateau is subject to moderate but often continuous rainfall. Temperate cereals, vegetables, and fruits are plentiful year round on the savanna ringing the city and provide half of its food supply.

Transportation. Until airplanes flew over its surrounding wall of mountains, Bogotá was one of the least accessible capitals in the world. Today it is the hub of air travel in Colombia and the home of Avianca (Aerovías Nacionales de Colombia), the first commercial airline in South America. Railroads connect Bogotá with the Caribbean coast and to sections of the departments of Boyacá and Santander to the east, but any trip by rail to Colom-



The cathedral on the Plaza Bolívar, Bogotá.
Hector Acebes—Photo Researchers

bia's western cities and the Pacific coast must be made via Puerto Berrío on the Río Magdalena. Bogotá is on the Colombian section of the Pan-American and Simón Bolívar highways and has road connections with all major Colombian cities.

Although automobiles are expensive in Colombia, Bogotá is jammed with them. In recent years traffic arteries have been widened in the downtown area, bypassing heavily congested sections, and a highway now crosses the city from south to north, built as part of the massive road improvement that began in 1968 in preparation for the visit of Pope Paul VI to Colombia in 1969. The city's public transportation consists of several interconnecting bus routes, metered taxicabs, and *colectivos*, or cabs with a fixed route and fare.

Population. Bogotá's population has increased fourfold in 20 years, from 660,280 in 1951 to 2,539,100 in 1971. The urban slums in the south of Bogotá testify to the scourge of that rapid growth. It has been estimated that more than 50 percent of the city's inhabitants were born elsewhere but migrated to the city in search of better living conditions. Substantial numbers of Indians, mestizos (persons of mixed white and Indian ancestry), and mulattoes (persons of mixed white and Negro ancestry) are represented in the population. Caste and class are still dominant features of Bogotá society, yet boundaries have tended to blur, and some social mobility exists.

Close to Spain in its religious orientation, Bogotá is one of the world's strongholds of orthodox Roman Catholicism, and has been an archiepiscopal see since 1564.

Architecture and housing. Its classical cathedral and several other early churches in the Baroque and Rococo style preserve some of the finest examples of colonial architecture in South America. Today, almost every architectural period and style exists in Bogotá. Public buildings of the late 19th century follow the Neo-Gothic, and recent office buildings exhibit the clean, functional lines of the International style. In spite of expanded building programs, housing in the city was unable to keep up with the influx of squatters and settlers beginning in the late 1940s, and thousands of such residents still live in squalid slums.

Economic life. Although Bogotá is the commercial, cultural, and political centre of Colombia, it does not completely dominate a dependent interior like the capital cities of most Latin American countries. It is the home of the country's tire, chemical, and pharmaceutical industries and accounts for 50 percent of the country's industry in volume of sales, but its chief activities are commercial. Over 30 banks have their main branches in Bogotá, including the central bank, the Bank of the Republic. A stock exchange was established in 1928.

Government. In 1954 Bogotá was made a Special District for administrative purposes, while it remained the capital of the Department of Cundinamarca. Bogotá is governed by a mayor appointed by the governor of the department. The mayor chooses the municipal secretaries of government, treasury, and public works. Together, they carry out the laws and orders of a popularly elected municipal council. Bogotá is also the seat of the department's elected assembly. It is the site of the national capitol, and many government offices and agencies are located in the city.

The Bogotá Special District

Utilities, health, and safety. Although the growth of electric power and water supply in Bogotá has made prodigious increases in the last 20 years, it has still not been enough to keep up with the mushrooming population. Power has been rationed at various times since the 1960s, and temporary blackouts are not uncommon. A pipeline to bring Bogotá fresh water from the neighbouring Department of Boyacá was scheduled for completion in 1972.

Bogotá has several public and private hospitals and many outstanding clinics. All employers are required by law to supply free health care and free group life insurance for their employees.

Since 1961 all civil police units in Bogotá have been staffed by the national police. The city is also the headquarters for the Department of Administrative Security (DAS)—a national security investigation force. Bogotá has a volunteer fire department with selected neighbourhood substations.

Education. Education receives a high priority in Bogotá, although standards of quality are not uniform. There is a plentiful supply of private schools, most of which are distributed in the northern suburbs and many of which are run by religious orders. The number of public schoolrooms has been increasing, and the city's goal is to provide elementary and secondary education to anyone who seeks it.

Higher education

Of Bogotá's several excellent universities, the Academia Javeriana and the Universidad de Santo Tomas de Aquino held the standings of public universities as far back as the early 17th century.

The media. Bogotá has five morning and two evening newspapers, each representing one of the major political parties. There are numerous local radio stations and three television stations. Of the two government-owned television stations, one is used for classroom educational programs during the day and for adult education at night. All media are government controlled.

Recreation and culture. Beautiful, spacious parks adorn the city and its outskirts. The city is filled with motion-picture theatres, for movies are the most popular form of public entertainment. A prime attraction for Bogotanos and tourists are the tram and cable car that climb more than 1,800 feet to the church and shrine atop Monserrate. Soccer, bullfighting, and automobile, bicycle, and horse racing always draw enthusiastic spectators to scheduled events.

Traditional cultural institutions in Bogotá, aside from the universities, include the Botanical Institute, the National Conservatory of Music, the National Museum, the National Astronomical Observatory, National Library, and the Columbus Theater, a showcase for opera, ballets, and plays of national and foreign companies. In recent years the city has been enriched by the establishment of another excellent library, a planetarium, a museum of natural history, several modern art galleries, and the Gold Museum, which houses the world's largest collection of pre-Colombian gold objects.

BIBLIOGRAPHY. There are no existing books in print in English exclusively devoted to Bogotá. General sources of information with some information on the city include THOMAS E. WEIL *et al.*, *Area Handbook for Colombia*, issued by the U.S. Department of the Army, rev. ed. (1970), a thorough work with up-to-date statistics and information; C.H. HARING, *The Spanish Empire in America* (1963), a good treatment of the ecclesiastical and political development of Colombia during the colonial period; and R.H. DIX, *Colombia: The Political Dimensions of Change* (1967).

(J.Sa.)

Bohemia and Czechoslovakia, History of

Bohemia (Czech: Čechy; German: Böhmen) is a geographical and historical unit in central Europe, bounded on the south by Austria, on the west by Bavaria, on the north by Saxony, and on the east by Silesia and Moravia. It existed for many centuries as the largest and richest province of a kingdom of the same name. Its links with Moravia, established in the early Middle Ages, were closer than those with territories acquired by the kings of later days. In 1918 Bohemia was included in the Republic of Czechoslovakia as its westernmost province. Administrative divisions have never affected too deeply its structure and the life of its inhabitants. Natural boundaries following mountain ranges and other geographical features have usually coincided with the political frontier; attempts to split Bohemia or to reduce its size by annexation have never met with lasting success.

This article is divided into the following sections:

I. Bohemia to 1914
- Early history
  - Unification of greater Moravia
  - The Přemysl rulers of Bohemia (895–1306)
- The late Middle Ages (1310–1526)
  - The Luxembourg dynasty (1310–1437)
  - The Hussite preponderance (1437–71)
  - The Jagiellonian kings (1471–1526)
- Habsburg rule (1526–1914)
  - The Habsburgs to 1848
  - From absolutism to constitutionalism (1848–1914)

II. Czechoslovakia since 1914
- The Republic of Czechoslovakia (1918–45)
  - The struggle for independence (1914–18)
  - The establishment of Czechoslovakia (1918–25)
  - Political consolidation (1925–35)
  - Moving toward the abyss (1935–38)
  - From Munich to the disruption of the republic (1938–39)
  - The struggles at home and abroad (1939–45)
- Czechoslovakia since 1945
  - The uneasy interim (1945–48)
  - The People's Democracy (1948–60)
  - Reform endeavours (1960–67)
  - The "Prague Spring"
  - Normalization and consolidation (1968–71)

I. Bohemia to 1914

EARLY HISTORY

The prehistoric people north of the Middle Danube were of uncertain origin. The Boii, a Celtic people, left distinct marks of a fairly long stay, but its time cannot be firmly established. The Latin name of the country, from which its names in major Western languages are derived, is of Celtic origin. The Celtic population was supplanted by Germanic tribes. One of them, Marcomanni, inhabited Bohemia; others settled in adjacent territories. No outstanding event marked the Marcomanni departure.

Movements of ethnic groups

Archaeological discoveries and incidental references to Bohemia in written sources indicate that the movements of ethnic groups were not always abrupt and turbulent but that the new settlers began to enter the territory before the earlier inhabitants had left it. It can be assumed, therefore, that the Slavic people were coming in groups before the southward migration of the Germanic tribes. In the 6th century Bohemia and the neighbouring territories were inhabited by the Slavs.

While mountains and forests offered protection to Bohemia, the tribes in the lowlands north of the Danube and along its tributaries were hard pressed by the Avars of the Hungarian plains. Attempts to unite the related Slavic tribes for resisting the Avars were successful only when directed by such strong personalities as the Frankish merchant Samo, who gained control of a large territory in which at least part of Bohemia was included. His death in 658 ended the loosely knit state. A more auspicious era dawned after Charlemagne defeated the Avars in the late 8th century.

There followed a period of comparative security, in which the concentration of the Slavs into political organizations advanced more promisingly. Soon after 800 three areas emerged as potential centres: the lowlands along the Nitra River; the territory on both sides of the Lower Morava (German: March); and central Bohemia, inhabited by the tribe of the Czechs. In time, the Czechs, protected from foreign intruders, rose to a dominant position. Governed by rulers descended from the mythical plowman Přemysl and his consort Libuše, the Czechs brought much of Bohemia under their control before 800 but failed to defeat the tribes in the east and northeast. Apart from occasional disturbances, such as Charlemagne's invasions (805), the Czech domain was not exposed to war and devastation, and little of the life there came to the notice of clerics who were recording contemporary events in central Europe.

Unification of greater Moravia. The Moravian Basin, through which ran a trade route from the Baltic to the Adriatic, became the scene of lively activities by princes of unknown origin. The first of them, Mojmír I, probably began to rule after Charlemagne's death (814) and maintained friendly relations with Louis I the Pious. Around 833 Mojmír attached the Nitra region to his domain. His successor (after 846), Rostislav, consolidated the country and defended it successfully. His relations with the Eastern Frankish Empire (since 843 under Louis the German) were determined by political considerations and by the advance of Christianity into the Slavic areas. The bishoprics of Regensburg, Passau, and Salzburg competed in trying to convert the central European Slavs but achieved only limited success. The Archbishop of Salzburg consecrated a church at Nitra around 828. In 845 Regensburg baptized 14 chieftains from Bohemia. Mojmír's Moravia had apparently more frequent contacts with Passau than with Salzburg. Recent archaeological discoveries indicate that missionaries made noticeable progress before 860; stone churches were built as places of Christian worship at Mikulčice and elsewhere.

Competition of bishoprics

Religious conflicts. But Rostislav was dissatisfied with the Frankish clergy and asked the Byzantine emperor Michael III for Slavic-speaking preachers. A group of clerics headed by two brothers of Greek origin, Constantine and Methodius, arrived from Constantinople in 863. They not only preached in Slavic but also translated the sacred books into that language and used them in divine services. To Constantine is attributed the creation of the first Slavic alphabet (Glagolitic). After some two and a half years the two brothers journeyed to Rome. There Constantine entered a convent, taking the name of Cyril; he died in 869. Methodius received the Pope's sanction for his work in Moravia and in Pannonia, Moravia's southern neighbour. The two territories were organized as a province and connected with the ancient archbishopric of Sirmium, restored by the Pope. Methodius angered the Frankish clergy, who regarded his archdiocese as their missionary field. He was captured and imprisoned until 873; he then returned to Moravia and put himself under the protection of Rostislav's successor Svatopluk. But relations between the ruler and the Archbishop were not harmonious. After Svatopluk's conciliation with the Franks at Forchheim (874), clerics of the Latin rite appeared again in Moravia, interfering with the Archbishop's work. In 880 Methodius obtained from Pope John VIII a formal sanction of his work, including the Slavic liturgy.

Political expansion. Svatopluk distinguished himself in the conduct of political affairs. After the death of Louis the German (876), he acquired large territories with Slavic population. He annexed some and left local princes who recognized his suzerainty in others. Such was apparently the case of the Czech prince Bořivoj I. Propagation of Christianity followed Svatopluk's advances. According to legends, Bořivoj was baptized by Methodius and then admitted clerics of the Slavic rite to his principality. But, while the Archbishop was engaged in missionary work in the annexed territories, advocates of the Latin rite, headed by a Frankish cleric, Wiching, bishop of Nitra, strengthened their position in Moravia. During Methodius' lifetime the Slavic clergy had the upper hand; but after his death (885) Wiching banned Methodius' disciples from Moravia and most of them moved to Bulgaria. Pope Stephen V reversed his predecessor's policy and forbade the Slavic liturgy.

Svatopluk continued his policy of expansion for several more years. But soon after 890 he made the East Frankish king Arnulf his enemy. Arnulf's expedition into Moravia in 892 opened a period of troubles, which increased when Arnulf made an alliance with the Magyars of Hungary. Svatopluk's successor, Mojmír II, tried unsuccessfully to protect his patrimony; sometime in 905–908 greater Moravia ceased to exist as an independent country.

The Přemysl rulers of Bohemia (895–1306). The Prince of Bohemia made an accord with Arnulf (895) and warded off the danger of invasion. The domain over which the descendants of Přemysl ruled from the Prague castle was, in the early 10th century, the largest unit in Bohemia. The tribal chieftains who opposed centralistic tendencies exercised control over the eastern and northeastern districts, but the extent of their power is not known. The most powerful of them, the Slavníks residing at Libice, remained defiant until the end of the 10th century.

Bohemia maintained close relations with neighbouring Bavaria. Both countries were threatened for several decades by the Magyars, and other developments in their vicinity also affected political and social life. The most important of these was the rise in Germany of the Saxon dynasty, which began with Henry I the Fowler; the imperial coronation of Otto I, in Rome (962), marked the restitution of the Holy Roman Empire, with which Bohemia was linked, thereafter, for many centuries. Bohemia's reorientation toward the Saxon dynasty began under the grandson of Bořivoj, Prince Wenceslas (Václav, ruled 921–929); it was symbolized by the dedication of a stone church at the Prague castle to a Saxon saint, Vitus. Both Slavic and Latin legends praise Wenceslas as a fervent believer but tell little about his political activities. He was murdered by his younger brother Boleslav. The legends present the murder as an outburst against Wenceslas' devotion to the new faith, but the conspiracy probably had also a political motivation.

Restitution of the Holy Roman Empire

Boleslav I (ruled 929–967) reigned as a Christian prince; his daughter married Prince Mieszko I of Poland and helped spread the gospel in that country. Boleslav attempted, unsuccessfully, to loosen the ties with the Saxon dynasty. Boleslav II (ruled 967–999) used friendly relations with the Pope and the Emperor to enhance his prestige. He attached new territories east of Bohemia to his father's annexations. In 973 a bishopric was founded in Prague and subordinated to the Archbishop of Mainz. The first bishop, Thietmar, was from the Saxon land but knew the Slavic language; he was succeeded in 982 by Adalbert (Vojtěch), a member of the Slavník family. Adalbert's promotion can be viewed as an attempt to harmonize relations between the Prague and Slavník princes, but that result did not materialize. Legends hint that Adalbert encountered considerable opposition when attempting to raise the standards of religious life in his diocese, and tension between the rival dynasties showed no signs of abating. In 895 Boleslav II moved against the Slavníks and broke their power. Adalbert enjoyed some initial success among the heathen Prussians on the shores of the Baltic but then suffered a martyr's death in 997.

Annexation of Moravia. Struggles among the descendants of Boleslav II plagued the country for about 30 years and reduced considerably its power. Most of the territories attached to Bohemia in the 10th century were lost. Prince Břetislav I, a grandson of Boleslav II, led a successful expedition into Moravia; he conquered only a minor portion of the former greater Moravia, but it was large enough to constitute a province, and it was linked from then on indissolubly with Bohemia. But the ambitions of Břetislav, who ascended the throne in 1034, ran higher. He invaded Poland in 1039 with only temporary success; he incurred the indignation of the German king Henry III and was forced to evacuate the conquered territory and to make an oath of fealty (1041). In the latter part of his reign, Břetislav cooperated with Henry III, thus protecting his domain against armed intervention.

The entire territory of Bohemia and Moravia was regarded as a patrimony of the Přemysl House, and no emperor attempted to put a foreign prince of his own choice on the throne. But the ruling family grew large, and after Břetislav's death (1055) it became entangled in competition for primacy. For about 150 years the course of public life in Bohemia was largely determined by dissensions among the adult princes, some of whom ruled in portions of Moravia under Prague suzerainty. The emperors and the landowning magnates exploited the conflicts to promote their selfish interests. A main problem was the absence of any strict law of succession; the principle of seniority usually conflicted with the reigning prince's desire to secure the throne for his oldest son.

The territory's minor obligations toward the emperors was a handicap under weak princes or when the male members of the ruling family were at odds, but a strong prince could turn friendly relations with the empire to his advantage. Břetislav's second son, Vratislav II (ruled 1061–92), as a compensation for services rendered, obtained from Emperor Henry IV the title of King of Bohemia (1085). Another able ruler, Vladislav I, gained the dignity of a cupbearer to the Emperor (1114), one of the highest court offices; and as its holder the Prince of Bohemia became one of the electors who chose the Holy Roman Emperor. Vladislav II (ruled 1140–73) participated in the campaigns of Frederick I Barbarossa in Italy. He was named king and crowned by the Emperor at Milan in 1158.

Active participation in imperial policies and military campaigns reduced markedly the Czech's isolation, caused by Bohemia's geographical position. Other contacts were made with foreign merchants and by clerics from abroad or travelling from Bohemia to Rome and to famous shrines. By the early 11th century the Latin rite prevailed. Cosmas of Prague, who recorded in his chronicle the history of Bohemia to 1125, was an ardent supporter of the Latin liturgy. Western orientation of the hierarchy and of the monastic orders was documented by the prevalence of Romanesque architecture, of which notable examples could be found in Prague and in the residences of appanage princes (lesser members of the ruling family). In social stratification and in economy the country reached such a degree of consolidation that it withstood, without serious damage, struggles that ravaged it in the late 12th century.

Frederick I Barbarossa helped foment discord among Přemysl's descendants. In 1182 he reduced the dependence of Moravia on the Prague princes and subordinated that province to the imperial authority. In 1187 he exempted the Prague bishop, a member of the Přemysl family, from the jurisdiction of the ruling prince and made the bishopric an imperial fief. These decisions had no lasting significance, however, and the Přemysl patrimony survived. The period of trials closed with Frederick's death (1190). Frequent subsequent changes on the imperial throne lessened the danger of intervention. During the same period the Přemysl family was reduced to one branch, so that the problem of succession lost its pressing importance. In 1198 Přemysl Otakar I received the royal title for himself and his descendants from one of the competitors for the imperial crown. A solemn confirmation occurred in 1212, when Frederick II issued a charter known as the Golden Bull of Sicily, which regulated the relationship between Bohemia and the empire. The king's obligations were reduced to a minimum, but as elector he was able to exercise perceptible influence, ranking first among the temporal members of the college of electors.

The Golden Bull of Frederick II

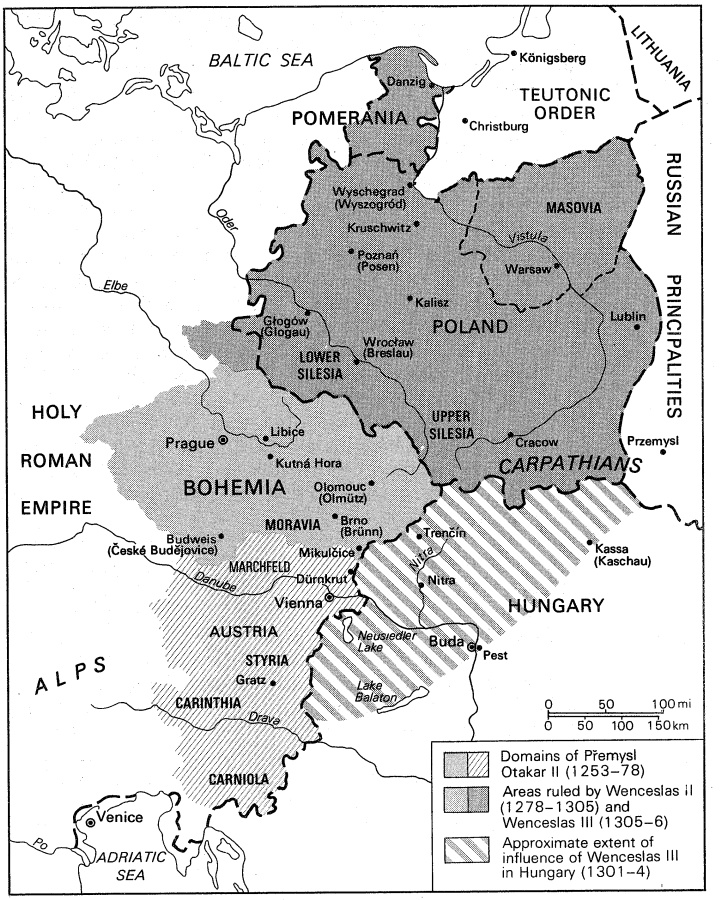
Political and economic growth. Under Otakar I and his successors, Bohemia moved from depression to political prominence and economic prosperity. The original socioeconomic structure was giving way to a more developed stratification. The clergy gained independence from temporal lords in 1221. The landowning class, made up of wealthy lords and less-propertied squires, claimed freedom in administering their domains and a more active role in public affairs. In the early 13th century the population of Bohemia and Moravia increased noticeably by immigration from overpopulated areas in

Germany. Many of the German-speaking newcomers were encouraged by the kings to found urban communities or to develop mining, especially of silver. The landowning magnates and the ecclesiastical institutions followed the royal example and settled the immigrants on their estates. There thus came into existence an urban middle class that enjoyed valuable privileges, especially the use of German law, and that became only slowly amalgamated with the native population. Apart from the townsfolk, German farmers moved into Bohemia and Moravia and transformed the less attractive border districts into prosperous areas. German immigration continued under Otakar I's successor, Wenceslas I (ruled 1230–53), and reached its peak under Přemysl Otakar II (ruled 1253–78). Bishop Bruno of Olomouc, in cooperation with the King, promoted colonization of large tracts in northern Moravia.

Otakar II was a strong and capable ruler who obtained possession of the Austrian lands through marriage and, in 1260, was invited by the nobility of Styria to become their lord. Personal bravery and financial resources facilitated his penetration into other Alpine provinces. Before his opponents could combine forces to check his advance, Přemysl Otakar II had exercised influence in Carinthia as well as in some territories along the Adriatic coast. His expansion aroused hostility of the kings of Hungary; but even more dangerous was Rudolf, count of Habsburg, who was elected king of the Romans in 1273 and, to get foothold in central Europe, claimed the Austrian lands as vacant fiefs of the empire. War followed and ended in Otakar's defeat in 1276. Unwilling to accept the loss of Austria as final, Otakar began a new campaign. Not only Rudolf's army but also Hungarian troops moved against the Czech forces, and a group of noblemen, mostly from southern Bohemia, sided with the enemy. Otakar was too weak to resist the unexpected coalition, and, on August 26, 1278, he lost, at Dürnkrut on the Marchfeld, both the battle and his life.

Otakar's only son, Wenceslas II (ruled 1278–1305), was too young to take over control immediately. During the period following Otakar's death (remembered as the evil years), Wenceslas was a mere puppet in the hands of ambitious lords; but in 1290 he emancipated himself from the tutelage and ruled with more success than had his father. The country was recovering quickly from both political and economic depression, and it again played an active role in international relations. Instead of resorting to wars, Wenceslas engaged in negotiations and soon achieved success in Upper Silesia. This was a prelude to his penetration into Poland, which culminated in 1300 with his coronation as its king. Diplomatic dexterity and enormous wealth quickly enhanced Wenceslas' prestige. In 1301 he was considered a candidate for the vacant throne of Hungary, but he recommended his son as a candidate instead. Meanwhile, the mines in various parts of the country, especially at Kutná Hora, yielded so much silver that the king was able to reform the monetary system and issue coins (grossus), which soon circulated within and outside his kingdom. But Wenceslas II's acquisitions were lost soon after his death; his son Wenceslas III took over Bohemia but was assassinated on his way to Poland (1306). Thus ended the long rule of the Přemysl family.

The end of the Přemysl dynasty



Přemyslid expansion under Přemysl Otakar II and his successors (1253–1306).

From *Grosser Historischer Weltatlas*, vol. II, *Mittelalter* (1970); Bayerischer Schulbuch-Verlag, Munich

THE LATE MIDDLE AGES (1310–1526)

The Luxembourg dynasty (1310–1437). After a four-year struggle for the throne, the Bohemian magnates decided for John of Luxembourg, son of Henry VII, the king of the Romans. John, only 14, married Elizabeth (Eliška), the second daughter of Wenceslas II. John confirmed the freedoms that the Bohemian and Moravian nobles had usurped during the interregnum and pledged not to appoint aliens to high offices. But a group of advisers, headed by Archbishop Petr of Aspelt, followed John to Prague and endeavoured to uphold the royal authority. In the resulting conflict a powerful aristocratic faction scored a decisive victory in 1318. Its leader, Jindřich of Lípa, virtually ruled over Bohemia until his death in 1329. John found satisfaction in tournaments and military expeditions, and he attached to Bohemia some adjacent territories; the extension of suzerainty over the Silesian principalities was his most significant achievement. He was assisted late in his reign by his oldest son, Wenceslas, who was brought up at the French royal court, where he changed his name to Charles. Charles endeavoured to raise the prestige of the monarchy but was hindered by John's jealousy and by lack of cooperation among the nobility. In 1346 both John, then blind, and Charles joined the French in an expedition against the English. John fell at Crécy.

The reign of Charles I. John and Charles benefitted from friendly relations with the popes at Avignon. In 1344 Clement VI elevated the See of Prague and made Arnošt of Pardubice its first archbishop. Clement VI also promoted the election, in 1346, of Charles as the king of the Romans. In Bohemia, Charles ruled by hereditary right. To raise the prestige of the monarchy, he cooperated with the nobility and the hierarchy. He made Bohemia the cornerstone of his power and, by a series of charters (1348), settled relations between Bohemia, Moravia, and other portions of his patrimony. He acquired several territories in the vicinity at opportune times by purchase or other peaceful means. At the end of his reign, four incorporated provinces existed in union with Bohemia: Moravia, Silesia, and Upper and Lower Lusatias. Charles also confirmed earlier documents defining the position of Bohemia in relation to the empire. In 1355 he was crowned emperor in Rome as Charles IV. After consultation with the electors, Charles issued the Golden Bull, which readjusted the problems of the empire, especially the election of the emperor.

Coronation as Charles IV

Under Charles, Prague became headquarters of the imperial administration. By the foundation of a new district (*nové město*), Charles facilitated expansion of the city and a rapid increase in its population; about 30,000

persons lived there by the latter part of his reign. In 1348 he founded in Prague a university with four traditional divisions (theology, law, medicine, and liberal arts); its members were grouped into four nations (Bohemian, Bavarian, Saxon, and Silesian-Polish). Prague attracted scholars, architects, sculptors, and painters from France, Italy, and German lands; the most distinguished among them was the architect Petr Parléř (died 1399), a native of Swabia. The flourishing of the late Gothic architectural style left a deep mark on both the royal residence and the countryside. Under Charles, Bohemia was spared entanglements in war and reached a high level of prosperity, shared by the upper classes and the peasantry. Charles was anxious to save the power and possessions accumulated since 1346. He succeeded in getting his son Wenceslas crowned king of the Romans in 1376. He also made provisions for dividing the Luxembourg patrimony, with the understanding that its male members would respect Wenceslas as their head. After Charles's death (1378) a smooth transition to Wenceslas' reign appeared to be assured. The country mourned Charles as "the father of the country."

Opposition to Wenceslas IV. His heir began to rule, without opposition, as Wenceslas IV. Though not without talents, he lacked his father's tenacity and skill in arranging compromise, and in less than a decade the delicate balance between the throne, the nobility, and hierarchy was upset. In a conflict with the church, represented by Archbishop Jan of Jenštein, the King achieved temporary success; the Archbishop resigned and died in Rome (1400). The nobility's dissatisfaction with Wenceslas' regime was serious, mainly over the selection of candidates for high offices, which wealthy families regarded as their domain and to which Wenceslas preferred to appoint gentry or even commoners. The struggle was complicated by the participation of other Luxembourg princes, especially Wenceslas' younger brother Sigismund. The nobles twice captured the King and released him after promises of concessions. But Wenceslas never took his pledges seriously and the conflict continued. Simultaneously with the troubles in Bohemia, discontent with Wenceslas was growing in Germany. In 1400 the opposition closed ranks; the Electors deposed Wenceslas and elected Rupert of the Palatinate as emperor.

Religious reform movement. The turn of the century was a watershed in reform endeavours in Bohemia. The movement arose around 1360 from various causes, one of which was the uneven distribution of the enormous wealth accumulated by the church in a comparatively short time. Moral corruption infected a large percentage of the clergy and spread also among the laity. Prague, with its large number of clerics, suffered more than the countryside. Both the King and the Archbishop showed favours to zealous preachers like Conrad Waldhauser and Jan Milíč of Kroměříž, but exhortations from the pulpit failed to turn the tide. The Great Schism in Western Christendom after 1378 weakened the central authority. Disharmony between Wenceslas and Jan of Jenštein, archbishop of Prague, hindered the application of effective remedies. In the late 14th century the reform movement was centred at Chapel Bethlehem (Betlémské Kapli); its benefactors stressed preaching in Czech as the main duty of its rector.

Appointment of Jan Hus to Bethlehem

The second, more dramatic, period of the reform movement began with the appointment in 1402 of Jan Hus to the Bethlehem pulpit. A scholar, he combined preaching in Bethlehem with academic activities and thus was able to reach the Czech-speaking masses and to group around himself scholars and students dedicated to the idea of reform. The university was split, because foreign members followed the conservative line. Another cause of division was the popularity of John Wycliffe, an English ecclesiastical reformer, among the Czech masters and students. Hus did not follow Wycliffe slavishly but shared with him the conviction that the Western church had deviated from its original course and was in urgent need of reform. Hus enjoyed the goodwill of Archbishop Zbyněk Zajíc of Hazmburk. But the atmosphere in Prague deteriorated rapidly; the German members of the university allied with Czech conservative prelates, led by Jan Železný ("the Iron"), bishop of Litomyšl. Because Wenceslas favoured the reform party, its opponents pinned hopes on Sigismund, king of Hungary; Wenceslas was childless, and Sigismund had a fair chance of inheriting Bohemia.

In the winter of 1408–09, a strong group of cardinals convened a general council at Pisa, which deposed the two rival popes and elected Alexander V to fill the vacancy. Wenceslas sympathized with the cardinals and invited the university to join him. When the Germans did not respond favourably, he issued, in January 1409 at Kutná Hora (Kuttenberg), a decree reversing the traditional distribution of votes. Thereafter, the three "foreign" nations had one vote and the Bohemian nation had three. The Germans rejected the decree and moved to Leipzig, where some of them unleashed a polemical campaign attributing to Hus more influence on the King than he actually had and depicting him as the chief champion of Wycliffe's ideas. Meanwhile, Alexander V issued a bull virtually outlawing Hus's sermons in Bethlehem and authorizing rigid measures against discussion of Wycliffe's ideas. Hus and his collaborators continued their activities. Neither Wenceslas nor any of the Czech prelates were experienced enough to achieve reconciliation between the church authorities and the reform party, and Bohemia was drawn into a sharp conflict. In 1412 Pope John XXIII became involved in a war with the King of Naples and offered indulgences for contributions to the papal treasury. When Hus and his friends attacked the questionable practices of papal collectors in Prague, John put Prague under interdict. Hus, hit by the sentence of excommunication, left Prague and moved to the countryside under the protection of benevolent lords.

In 1414 the Pope, acting in harmony with Sigismund, since 1411 the king of the Romans, called a general council to Constance. Hus went there hoping to defend himself against accusations of heresy and disobedience. A safe conduct from Sigismund did not protect him in Constance. Late in November he was imprisoned and was kept there even after the Pope, who had lost control of the council, had fled and been condemned by the cardinals. In the spring of 1415, Hus was called three times before the council to hear charges, supported by depositions of the witnesses and by excerpts from his own writing. The council paid no attention to Hus's protests that many of the charges were exaggerated or false. Hus refused to sign a formula of abjuration; he was then condemned and handed over to temporal authorities for execution. He was burned at the stake on July 6.

Execution of Hus

Some scholars reduced to a small number the points on which Hus had deviated from the official doctrine. But his followers, not interested in doctrinal subtleties, reacted emotionally against the council, Sigismund, and the conservative clergy. A wave of indignations swept over Bohemia and Moravia, and this movement, taking the name Hussite from the martyred leader, grew rapidly. A letter of protest, signed by 452 members of the nobility, was dispatched to Constance in September 1415. The condemnation and burning of Hus's friend Jerome of Prague (May 1416) increased the discontent.

Hus had not evolved a system of doctrine nor had he designated his successor. The most faithful of his disciples, Jakoubek of Stříbro, was not strong enough to keep the movement under his control. Ideological differentiation set in and resulted in divisions and polemics. The moderate Utraquists were entrenched in Prague; the radicals came mostly from smaller boroughs and the countryside. The Germans in Bohemia and in the incorporated provinces remained faithful to the church, and, thus, the deep-seated ethnic antagonism was accentuated.

Struggle between Sigismund and the Hussites. After the death of Wenceslas IV (1419), political issues gained in importance. The Hussites were resolutely opposed to Sigismund, but the Czech Catholics and the Germans were willing to recognize him. Sigismund, determined to break the Hussite opposition, initiated a period of bitter struggles that lasted more than ten years. Sigismund had

the support of opponents of Hussitism within the kingdom, of many German princes, and of the papacy. Invasions of Bohemia assumed the character of crusades but were pushed back by the Hussites, who pulled together in time of danger.

The moderate Utraquists and the radicals reached agreement on the fundamental articles of their faith. The radicals built themselves a centre, given the biblical name Tábor. The accord, concluded in 1420 in the nation's capital, became known as the Four Articles of Prague; it stressed that: (1) the word of God should be preached freely; (2) the communion should be administered in both kinds to clerics and laymen; (3) worldly possessions of the clergy should be abolished; (4) public sins should be exposed and punished. A wide range of disagreements between Prague and Tábor was left open and often resulted in mutual accusations and embitterment. A third party arose in northeastern Bohemia, around a newly founded centre at Oreb, but it had a much smaller following than either Prague or Tábor.

Meetings were held at which attempts were made to give the country a national government; the most significant was an assembly at Čáslav (June 1421). A regency council was set up, but it lacked sufficient authority; and the virtual master of the country was the leader of the "warriors of God," Jan Žižka. He was originally attached to Tábor, but he became disgusted with the endless disputes of its theologians and left the radical stronghold to organize a military brotherhood in northeastern Bohemia (1422); its members became so devoted to Žižka that after his death (1424) they called themselves Orphans.

Žižka strove tenaciously for two goals—the protection of Bohemia from Sigismund and the suppression of the enemies of the law of God within Bohemia and Moravia. He scored brilliant victories in battles against Sigismund's forces but could not unite the country under his banner. A Catholic minority, stronger in Moravia than in Bohemia, resisted the overtures of the Hussite theologians and Žižka's attacks. After Žižka's death, his heirs, headed by Prokop Holý the Shaven, lost interest in protracted warfare with Catholic lords and undertook instead foraging raids into the German territories bordering on Bohemia. But, whenever a crusade menaced Bohemia, the radical military brotherhoods joined the conservative forces to push back the invader. The last encounter at Domažlice was bloodless; the crusaders fled in panic upon hearing of the Hussite strength.

Meanwhile, a general council of the church met at Basel in 1431 and determined to find a peaceful settlement. At a conference at Cheb (1432), the delegates from Basel and Hussite spokesmen resolved that in controversial matters "the law of God, the practice of Christ, of the apostles and of the primitive church" would be used to determine which party holds the truth. The Hussite envoys reached Basel and opened debate on the cardinal points of their doctrine. It soon became clear, however, that the council was unwilling to abide by the Cheb agreement and that theologians representing the Tábor and Orphan brotherhoods would not acquiesce to a lean compromise. A drastic change occurred in Bohemia in 1434. In a fratricidal battle at Lipany in May, combined Catholic and Utraquist forces defeated the radicals and took over the control of the country. Under the leadership of Jan Rokycana, dealings with the council advanced markedly. The final agreement came to be known as the Compacts (Compactata); it followed the Four Articles of Prague but weakened them with subtle clauses (*e.g.*, the council granted the Czechs the communion in both kinds but under vaguely defined conditions). After the promulgation of the Compacts (July 1436), an agreement followed with Sigismund. But he died in 1437, and Bohemia was neither united in religion nor consolidated politically.

The Compacts

Various forces hindered religious pacification. The Catholic clergy refused to respect the Compacts, because they were not sanctioned by the pope, and would not accept Rokycana as archbishop. The radical parties, although gravely weakened at Lipany, existed as an uncompromising opposition in Rokycana's rear. His bid for recognition was also defied by the right Utraquist wing, which had seized the key positions during Sigismund's brief reign.

The Hussite preponderance (1437–71). Sigismund had no son, and the problem of succession caused a split among the nobility, which had been enriched during the revolutionary era by the secularization of church properties and had grown accustomed to the absence of monarchy. The conservatives accepted Sigismund's son-in-law Albert II of Austria, but the more resolute Hussites favoured a Polish candidate. Albert's death in 1439 ushered in another interregnum. In January 1440 an assembly was held to set up provincial administration for Bohemia; its composition demonstrated clearly the steady rise in the importance of the wealthy lords, functioning as the first estates. The lesser nobility, recognized as the second estate, was numerous; although the percentage of Catholics among the lords was rather high, the second estate was predominantly Hussite and conscious of its contributions to the Hussite defense. The upper classes recognized the royal boroughs as the third estate but were more and more reluctant to share power with them. In the January assembly the political alignments were not identical with religious divisions. Some moderate Catholics cooperated with the Utraquist majority, headed by Hynek Ptáček of Pirkštejn; a group of conservative Utraquists joined the Catholic lords, among whom Oldřich of Rožmberk held the primacy. The actual leader of the conservative bloc was Menhart of Hradec, nominally an Utraquist. No one was elected governor of Bohemia. Instead, in the counties into which Bohemia was subdivided, leagues were organized to promote cooperation of local lords, knights, and royal boroughs, irrespective of religious orientation.

The problem of succession became urgent when Albert's widow, Elizabeth, gave birth to a boy, baptized Ladislas and called Ladislas Posthumus. Several foreign princes showed an interest in the throne, but in 1443 the estates recognized Ladislas' claims. As he resided at the court of his guardian the German king Frederick III, the interregnum was extended. Ptáček, who headed the majority, died in 1444, and the party acclaimed George of Poděbrady as its leader. For several years the destinies of Bohemia were determined by efforts of Oldřich of Rožmberk and his allies to obstruct George's endeavours. Apart from political and economic stabilization, George strove for a papal sanction of the Compacts and for the confirmation of Rokycana as archbishop. George realized that Menhart's domination of Prague was a more serious obstacle than Rožmberk's intrigues; in 1448 George attacked and took Prague without bloodshed. Rokycana also entered the city and took over from the archconservatives, the Utraquist (or Lower) consistory. Although Frederick III was of the same religion as Rožmberk, he realized that an alliance with George would improve Ladislas' chances; in 1451 Frederick designated George as governor of Bohemia. From that position of strength, George moved energetically against both the Rožmberk coterie and the remnants of the radicals, entrenched at Tábor.

In October 1453 Ladislas was crowned king in the Church of St. Vitus, and George served as his chief adviser. Ladislas had been brought up as a Catholic, and German was his mother tongue. George hoped the King could re-establish Bohemia's connection with the incorporated provinces, especially the populous and rich Silesia. Ladislas died suddenly in November 1457. Several foreign princes competed for the throne, but the estates of Bohemia reaffirmed the elective principle and decided unanimously for George (March 1458).

Election of George of Poděbrady

Although attached to the Utraquist party, George endeavoured to rule as a king of "two peoples": the Utraquists and the Catholics; the Czechs and the Germans. He was anxious to be crowned according to the rites prescribed by Charles IV. George's son-in-law, King Matthias of Hungary, sent two bishops to Prague; George took a secret oath in their presence, by which he obliged himself to defend the true faith and to lead his people

from errors, sects, and heresies. Because the Compacts were not mentioned, George did not hesitate to make his pledge; since the agreement with the Council of Basel, the Utraquists considered the communion in both kinds as a lawful concession and not a heresy. Because both the election and coronation took place in Prague, George's principal concern was to have his title recognized by the estates of the incorporated provinces. He was mostly successful, but he had to accept the friendly help of papal envoys to get at least a provisional recognition by the Catholic and predominantly German city of Breslau (Wrocław) in Silesia (1459). During the next three years George enhanced his prestige both at home and abroad. Feeling that no lasting peace could be achieved without the speedy settlement of religious issues, George attempted in 1462 to have the Compacts sanctioned by Pope Pius II. Instead of approving the Compacts, the Pope declared them null and void. When informed of the Pope's action, George held a solemn assembly in Prague in August and affirmed his devotion to the communion in both kinds. Although neither the Pope nor the King showed any intention of retreating from his position, armed conflict was not inevitable, and several princes, including Frederick III, were willing to use their influence to arrange a compromise. But the new pope, Paul II, was elected in 1464 and soon adopted an aggressive policy that encouraged George's foes, especially the city of Breslau. A group of Catholic noblemen from Bohemia, headed by Zdeněk of Šternberk, formed a hostile league at Zelená Hora (1465) and entered into negotiations with Breslau and other Catholic centres. Shortly before Christmas 1466 the Pope excommunicated George and released his Catholic subjects from their oath of allegiance. In spring 1467 George's troops attacked the rebel forces. George was, on the whole, successful in desultory campaigns against the castles of the insurgents, but his position became more awkward in the spring of 1468 when Matthias of Hungary brought support to the Czech rebels. The Hungarians invaded Moravia and, by tying down a considerable portion of the royal army, they facilitated rebel successes in other parts of the kingdom. In May 1469 the opposition proclaimed Matthias king of Bohemia. In 1470 George achieved some successes over his rivals, but he was unable to consolidate them because of deteriorating health. He died on March 22, 1471, mourned both by the Utraquists and loyal Catholics.

Rejection of Compacts by the Pope

The Jagiellonian kings (1471–1526). Emperor Frederick III had observed benevolent neutrality. And George had also derived comfort from the friendly disposition of Casimir IV, the Jagiellonian king of Poland. Contacts with the Polish court continued after George's death and resulted, in May 1471, in the election of Casimir's son, known in Bohemia as Vladislav II, as king of Bohemia. Vladislav was supported by George's partisans irrespective of religious affiliation. George's foes adhered to Matthias, who possessed Moravia, Silesia, and the Lusatias. Vladislav's forces were not strong enough to defeat the rival, and an agreement concluded in 1478 enabled Vladislav to consolidate his position in Bohemia but left Matthias in temporary possession of the incorporated provinces. After Matthias' death (1490) Vladislav was elected king of Hungary (as Ulászló II); thus, he was able to reunite the incorporated provinces with Bohemia. Vladislav's successor was his only son, Louis, a sickly boy of 9 at his father's death.

Decline of royal authority

The reign of the two Jagiellonians was marked by a decline of royal authority. Vladislav II had been brought up as a Catholic and made no secret of his dislike of the Utraquist rites. But by his coronation oath he obligated himself to respect the Compacts. As long as Matthias was alive, Vladislav was supported chiefly by the Utraquists. After 1490 he spent more time in Hungary than in Bohemia, as did Louis. In this latter period the Catholic lords attached themselves to the royal court and exercised strong influence on public affairs of Bohemia.

The Jagiellonian era, when compared with the times of stronger monarchs, appears to have been an unbroken chain of aristocratic feuds and rivalries in which personal ambitions triumphed over patriotic sentiments, but a closer examination reveals brighter spots and concrete examples of constructive cooperation. The king stood aloof, and the Catholic and Utraquist factions of the estates concluded an agreement at Kutná Hora (March 1485) that reaffirmed the Compacts, recognized the existing divisions in Bohemia, and forbade attempts by either party to extend its sphere of influence at the expense of the other. The accord lasted until 1516 but was renewed in 1512 as "of perpetual duration." The Unity of the Czech Brethren, which had come into existence in 1457–58 as a new expression of Hussite rigourism, was not granted legal protection. In 1508 Vladislav II issued a stern decree, ordering persecution of the Unity, but it was not applied too rigidly.

The provincial diet rather than the royal court held primacy under the Jagiellonians, especially when the kings resided at Buda. The lords dominated the diet and were supported by the lesser nobility when attempting to limit royal power or when introducing restrictive measures against the lower classes. Both the mighty lords and the less propertied knights viewed with displeasure the political aspirations of the royal boroughs, their competitors in commerce and production. The diets passed several resolutions to remove the third estate from the positions acquired during the Hussite revolution. Because the boroughs obtained little help from the sovereign and his officers, the nobility encountered little resistance. A land ordinance adopted by the diet in 1500 limited considerably participation of the boroughs in the diet. The boroughs were also hit by several decrees, approved by the diet (especially those of 1487 and 1497), by which the landowners attached the peasantry to their estates. They thus reduced the possibility of migration into the towns and deprived the towns of cheap labour.

The boroughs, prosperous and self-confident, resisted the limitations and sought allies wherever they could be found. They obtained some concessions under Vladislav II, but a general compromise was made by the diet held in 1517 by which the boroughs joined concessions in political and administrative matters and surrendered some of the earlier privileges on which their economic prosperity was based. The higher estates tacitly recognized the right of the royal boroughs to participate in the diet as the third estate but reserved for themselves the positions on the board of provincial officers, including that of the vice chamberlain, who, in the king's name, supervised municipal administration. Although the boroughs gained some reasonable satisfaction, the landowning nobility was permitted to engage in production of articles that were previously the monopoly of the royal boroughs.

Compromise between the nobles and the boroughs

The agreement of 1517 did not end feuds and conflicts among the aristocratic factions and their partisans in the lower classes. In 1522 King Louis I left Hungary for Prague, intending to heighten the royal authority. With the help of loyal lords, in February 1523, he relieved Zdeněk Lev of Rožmitál of the office of supreme burgrave and appointed Karel of Minstrberk, a grandson of George of Poděbrady, to that key position in provincial administration. But, soon after the King's departure, Rožmitál resumed political activity and searched for allies. Religious controversies that flared up soon after Luther's attack on indulgences (October 1517) increased tensions in Bohemia. Rožmitál posing as a staunch supporter of the old faith, ingratiated himself with the King and regained his office. Louis, fully occupied with Hungarian affairs, was preparing for a campaign against the Turks. Meeting the Sultan's army with inadequate forces, Louis was defeated; he drowned in the marshes near Mohács while retreating from the battlefield (August 29, 1526).

HABSBURG RULE (1526–1914)

The Habsburgs to 1848. Ferdinand I of Habsburg, the husband of Louis' sister Anne, presented his claims to the vacant throne. He made substantial concessions to the Bohemian magnates and was elected king in October 1526; the coronation took place in February 1527. Fer-

dinand ruled also in other countries and, beginning in 1531, he assisted his brother, the emperor Charles V, in imperial affairs. After Charles's resignation (1558) Ferdinand was elected emperor. He considered Bohemia his most precious possession.

The dynasty and the estates (1526–1620). Early in his reign, Ferdinand was frequently absent; but, when in Bohemia, he endeavoured to dilute his precoronation pledges and curtail the privileges of the estates. He was obliged by the coronation oath to observe the Compacts and to treat the Utraquists as equal with the Catholics. But since 1517 Bohemia had been open to ideas emanating from Wittenberg and other Reformation centres. Lutheranism had adherents among the Utraquists and among the German-speaking inhabitants of Bohemia and Moravia. The Unity of the Czech Brethren resisted successfully repeated attempts at its extermination; although not protected by the Compacts, the Unity increased in numbers and was shielded by sympathetic landowners, some of whom became members. The teachings of radical reformers also had echoes in Ferdinand's domains.

An opportunity to settle controversial problems arose in 1547. During the Schmalkaldic War (1546–47), between the Habsburgs and the Protestant Schmalkaldic League, the estates of Bohemia pursued an inconsistent policy, and, after the Habsburg victory at Mühlberg (April 1547), Ferdinand moved quickly against them. The high nobility and the knights suffered comparatively mild losses. But the royal boroughs virtually lost their political power and were subordinated more rigidly to the royal chamber. Another target of the King's wrath was the Unity; significantly, Ferdinand's vindictive policy did not apply to Moravia, the estates of which were more cooperative during the Schmalkaldic War than those of Bohemia. After 1547 the Unity flourished in Moravia; its members, driven from Bohemia, moved to Moravia or emigrated to Poland.

Ferdinand's moves against the Bohemian estates

The Diet of 1549 approved Ferdinand's request that his firstborn son, Maximilian, be accepted as the future king. Ferdinand also resumed his scheme of religious reunion on the basis of the Compacts, but he soon realized that few Utraquists adhered to that outdated document. The majority, called Neo-Utraquists by modern historians, professed Lutheran tenets as formulated by Martin Luther's associate, Philipp Melanchthon. Disheartened by the meagre results of his policy, Ferdinand turned toward the Catholic party to consolidate its organization. He introduced the newly founded and militant Society of Jesus into Bohemia (1556) and obtained from Rome consecration of Antonín Brus of Mohelnice as archbishop (1561). Shortly before his death, Ferdinand succeeded in getting from Pius IV a sanction of the communion in both kinds, but the Pope insisted on so many restrictions that his bull satisfied only the Utraquist extreme right.

Maximilian II (ruled 1564–76) was reluctant to grant free exercise of the Lutheran faith, which the majority of the estates requested in 1571. After several years of futile efforts, the estates adopted a more flexible policy. Both the Czech Neo-Utraquists and the German-speaking Lutherans came together and prepared a summary of their faith, known as the Bohemian Confession; it agreed in the main points with the Augsburg Confession. The Brethren cooperated with the adherents of the Bohemian Confession but preserved both their doctrine and their organization. In 1575 Maximilian II approved the Bohemian Confession, but only orally; it was commonly assumed that his oldest son, Rudolf, who was present in the session, would respect his father's pledge.

The early stage of Rudolf II's (ruled 1576–1612) long reign was simply an extension of Maximilian's regime. But in 1583 Rudolf transferred his court from Vienna to Prague, bringing with him the high offices and foreign envoys; the Bohemian capital became once more an imperial residence and a lively political and cultural centre. Rudolf, brought up in Spain, had sympathy only for the Roman Catholic faith. Because the crown possessions were too small to yield adequate income, he depended mostly on the estates, whose majority was Protestant; only the provincial diets had the power to approve increased taxation and to grant subsidies for interminable wars against the Turks. The Catholic party, stronger among the lords than among the lesser nobility and burghers, came under the influence of militant elements, trained in Jesuit schools, and listened attentively to the papal nuncios and Spanish ambassadors. Because of its long antipapal tradition and its political prominence, Bohemia had an important place in the strategy of the Counter-Reformation. The Catholics singled out the Unity as their first target. Although numerically weak, the Brethren exercised a strong influence on Czech religious life and developed lively literary activities (in Rudolf's reign they produced a translation of the Bible from the original languages, which was printed in a hamlet of Kralice on the domains of the lords of Žerotín and which came to be known as the Kralice Bible). The Catholics sought to create a breach between the majority party of the Bohemian Confession and the Unity.

The Counter-Reformation in Bohemia

By a succession of new appointments, Catholic radicals around 1600 occupied the key positions in the provincial administration of Bohemia; their head, Zdeněk Vojtěch of Lobkovice, served as the supreme chancellor and enjoyed Rudolf's confidence. In 1602 Rudolf issued a rigid decree against the Unity, which was enforced not only in the royal boroughs but also on the domains of fervent Catholic lords. The Brethren and also the more resolute adherents of the Bohemian Confession realized that the days of peaceful coexistence were gone. They closed ranks under the leadership of Václav Budovec of Budov, a prominent member of the Unity. Dissatisfaction with Rudolf's regime was growing rapidly in other Habsburg domains. His younger brother, Matthias, made contacts with the Austrian and Hungarian opposition; the Moravian estates, headed by Karel the Elder of Žerotín, joined Matthias. In 1608 rebel forces advanced to Bohemia; Rudolf was unable to resist them, and he made peace and transferred to Matthias the dissatisfied provinces. The Protestant estates of Bohemia used Rudolf's weakness for their purposes. Although reluctant, Rudolf issued, in July 1609, a charter, known as *Majestát* (Letter of Majesty), that granted freedom of worship to the Catholics and to the party of the Bohemian Confession, with which the Brethren closely cooperated. Some passages of the charter were vague, and so the Protestant and Catholic estates concluded an agreement, stipulating that future conflicts should be settled by negotiation. The Catholic radicals, too weak to upset the agreement, were unwilling to accept the *Majestát* as the final word in religious controversies.

In 1611 Rudolf was deposed, and Matthias was crowned king of Bohemia. Because he was childless, the question of succession was debated both in the court circles and among the estates. In 1617 Matthias presented his nephew Ferdinand of Styria to the Diet of Bohemia as his successor. The resolute faction among the Protestant nobility was caught unprepared and acquiesced in Ferdinand's candidacy; he was accepted and crowned in the Church of St. Vitus. Opposition grew quickly to Ferdinand, who was suspected of cooperation with the irreconcilable opponents of the *Majestát.* In the spring of 1618 the Protestant estates decided on an action. Two governors of Bohemia, William Slavata and Jaroslav Martinic, were accused of violation of the *Majestát;* after an improvised trial they were thrown from a window of the Royal Chancellery (May 23, 1618) but escaped with minor injuries. The act of violence, usually referred to as the Defenestration of Prague, sparked a rebellion in Bohemia. The estates replaced the board, or royal governors, with 30 directors, who assembled troops for defensive purposes and gained allies in the predominantly Lutheran Silesia and in the Lusatias; the estates of Moravia were reluctant to join. The death of Matthias (March 1619) changed the situation profoundly. The directors refused to admit Ferdinand II into Bohemia. In Moravia the militant Protestant party overthrew the provincial government, elected 30 directors, and made an accord with Bohemia. At a general assembly of representatives of all five provinces, a decision was made to form a federal system. Ferdinand II was deposed, and

The Defenestration of Prague

Frederick V, elector of the Rhine Palatinate, a son-in-law of James I, king of England and Scotland, was offered the crown. He accepted and early in November 1619 was crowned king according to an improvised Protestant rite. Frederick's chances for success were slight; the population of Bohemia, especially the peasantry, was unenthusiastic in its support of the rebellion. Frederick received some financial help from the Netherlands, but German Protestant princes hesitated to become involved in a conflict with the Habsburgs, among whose allies were not only Catholic Bavaria but also Lutheran Saxony. In late summer 1620 Maximilian I of Bavaria coordinated the Catholic forces; although short, the battle on the White Mountain, at the gates of Prague (November 8, 1620), had a decisive effect and delivered Bohemia to Ferdinand II. Frederick and his chief advisers fled from Bohemia. Fighting continued in 1621 at some isolated places and in Moravia, but no one succeeded in pushing back Ferdinand's troops.

In imposing penalties Ferdinand treated Bohemia more harshly than the incorporated provinces. On June 21, 1621, 27 leaders (three lords, seven knights, and 17 burghers) were executed. Landowners who had participated in any manner in the rebellion had much of their property confiscated. The upper estates and the royal boroughs were ruined; they ceased to function as centres of economic and cultural activities. Ferdinand rescinded the *Majestát* and declared his intention to promote the program of re-Catholicization of Bohemia and Moravia. The Society of Jesus, banned in 1618 by the directors, returned triumphantly and acted as the vanguard in the systematic drive against the non-Catholics, including the moderate Utraquists.

Under absolutist rule (1627–1848). In 1627 Ferdinand II promulgated the Renewed Land Ordinance, a collection of basic laws for Bohemia that remained valid, with some modifications, until 1848; he issued a similar document for Moravia in 1628. Ferdinand settled, in favour of his dynasty, issues that had disturbed Bohemian public life since 1526: the kingdom was declared hereditary in both the male and female branches; the king had the right to appoint supreme officers; in the provincial diet the higher clergy was constituted as the first estate, and all the royal boroughs were represented by one delegate only; the diet lost legislative initiative and could meet only upon the king's authorization to approve his requests for taxes and other financial subsidies; the king could admit foreigners to permanent residence; the use of German besides the traditional Czech was authorized. No faith other than Roman Catholicism was permitted.

Royal decrees pertaining to religion granted the upper classes the right to choose either conversion or emigration. A rather high percentage decided for the latter and settled abroad, mostly in Saxony. Many peasants left the country illegally, especially during the Protestant invasions of Bohemia. The Czech's most significant representative abroad came to be a scholar, John Amos Comenius. The majority of the population remained in the homeland and gradually converted to the Roman faith.

Conversion of the Czechs to the Roman faith

The Jesuits became the most important force in Czech spiritual life. Their leading school, the Clementinum, was, in 1654, united with the remnants of the ancient Charles University. The Jesuits controlled not only higher education but also literary production. With an increasing number of Czech novices, the Society could reach the common people, whose majority spoke only the Czech language.

The vacated places among the upper social classes were gradually filled by newcomers, most of whom obtained land as a compensation for services rendered to Ferdinand II and his successor, Ferdinand III (ruled 1637–57); some enterprising individuals purchased land in Bohemia either during or after the Thirty Years' War (1618–48). The old families and the newcomers had in common their attachment to the Roman church and to the dynasty; they intermarried and became amalgamated over the next several decades. German became the language in which public affairs were transacted. But language was not the only barrier separating the peasantry and lower middle class from the propertied noblemen and burghers. Both the victorious church and the wealthy laymen regarded the Baroque style as the most faithful expression of their religious convictions and their worldly ambitions. For about 100 years, the Baroque dominated in architecture, sculpture, and painting and influenced literature, drama, and music. The external appearance of Prague and the smaller boroughs and towns changed markedly. In the countryside sumptuous aristocratic residences contrasted sharply with the modest dwellings of the peasantry.

Leopold I (ruled 1657–1705) soon became involved in long and costly wars against the Turks and the French. Although Bohemia was not threatened by either of his enemies, its population had to share the financial burdens; the landed nobility was reluctant to accept financial obligation, so the major part of the contributions was expected to come from the burghers and the peasants. The urban communities, which had been impoverished during the Thirty Years' War, made no progress toward social and economic recovery. The lot of the peasantry was so heavy that risings occasionally flared up, though with no chance of success. For the common people the short reign of Joseph I (ruled 1705–11) brought some relief, but under his brother and successor, Charles VI (ruled 1711–40), their plight reached appalling dimensions. The court and the residences of the ranking aristocrats consumed vast sums of money, which had to be squeezed from the depopulated towns and poorly managed domains. At this time, alienation of the masses of people reached its apex.

Retention of autonomy under Habsburgs

The Habsburgs, ruling over Bohemia from 1620 to 1740, did not insist on its close union with their other domains. The kingdom of Bohemia, though under an absolutist regime, retained its autonomy. The two Lusatias were ceded in 1635 to Saxony; Bohemia, Moravia, and Silesia retained their provincial administration. Members of the local nobility were appointed to high offices. The supreme chancellor of Bohemia served as a link between the kingdom and the sovereign and resided in Vienna to facilitate communication with the court and various central agencies attached to it.

Most of the reforms of Charles's daughter Maria Theresa (ruled 1740–80), although motivated primarily by dynastic interests, improved the living conditions of the population. Soon after her accession Bavaria and Prussia invaded the Habsburg territories. Charles Albert, elector of Bavaria, occupied a major part of Bohemia and was acclaimed king by a fairly strong party among the estates; but he could not establish himself permanently, and in 1742 he pulled his forces back. Three wars fought against Frederick II the Great of Prussia in 1741–63, mostly in Bohemia and Moravia, were more serious and costly; finally, Maria Theresa acquiesced in the loss of the major part of Silesia. Small duchies that she was able to retain were constituted as a crown land of Silesia and remained closely connected with Moravia and Bohemia.

Realizing that the system inherited from Charles VI was the main source of weakness, Maria Theresa launched a program of administrative reforms (1749); its principal point was a closer union of the Bohemian crown land with the Alpine provinces. The Queen's staunchest opponents were members of the landowning nobility who, up to that time, had controlled the provincial administration. In 1763 Maria Theresa made some concessions but would not abandon her centralistic policy. The opposition did not remain united. The conservative faction remained unreconciled to the new course, but more flexible individuals accepted high positions in Vienna or in the provincial capitals and helped to build up the system, which Joseph II (ruled 1780–90) inherited from his mother and subordinated more rigidly to the sovereign's will and discretion.

Maria Theresa, partly under the influence of her husband, Francis Stephen of Lorraine, had adopted the idea of curtailing the privileges of the upper social classes so as not to conflict with the interest of the state, of which the ruler was the supreme representative. Joseph II had grown up in this enlightened atmosphere, and, when con-

fronted with conservative opposition as king, he went far beyond his mother's limits. Apart from the administrative reforms, the judicial and fiscal systems were revamped to serve the enlightened monarch more adequately. The state extended its influence in such other fields as education, religion, landowner–tenant relationships, the economic recovery of the royal boroughs, and a more adequate distribution of the burden of taxes. The reforms did not aim at total abolition of social and economic distinctions, but they generally improved the lot of the lower middle class and of the peasant. Two decrees of 1781 made Joseph popular among the masses: he abolished restrictions on the personal freedom of the peasants, and he granted religious toleration. After the long period of oppression, these were hailed as beacons of light, although they did not go so far as enlightened minds expected. The edict of toleration in Bohemia and Moravia was not followed by a mass defection from the Roman church, partly because it did not refer to either the Utraquism or the Unity but authorized worship according to either the Augsburg or Reformed Confession.

Joseph's conservative successors, Leopold II (ruled 1790–92), Francis I (ruled 1792–1835), and Ferdinand V (Ferdinand I of Austria; ruled 1835–48), left intact the centralistic system inherited from Maria Theresa and Joseph II, but they did engineer gradual transition from the manorial system to the full ownership of land by the peasants. They made peace with the landowning nobility, seeing in it their most faithful ally; but the provincial diets of Bohemia and Moravia still had no more than a decorative function. A fairly large number of persons of rank distinguished themselves as patrons of learning, lovers of theatre and music, promoters of new and more profitable methods in agriculture, and, in the early 19th century, as pioneers of industry. In these activities they made contacts with gifted men of middle-class or of peasant origin, gave them financial support, and shielded them from the ubiquitous police and rigid censorship. Provincial loyalties were stronger than ethnic differentiation, which emerged as a new factor in Bohemia around 1800, partly out of opposition to the centralistic tendencies of the Vienna court, partly under the impact of the French Revolution. Institutions destined to play an important role in the Czech national renascence, such as the Royal Bohemia Society of Sciences or the National Museum (1818), were bilingual and drew support both from the propertied German population and from a small fraction of the Czechs who became conscious of their origin, of the brighter periods, and of their kinship with other Slavic peoples.

Czech national renascence

From absolutism to constitutionalism (1848–1914). In 1848 the German-speaking population of Bohemia and Moravia had a distinct advantage over the Czechs. The upper classes of these two provinces were almost entirely German and the rural areas in which, after 1620, the Germans gained predominance extended from the mountain ranges deep into the lowlands, once purely Czech. There were, however, limited opportunities for Czechs of middle-class or peasant origin, who prepared for more lucrative occupations through higher studies or who acquired special skills. Some improvement could be observed in the last stage of Habsburg absolutism, from about 1830 onward. The efforts of scholars, writers, clergymen, and schoolmasters, aware of their Czech origin, stirred a national consciousness among the common people. Not only the countryside but also the urban communities witnessed an awakening. The Habsburgs, symbolized by Prince Metternich, tolerated no political activities but did not hinder cultural activities, the printing and distribution of nonpolitical books in Czech, theatrical performances, and gatherings for other than political purposes. The Czechs had their social and intellectual elite, small in numbers but devoted to the national cause; they were shielded by a group of aristocrats, who thus manifested opposition to Vienna.

Similar conditions existed in the Hungarian counties inhabited by the Slovaks. Contacts between these two ethnic groups were hindered by the existence of provincial boundaries, but the groups were close enough to permit cultural exchanges. Up to 1840 the Czech language, regenerated by such eminent linguists as Josef Dobrovský and Josef Jungmann, was used by both Czech and Slovak authors. But the growing national awareness gave rise to endeavours to develop a literary language for the Slovaks in order to reach people with no more than elementary training. The literary Slovak gradually replaced the Czech. Thus, the mounting wave of nationalism, instead of cementing the traditional relationship, created conditions for differentiation and for the establishment of two closely related but distinct ethnic units.

Revolution of 1848. In opposing Metternich's oppressive regime, the Czech intellectuals were allied with the progressive forces among the Germans. When the revolutionary wave reached Bohemia and swept away its outdated institutions in March 1848, leaders of the two nationalities worked together in an attempt to shift from absolutism to constitutionalism. Both parties had a vague notion that Bohemia should return to its autonomous status and become a constituent part of the regenerated monarchy, but they could not resolve specific problems. The Germans saw advantages in cooperating with their kinsmen in other Habsburg lands; moreover, they were keenly interested in the idea of German unification, debated in the German constituent assembly at Frankfurt. The Czech voters looked to František Palacký as their leader; he had written several volumes of *A History of the Czech People* and was a respected political thinker. He was assisted by Karel Havlíček Borovský, a journalist, and by František Rieger, a student of political science and economics. The Czech leaders sensed danger in the schemes laid before the Frankfurt assembly and in plans for a modernized but highly centralized Austria. Their primary concern was the Diet of Bohemia, and, at times, they included among their desiderata a general assembly of deputies from Bohemia, Moravia, and Silesia, to stress a continuity of modern political efforts with the ancient kingdom. But, despite some support from the aristocratic circles, the Czechs were unable to change the movement toward centralization.

Movement toward centralization

A good deal of vacillation in and after 1848 was caused by the inability of Palacký and others to harmonize the emphasis of historical rights with genuine devotion to the principle of nationality. In late spring 1848 the idea of an elected diet for Bohemia was obscured by a loftier project, an assembly of spokesmen of the Slavic peoples from all parts of the Habsburg Empire. No matter how sincerely Palacký and other prominent figures professed their loyalty to the ruling house, the Slavic congress was viewed with displeasure by the Germans and the Magyars and was finally dispersed by the archconservative Alfred, Fürst zu Windischgrätz, who ordered that no election for the provincial diet could be held. The Czech leaders recognized that the constituent assembly meeting in July 1848 in Vienna was the only representative body before which they could express their aspirations. They participated in the late summer and early autumn sessions and worked with even more vigour when the assembly reconvened at Kroměříž (Kremsier). They made themselves allies of all factions that attempted to prepare the ground for a constitutional and federal system. Rieger, in particular, rose to the occasion when defending the principle that all power comes from the people. But the draft of a constitution for the Habsburg monarchy ran counter to ideas prevailing among the advisers of the new king Francis Joseph I (ruled 1848–1916). Early in March the Kroměříž assembly was dispersed. The Habsburg government, headed by Felix, Fürst zu Schwarzenberg, ruled for some time in accordance with a constitution drafted by the crown advisers; but, on December 31, 1851, Francis Joseph abolished the last vestiges of constitutionalism and began to rule as absolute master.

The regime, allied with the church and supported by the army, police, and bureaucracy, was rigid and effective but tolerated no opposition. Its weakness was revealed, however, by the poor show of its armies in a war with Sardinia in 1859. In October 1860 Francis Joseph issued a diploma burying the absolutist rule and inaugurating a constitutional era. But it soon became clear that no

scheme forwarded by the crown advisers could reconcile the federalist tendencies with the monarch's desire to concentrate as much power as possible in Vienna.

Division of Czechs and Slovaks. After a war with Prussia and Italy in 1866, Francis Joseph sought a solution that would promise speedy recovery and the stabilization of internal affairs. The monarchy was transformed, in 1867, into a dual system. The Magyars obtained the dominant position in Hungary; in the conglomeration of other provinces, which was briefly called Austria, the Germans were the strongest single group, followed by Czechs, Poles, and other nationalities. The dual system passed through successive crises but remained in existence until 1918.

Like other nationalities the Czechs resumed political activities after the promulgation of the October Diploma. Palacký was recognized as a dominant figure, but the actual leadership passed into Rieger's hands. Two courses were open to the Czechs: to apply the principle of nationality or to emphasize historical continuity. Palacký and Rieger decided for the latter and were supported by their conservative collaborators; clearly they had no chance for success without a close alliance with the conservative landed aristocracy, to which the electoral system granted a strong position in the provincial diets and in the parliament. But this alliance was exploited by Rieger's progressive opponents. Differentiation within the National Party began in 1863 and continued more rapidly after 1867. The Czechs, irrespective of ideological orientation, opposed the dual system and boycotted institutions that Austria received after the promulgation of a new constitution in December 1867. After two stormy years, an attempt was made to devise a solution that would give Bohemia autonomy within the Austrian half of the monarchy. In agreement with the historically minded nobility, Rieger negotiated in 1870 and 1871 with the Vienna cabinet and consented to a compromise. But Francis Joseph, although originally sympathetic, yielded to heavy pressure from many sides in October 1871 and refused to sanction the compromise. No attempt was made after 1871 to revive the project.

Internal differences in National Party

Despite the setback, Rieger was able to retain leadership for some 20 more years. Most official statements in either the Vienna Chamber of Deputies or in the provincial diets of Bohemia and Moravia contained a formal declaration in favour of the state right. The idea of restitution of the kingdom of Bohemia to its former rank, similar to that of Hungary, was never given up; but its chances of realization declined with the consolidation of the dual system, and Francis Joseph showed no intention of going to Prague to be crowned with the ancient crown of St. Wenceslas. After 1871 the Czech political leadership was confronted with a dilemma: whether to boycott the parliament and the diets or to join the government majority for concessions in education and economic life. In 1874 the National Party split; the progressive wing, commonly called the Young Czechs, gained in popularity among the urban middle class and well-to-do peasants. Rieger found it more and more difficult to defend his alliance with the big landowners, because it brought no tangible results and obstructed the flow of progressive ideas. The Young Czech deputies insisted on its dissolution and were applauded by their supporters, to whom progress in education, emancipation from clerical influences, and improvement of living standards were more vital than the continued emphasis on unforfeited state right. The Old Czechs lost ground in the 1880s and suffered a total defeat in the parliamentary election of 1891.

German–Czech rivalry. The most determined opponents of the state right scheme in 1871 and thereafter were the spokesmen of the German-speaking population of Bohemia and Moravia, later known as the Sudeten Germans, who realized the losses they would suffer with any decentralization of Austria. In the Vienna parliament they cooperated with their kinsmen from the Alpine provinces and helped determine the composition of the cabinets. An alliance between Austria-Hungary and Hohenzollern Germany (1879) increased their sense of belonging to one of Europe's strongest ethnic units. But their population was losing in Bohemia and Moravia in proportion to the Czechs. The losses were not spectacular and were largely neutralized by Vienna's reluctance to change the traditional practices of giving preference to German over Czech candidates in civil service and especially in the army. The electoral system for the provincial diet, introduced in 1861, was not changed, although the right to vote in parliamentary elections was extended several times to benefit less propertied voters. The immediate cause of Rieger's fall was dissatisfaction over concessions he was willing to make to the Germans in 1890. Thereafter, no attempt was made to achieve general agreement on problems of coexistence of the two ethnic blocs. The largest and richest crown land, in fact, became a trouble spot second, after 1908, only to the southern Slavic provinces.

But the Young Czech leaders were soon caught in the same dilemma that had plagued Rieger. Solemn declarations of adherence to the state right scheme were followed by bargaining with the prime ministers, who sought potential members of a government coalition and offered tempting concessions, including Cabinet posts. Graf Kazimierz Badeni, who headed the Austrian Cabinet in 1895–97, promised administrative measures that would sanction wider use of Czech in Bohemian civil service and law courts. But he encountered vigorous opposition, organized by German nationalists, in the parliament and lost the Emperor's confidence. He resigned, and his successor recognized the futility of trying to adjust the outdated laws in favour of the Czechs, whose members in relation to the Germans amounted to almost two-thirds.

Decline of the Young Czechs

The changing social and economic stratification also sped the decline of the Young Czechs. They unsuccessfully courted industrial workers, who were more attracted by the Social Democrats and voted for their candidates. Václav Klofáč, a talented journalist, after several years of cooperation with the Young Czechs, founded the National Socialist Party. The peasants, dissatisfied with the increasing influence of big business and the upper middle class, turned away from the Young Czechs after 1890. An agrarian movement soon became the Young Czechs' most dangerous rival, because the peasants predominated in the Czech-speaking areas of Bohemia and Moravia. The young Czech political program was pervaded by liberal principles, which included anticlericalism; that made it unpalatable to the conservative groups, which favoured close cooperation with the Roman Catholic Church and which were stronger in Moravia than in Bohemia. Finally, voters led in Moravia by Adolf Stránský and in both provinces by Tomáš Garrigue Masaryk came to feel that the Young Czechs were not seriously carrying out the progressive ideas included in their program. Parties that developed out of ideological opposition were small when compared with the Agrarians, the Socialists, and the Young Czechs; but their ideas reached the noncommitted voters. The grant of universal manhood suffrage in 1906 greatly improved the chances of parties appealing to the less propertied voters; instead of helping to consolidate the parliament, it caused such differences that the prime ministers, following each other in quick succession, found it increasingly difficult to form a solid majority block. Thus, from the election in 1907 to the outbreak of World War I in 1914, the Chamber of Deputies could easily be bypassed by the court and by the ministries of Foreign Affairs and War, over which Francis Joseph exercised stronger control than other constitutional rulers. The dual monarchy, instead of experiencing regeneration, was moving toward more dangerous involvements in international affairs and, finally, toward catastrophe.

II. Czechoslovakia since 1914

THE REPUBLIC OF CZECHOSLOVAKIA (1918–45)

The struggle for independence (1914–18). World War I brought about a total estrangement between the Germans and the Czechs and Slovaks within the country. The Germans lent full support to the war effort of the

Central Powers, but among the Czechs the war was unpopular. Opposition to the war, however, was uncoordinated, because Czech political leaders were unable to agree on a program. In December 1914 Tomáš Garrigue Masaryk, a representative in the Vienna parliament, left Prague to organize activities that could not be developed at home because of the suspension of civil rights and political persecution. After staying some months in neutral countries, Masaryk moved to London. In 1915 he had been joined in Switzerland by his former student Edvard Beneš and by Josef Dürich, a member of the conservative Czech Agrarian Party. Masaryk at first had rather vague notions of the tasks ahead of him. But, after conferring with distinguished experts in central European affairs, he eventually came to advocate a program of political union of the Czechs and Slovaks. A young Slovak astronomer, Milan Rastislav Štefánik, offered his support. Masaryk established contacts with the Czechs and Slovaks living in Allied and neutral countries, especially the United States. In 1916 a Czechoslovak National Council was created under Masaryk's chairmanship. Its members were anxious to maintain contacts with the leaders at home in order to avoid disharmony, and an underground organization called the Maffia served as a liaison between them.

At home the military regime headed by Archduke Frederick curbed the press, forbade public meetings, and imprisoned those suspected of disloyalty. Among those arrested were the pro-Russian Young Czech leader Karel Kramář and the economist Alois Rašín. Dissatisfaction among the Czech soldiers on the Eastern front became more articulate in 1915, and whole units often went over to the Russian side.

Francis Joseph died in November 1916 and was succeeded by Charles I. The new emperor abolished the most irritating restrictions on the freedom of expression and granted amnesty to political prisoners. In the spring of 1917 he called the parliament to session. Charles's reforms, although in many respects gratifying, called for more intensive activities abroad in order to convince the Allied leaders that partial concessions to the Czechs were inadequate to the problems of postwar reconstruction. The position of the Slovaks was not getting better, and the Hungarian government showed no inclination to reorganize the kingdom in accordance with the principle of nationality. Two major events coincided with Charles's new course in home affairs and with his discreet exploration of the chances of a separate peace: the Russian Revolution (March 1917) and the United States declaration of war on Germany. In May 1917 Masaryk left London for Russia to speed up organization of a Czechoslovak army. While small units of volunteers had been formed in the Allied countries during the early part of the war, thousands of prisoners of war were now released from Russian camps and trained for service on the Allied side. A Czechoslovak brigade participated in the last Russian offensive and distinguished itself at Zborov in July 1917. From the United States came moral encouragement, but Pres. Woodrow Wilson's early statements pertaining to the peace aims were rather hazy. Several weeks after the United States declaration of war on Austria-Hungary, President Wilson promulgated his celebrated Fourteen Points (January 8, 1918), the tenth of which called for "the freest opportunity of the autonomous development" for the peoples of Austria-Hungary.

Moderate reform under Charles I

United States encouragement

After the Bolshevik Revolution, the National Council decided to transfer Czechoslovak troops from Russia to France. As other routes to the West were blocked, they started moving toward Vladivostok but got involved in struggles between the Bolsheviks and the Conservative forces for the control of the Siberian railroad. Their achievements, noticed favourably in the Allied press, gave the Czechoslovak cause wide publicity, and Masaryk left Russia for the United States, where, in May 1918, he gained solid support from Czech and Slovak organizations. A declaration favouring political union of the Czechs and Slovaks was issued at Pittsburgh, Pennsylvania, on May 31, 1918 (called the Pittsburgh Convention).

In 1918, dealings with the Allies progressed more successfully. Not only the Siberian campaigns but also increased activities at home were used to get the struggle for independence endorsed by the Allied governments. A demand for a sovereign state "within the historic frontiers of the Bohemian lands and of Slovakia" was made in Prague at the Epiphany Convention (January 6, 1918) and repeated later with more vigour. In May 1918 not only the Czechs but also the Slovaks made statements to which Masaryk and his collaborators could point when pressing for an official recognition. The anti-Austria resolution, adopted at the Congress of Oppressed Nationalities at Rome (April 1918), helped in disarming conservative circles in the Allied countries who opposed a total reorganization of the Danubian region. After several encouraging statements came the recognition by France of the Czechoslovak National Council as the supreme body controlling Czechoslovak national interests; the other Allies soon followed the French initiative. On September 28, 1918, Beneš signed a treaty whereby France agreed to support the Czechoslovak program in the postwar peace conference. To preclude a retreat from the earlier Allied declarations, the National Council constituted itself as a provisional government (October 14, 1918). On October 18 Masaryk and Beneš issued a declaration of independence simultaneously in Washington and Paris. Events were moving rapidly toward total collapse of the Habsburg monarchy. The last attempt to avert it, the manifesto issued by Emperor Charles on October 16, 1918, brought no positive results. After that, Vienna had no choice but to accept President Wilson's terms. The surrender note, signed by Count Gyula Andrássy, last foreign minister, was accepted as a sanction of the idea of independence. The Prague National Committee proclaimed a republic on October 28, and, two days later, the Slovak National Council at Turčiansky Svätý Martin acceded to the Prague proclamation.

The establishment of Czechoslovakia (1918–25). Despite all efforts to maintain contacts between the leaders abroad and those at home, the early years of the republic were hindered by differences of opinion and occasional frictions. Masaryk returned to Prague on December 21, 1918. Beneš stayed in Paris and was joined by Karel Kramář, prime minister since November 1918. The Slovak leader Štefánik decided to return home but died in an air accident in May 1919. Masaryk and Beneš conducted external relations, and the leaders of five major parties controlled home affairs.

Of the many tasks facing the new government, negotiations at the peace conference, though complicated by dissensions among the great powers, were the least onerous. The frontiers separating Bohemia and Moravia from Germany and Austria were approved, with minor rectifications, in favour of the republic. The Slovak boundary was also felt to be satisfactory. The dispute over the Duchy of Teschen strained the relations with Poland; the partition of the duchy in 1920 was opposed by powerful Polish groups, and the Polish senate did not ratify the treaty. The northeastern counties of prewar Hungary (Carpathian Ruthenia) were attached to the new state. The area was inhabited by Slavic peoples, the majority of whom were keenly aware of their kinship with the Ukrainians.

Boundary settlements after war

Consolidation of internal affairs proceeded slowly. The winter of 1918–19 was critical. The most urgent task of the new government was to replace the wartime economy by a new system. The network of railroads and highways had to be adjusted to the new shape of the republic, stretching from the Cheb (Eger) region in western Bohemia to the Carpathians in the east. The first minister of finance, Alois Rašín, saved the Czechoslovak currency from catastrophic inflation, and his death in February 1923, after being shot by a young revolutionary, was a shock to the new republic.

In the chaotic conditions prevailing in central Europe after the armistice, a parliamentary election appeared to be impossible. The Czech and Slovak leaders agreed on the composition of the National Assembly. The Assembly's main function was the drafting of a constitution.

The new, democratic constitution was adopted on February 29, 1920, and was modelled largely after that of the French Third Republic. Supreme power was vested in a bicameral National Assembly. The Chamber of Deputies and the Senate had the right to elect, in a joint session, the president of the republic for a term of seven years. The Cabinet was made responsible to the Assembly. Fundamental rights of the citizens, irrespective of ethnic origin, religion, and social status, were defined generously.

Opposition from Sudeten Germans

Large segments of the population gave wholehearted support to the republic; the most resolute opposition, however, came from an ethnic minority that soon came to be known as the Sudeten Germans. The age-old antagonism between Germans and Slavs, accentuated during the war, prevented cooperation during the opening stages of the republic. The Germans issued protests against the constitution but participated, nevertheless, in parliamentary and other elections. In 1925 two German parties—the Agrarian and Christian Socialist—joined the government majority, thus breaking the deadlock. Disagreement with the trend toward centralism was the main source of dissatisfaction among the Slovak Populists, a clerical party headed by Msgr. Andrej Hlinka. Calls for Slovak autonomy were counterbalanced by other parties seeking closer contacts with the corresponding Czech groups; the most significant contribution to that effort was made by the Agrarians under Milan Hodža and by the Social Democrats under Ivan Dérer. The strongest single party in the opening period, the Social Democracy, was split in 1920 by internal struggles; in 1921 its left wing constituted itself as the Czechoslovak section of the Comintern. After the separation of the Communists, the Social Democracy yielded primacy to the Agrarians. The Republicans, as the peasant party was called officially, became the backbone of government coalitions until the disruption of the republic; from its ranks came Antonín Švehla (prime minister 1921–29) and his successors.

Political consolidation (1925–35). Foreign relations were largely determined by wartime agreements. Czechoslovakia adhered loyally to the League of Nations. Treaties with Yugoslavia and Romania gave rise to the Little Entente. France was the only major power that concluded an alliance with Czechoslovakia (January 1924). Relations with Italy, originally friendly, deteriorated after Benito Mussolini's advent to power in 1922. Czech anticlerical feeling precluded negotiation of a concordat with the papacy until 1928, when an agreement was worked out providing for settlement of the most serious disputes between church and state. But it was Germany that most strongly influenced the course of Czechoslovak foreign affairs. The relations between the two neighbours, correct but cool, improved slightly in 1925 after the Locarno Pact, a series of agreements among the powers of western Europe to guarantee peace. In the milder climate of the late 1920s, a third party, the Social Democrats, joined the German activists. Attempts to change the attitude of the Slovak Populists met with partial success. Reorganization of public administration in 1927, while marking a retreat from rigid centralism, did not go far enough to meet demands for Slovak autonomy. Monsignor Hlinka and his chief collaborator, Msgr. Josef Tiso, tenaciously pursued the program of decentralization and only at short intervals supported the Prague government.

Impact of Great Depression

When the impact of the Great Depression reached Czechoslovakia, soon after 1930, the highly industrialized German-speaking districts were hit more severely than the predominantly agricultural lowlands. The ground was thus prepared for the rise of militant nationalism. Parties supported by middle-class German voters and persisting in opposition to Prague gained in popularity. Soon after Hitler's rise to power, in 1933, efforts were made to organize noncommitted voters, especially hard-pressed businessmen. On October 1, 1933, Konrad Henlein, a gymnastics teacher, launched his Sudeten German Patriotic Front. Professing loyalty to the democratic system, he asked for recognition of the German minority as an autonomous body. In 1935 Henlein changed the name of his movement to the Sudeten German Party so as to enable its active participation in the parliamentary election (May 1935). The party captured nearly two-thirds of the Sudeten German vote and became a political force second only to the Czech Agrarians.

Moving toward the abyss (1935–38). A stable interlude of little more than two years followed the landslide victory of the Sudeten Germans. In December 1935 Masaryk retired from the presidency, and Beneš was elected his successor by an overwhelming majority, including Hlinka's party. A treaty with the Soviet Union in 1935 enhanced the sense of national security. The program of the Communist Party was determined not only by this treaty but also by the general reorientation of the Comintern, which now urged cooperation with anti-Fascist forces in popular fronts. The Czechoslovak Communists did not, however, seek Cabinet posts. The erection of an elaborate system of fortifications along the German frontier modelled on France's Maginot Line was commonly interpreted as an unwritten pledge of the French army to aid Czechoslovakia in the event of an unprovoked attack. Their capture would have given (and did give) the Germans the key to the French defense system. In February 1937 Prime Minister Milan Hodža made significant progress toward gaining the cooperation of those segments of the German population that were attached to the principles of democracy. The hope that Czechoslovakia would be able to withstand pressure from the Third Reich seemed, for a while, to be justified.

Hitler and eastward expansion of Germany

But, soon after the death of Masaryk, in September 1937, Hitler embarked on his program of eastward expansion. As early as November 1937, he informed his military chiefs of his intention to move against Austria and Czechoslovakia. After the annexation of Austria in March 1938, the Czechoslovak crisis became acute.

The Czechoslovak leaders divided their energies. Prime Minister Hodža devoted all his talents to a search for a compromise that would satisfy the Sudeten Germans and held long conferences with Henlein's lieutenants. President Beneš, assisted by his foreign minister, Kamil Krofta, maintained contacts with foreign powers. Henlein played his hand so skillfully that the influential circles, especially in London, believed that he was a free agent and not Hitler's stooge. The advocates of "appeasement," then rapidly gaining ground in Great Britain and in France, failed to realize that the Sudeten German negotiators had no intention of compromise and acted on instructions from Berlin. The main task of Henlein's party was to give Hitler a better chance to dislocate the republic without recourse to war. To invalidate critical comments from London and Paris, Beneš consented late in July to the mission of Lord Runciman, whose avowed purpose was to observe and report on conditions within the country.

The crisis culminated in September 1938. Armed with information supplied by Lord Runciman, British Prime Minister Neville Chamberlain visited Hitler at Berchtesgaden and Godesberg. Chamberlain assured Hitler that the German objectives could be achieved without fighting. The French consented to Chamberlain's policy, thus abandoning their former commitments. The Soviet Union was under no treaty obligation to assist Czechoslovakia, since the treaty of 1935 was to be operative only if the French would honour their pledges. Thus, the stage was set for a meeting between Hitler, Mussolini, Chamberlain, and Edouard Daladier, at Munich on September 29, 1938. They agreed on a document enjoining the Prague government to cede to the Third Reich all districts of Bohemia and Moravia with 50 percent or more of Germans; October 10 was set as the deadline for the transfer of these territories. Although presented as a measure to make Czechoslovakia more homogeneous and viable, the pact and its ruthless implementation sealed the fate of the country.

From Munich to the disruption of the republic (1938–39). Beneš resigned the presidency rather than agree to the German annexation. After several weeks he left Prague, first for London and then for Chicago. The lead-

ers who took over had to face mounting difficulties. The annexations completed according to the Munich timetable were not Czechoslovakia's only territorial losses. Poland obtained the Duchy of Teschen as a reward for its menacing attitude during the Munich crisis. By the Vienna Award (November 2, 1938) Hungary was granted large portions of Slovak and Ruthenian territories. By all these amputations Czechoslovakia lost about one-third of its population. Communications were totally disrupted, and the country was rendered defenseless.

The chances of recuperation were greatly reduced by the rapid growth of centrifugal tendencies. The Slovak Populists, headed since Hlinka's death by Monsignor Tiso, presented Prague with urgent demands for autonomy, which the government accepted. A similar request came from Carpathian Ruthenia. A cumbersome system composed of three autonomous units (the Czech lands, Slovakia, and Ruthenia) united by allegiance to the Prague government was introduced late in the fall. On November 30, 1938, Emil Hácha was elected president; an Agrarian leader, Rudolf Beran, formed the federal Cabinet. Under German pressure the complicated party system was changed drastically. The right and centre parties in the Czech lands formed the Party of National Unity; the Socialists organized the Party of Labour. In Slovakia the Populists absorbed all other political groups. Despite all efforts of the loyal elements, stabilization of political and economic life made little progress. Moreover, the public knew little of the confidential negotiations being conducted in Vienna and Berlin by Tiso's aides, who went along with Hitler's preparation for the final takeover. In the early months of 1939, Tiso's group prepared the ground for the secession of Slovakia, and, on March 14, 1939, the Slovak National Assembly voted for independence. On the same day, Hácha and his foreign minister, František Chvalkovský, journeyed to Berlin, where they were informed of Hitler's decision to annex Bohemia and Moravia; on the following day, Czechoslovakia was proclaimed a protectorate of the Third Reich.

The struggles at home and abroad (1939–45). The basic laws regulating the status of Bohemia and Moravia were drafted hastily, and many loopholes were left in them to facilitate German intervention. Hitler installed a Reich Protector in Prague as his personal representative. He was assisted not by Konrad Henlein but by Karl H. Frank, a Sudeten German notorious for his brutality. The Cabinet under President Hácha operated with limited rights and powers. Matters that in any way touched on the Third Reich, especially the war effort, were excluded from Czech jurisdiction and handled either by the Reich Protector or by direct order from Berlin. For some two years the protectorate kept the semblance of an autonomous body. But in September 1941 Reinhard Heydrich replaced Konstantin von Neurath as Reich Protector and inaugurated a reign of terror. After Heydrich's assassination (May 1942) the Germans virtually took over the country. Hácha stayed on as president, but the Cabinet was reconstructed in such a manner that it served only as a screen behind which the Germans carried out retaliatory measures and exploited the country's economy for their own purposes. Mass executions, consignment of Czech patriots to concentration camps, and recruitment of young people for work in Germany or behind the front continued until the collapse of the Nazi regime.

Reich Protector in Prague

Several months after the proclamation of the protectorate, Beneš moved from Chicago to London to resume his political activities. His position was originally rather awkward, as neither French nor British statesmen were keen on dealing with him. But, after the fall of France, in spring 1940, the British prime minister, Winston Churchill, granted Beneš recognition; a provisional government, with Msgr. Jan Šrámek as prime minister, began to function in London. In July 1941 Great Britain and the Soviet Union granted Beneš and his government-in-exile full recognition. Beneš maintained underground contacts with Prague and, until Heydrich's assassination, exercised indirect influence on the protectorate government. But Beneš' main occupation was with diplomacy. He devoted a good deal of energy to get the Munich Agreement denounced as invalid. While London and Washington were reluctant to make statements that might prejudice the outcome of the future peace conference, Moscow did not hesitate to condemn the past and open bright prospects for cooperation in the war and in the postwar reconstruction. In the desire to reap benefits of a timely decision for cooperation, Beneš visited Moscow in December 1943 and signed a treaty of alliance for 20 years, the terms of which far exceeded the pact of 1935. Not only the treaty but also conversations with Klement Gottwald, the leader of the Czechoslovak Communists, from then on determined both the policy of the exiles and the underground movement in the protectorate and in Slovakia. The Communist groups, passive before the Nazi invasion of the Soviet Union, now became intensely active and gradually seized the leadership from other clandestine organizations. They played a significant part in the Slovak uprising (late summer 1944) and in the activities of Czech patriots (spring 1945). It was of decisive importance that the Red Army penetrated deep into the territory of the republic several months earlier than the Western Allies were able to cross the traditional borderline between Germany and Bohemia. In March 1945 Beneš and other political figures journeyed from London to Moscow to make a final accord with Stalin and Gottwald. A program of postwar reconstruction was worked out under decisive Communist influence and a Cabinet was formed with Zdeněk Fierlinger as prime minister. The new government, set up at Košice in Slovakia on April 3, 1945, exercised jurisdiction in the eastern portion of the republic; in Moravia and Bohemia, fighting continued until early May. Underground activities, guided by the Czech National Committee, were intensified. On May 5 the people of Prague launched an uprising against the German troops concentrated in central Bohemia and fought them bravely for four days. Their appeals for Allied help were largely ignored. The U.S. general George S. Patton, though sympathetic, did not move from Plzeň, complying with instructions from Gen. Dwight D. Eisenhower. On May 9 the forces of the Soviet general Ivan Konev entered Prague. Thus, the Soviet Union secured both a military and political victory.

Communist role in underground

CZECHOSLOVAKIA SINCE 1945

The uneasy interim (1945–48). The new regime at Košice did nothing in foreign or in home affairs that would conflict with the 1943 treaty of alliance with the Soviet Union. Immediately after the end of fighting, a purge of those suspected, rightly or wrongly, of collaboration with the Nazis was carried out with utmost severity. The main target was the Sudeten Germans, whose transfer to Germany was sanctioned at the Potsdam Conference on August 2, 1945. A small fraction were allowed to stay in the republic. The transfer of the Magyars was discontinued after an agreement with Hungary. Since the Carpathian Ruthenia was ceded to the Soviet Ukraine, the republic became ethnically more homogeneous than it had been prior to 1938. The Slovak and Czech parties accused of either subservience to Hitler or of irresolute policy before Munich were not allowed to resume activities. The strongest among the former was the Slovak Populists, among the latter the Agrarians. Five major parties agreed on basic principles and established the National Front; they were the Communists (Czech and Slovak), Czech National Socialists, Czech People's Party (Catholic), Social Democrats, and Slovak Democrats. In the early period the government ruled by decrees, the most important of which called for the nationalization of key industries and banks. "National committees" were established on the village, town, district, and provincial levels; and on October 28, 1945, a provisional National Assembly was formed and unanimously endorsed a program of gradual transition from capitalism to Socialism.

Both in the local committees and on the higher levels, the Communists recognized other parties as equal partners. But by skillful manoeuvring the Communists es-

tablished themselves as the dominant element not only in the political assemblies but also in a large number of other organizations including the trade unions. The press, though ostensibly free, operated under invisible restrictions, and, on a number of vital problems, no open exchange of view was tolerated. In foreign affairs, national interests were subordinated to those of Moscow.

Communist election victory

The election of the Constituent Assembly in May 1946 revealed the proportional strength of the Communist and non-Communist movements. The Communists, stronger in Czech lands than in Slovakia, polled approximately 38 percent of the vote, making them the strongest party in its country, and, in accordance with parliamentary practices, Klement Gottwald became the prime minister. Seats in the Cabinet were distributed among the members of the National Front. Jan Masaryk, the son of T.G. Masaryk, retained the portfolio for foreign affairs that he had held in London; Gen. Ludvík Svoboda, whose troops had fought on the side of the Red Army, continued to serve as minister of defense. But not more than one year was allowed to pass in continuous efforts to keep the National Front in a reasonable balance. In June 1947 the United States invited Czechoslovakia to participate in the European Recovery Program (Marshall Plan). The Prague government accepted the offer in the belief that the Soviet Union would adopt a friendly attitude. But several days later Stalin warned the Czechoslovak delegation that any participation in the Marshall Plan would be interpreted in Moscow as a hostile act. Thereupon, the government reversed its decision. Seemingly, nothing hindered continued cooperation within the National Front. But the appearances were deceptive. Few people were aware of President Beneš' deteriorating health, an important factor in view of his key position.

With the term of the Constituent Assembly due to expire late in the spring of 1948, all parties opened campaigning for the coming election. Several ugly incidents in the Czech lands and in Slovakia caused uneasiness in democratic ranks, but the leaders of the parties were convinced that they had been motivated only by the Communist determination to increase the number of seats. A decision not to wait for the outcome of the election was made by the Communist leaders in utmost secrecy sometime before the end of 1947, and a timetable was set up to synchronize Communist policy with the tactics of other parties. The crisis came to a head in February 1948, precipitated by the decision of the minister of the interior, Václav Nosek, to replace non-Communists in the security forces by party members. Attempts of other leaders to change the Communist policy came to nothing. On February 20, 1948, the ministers of three parties—National Socialist, People's, and Slovak Democratic—resigned in protest and in the belief that the Social Democrats would eventually back them and that the president would take a firm stand against the Communists' tactics. But the Social Democratic leadership, weakened by discord, did not take the expected action. The Communists, advised in their policy by Valerian Zorin, who had gone to Prague ostensibly to supervise shipments of wheat from Russia, mobilized the worker's militia in Prague and other centres. Minister of Defense Ludvík Svoboda expressed his sympathy with the Communist aims. Action committees were organized in the factories, workshops, and offices to protect the Socialist cause. Messages from all parts of the country poured to the Hradčany Castle urging Beneš to drop the resigning ministers. He yielded to the pressure, and, on February 25, 1948, approved a list of Cabinet members submitted to him by Gottwald. In the new Cabinet the Communists held one-half of the portfolios; the other half was made up of members of other parties, which by now had been thoroughly purged. Jan Masaryk remained in office, but on March 10 he was found dead in the courtyard of his residence at Černín Palace. It was strongly suspected that he had been murdered, though his death was made to look like suicide.

The People's Democracy (1948–60). The new regime was ruled under the aegis of the Communist-dominated Renewed National Front (RNF). The Social Democrats, headed by Zdeněk Fierlinger, merged with the Communists in June 1948. The National Socialists and the People's Party continued in existence but with drastically reduced membership. The party of Slovak Regeneration replaced the Democrats. Several nonpolitical organizations were also admitted to the RNF. For several months the action committees were busy purging the civil service, army, mass organizations, and cultural institutions. Most non-Communist leaders fled to Paris, London, or the United States. On May 9, 1948, a new constitution was promulgated codifying the principles of a "people's democracy." Beneš disagreed with some articles and resigned on June 7; he died on September 3, 1948.

Resignation of Beneš

On June 14 the National Assembly elected Gottwald as president. Antonín Zápotocký, since 1945 the leading figure in the trade unions, headed the Cabinet. Rudolf Slánský continued as the secretary general of the Communist Party. The new leadership pushed ahead the February program of socialization. In October 1948 the first five-year plan set the lines for a thorough transformation of national economy, stressing rapid development of heavy industry above the production of consumer goods. Two subsequent plans (1954–58 and 1959–63) were drawn along similar lines, increasing the targets for metallurgy and heavy machinery. A law creating uniform agricultural cooperatives was passed in February 1949, but its application was hindered by farmers' opposition. Czechoslovakia joined the Council for Mutual Economic Assistance (Comecon), set up in 1949 as a counterpart to the European Recovery Program (Marshall Plan).

Gottwald and his aides loyally followed Stalin's line and were determined to crush any kind of opposition. Peasant and urban middle-class reluctance to accept a new social and economic order was suppressed by stern measures. Potential centres of discontent were destroyed by mass arrests and spectacular trials before "people's courts." The religious bodies, especially the Roman Catholic Church, were among the principal targets. In October 1949 a state office for church affairs was set up. In exchange for nationalization of church properties, the state accepted obligation to support the clergy from public funds. The main cause of conflict was the oath of allegiance. While the non-Catholic churches eventually complied with the request, the Catholic hierarchy persisted in its opposition. Several trials were held in 1950 and 1951 with the bishops, superiors of religious orders, as well as members of the lower clergy, as defendants. The monasteries were closed and the monks and nuns either imprisoned or forced to seek other occupations. Between August and October 1950, most of the bishops were moved from their residences; some of them were sentenced to prison terms. In March 1951, five prelates were finally induced to take the oath of allegiance. Other vacancies were filled from the ranks of "patriotic priests."

Political opposition was thrown into confusion by the escape to foreign lands of almost all pre-February leaders and by the elevation of pliable figures to mostly imaginary posts. But the omnipresent security forces sniffed out underground activities and attacked them mercilessly. In late spring 1950, Milada Horáková, a prominent figure in the feminist movement was brought, along with several others, before the people's court; one of them, Záviš Kalandra, an ardent Socialist, was accused of Trotskyism. The Communists soon discovered that their enemy was not alone to be sought in bourgeois or clerical centres. After Tito's feud with Moscow and the execution of László Rajk in Hungary, the high party echelons were put under surveillance. The demotion, in March 1950, of Vladimír Clementis from the foreign ministry was followed by an action against his two friends—Laco Novomeský, a poet of distinction, and Gustáv Husák, since 1946 a chairman of the Board of Trustees for Slovakia. The purge of the party proceeded with increasing vigour. In December 1951 Rudolf Slánský fell from grace; he and ten of his associates, including Clementis, were sentenced to death in November 1952. In April 1954 Gustáv Husák was sentenced to life imprisonment.

Party purges

The death of Gottwald (March 1953) brought no

change in the party's policy. The new president, Antonín Zápotocký, was personally more popular but showed no intention to inaugurate a milder course. The head of the Cabinet, Viliam Široký, a Slovak, was a staunch Stalinist, as was the acting secretary, Antonín Novotný. After a reorganization of the party's upper echelons, Novotný became its first secretary. Purges and public trials continued in 1954. The unveiling (May 1, 1955) of a huge Stalin monument on heights overlooking Prague showed that the Czechoslovak Communist Party was persisting in the policy that came soon to be known as the "cult of personality." The 20th Congress of the Communist Party of the Soviet Union and the policies of Stalin's successor, Nikita Khrushchev, were merely recorded in the Czechoslovak press. The conservative tendencies were strengthened by the election of Novotný to the presidency after Zápotocký's death in November 1957.

With closed ranks and concentration of power in the political bureau, the party carried out unpopular measures that were resented not only by whatever was left of the middle class but also by the working people. A monetary reform in May 1953 wiped out savings and caused disastrous inflation. The main industrial centres of Plzeň, Kladno, and Ostrava seethed with discontent, but security forces were strong enough to keep the situation under control. The collectivization of land was resumed in 1956 with greater energy, and by 1960 about 90 percent of cultivable land had been converted into either agricultural cooperatives or state farms. At about the same time the last vestiges of private business and of liberal professions disappeared. The educational system was tied to the party and adjusted to the Soviet pattern. In the creative arts so-called Socialist Realism was proclaimed the official creed. In July 1960 the constitution of May 1948 was replaced by a new charter, which gave the country the title of Czechoslovak Socialist Republic (CSSR). Like its model, the Soviet constitution of 1936, the Czechoslovak charter subordinated the state to the party and its complex machinery.

Reform endeavours (1960–67). Outbreaks of dissatisfaction in Poland and a Hungarian rising in 1956 played into the hands of the old-line Czechoslovak leaders. Neither Khrushchev nor his successors, Leonid Brezhnev and Aleksey Kosygin, insisted on replacement of the Stalinists by people of their own choice. Opposition was born in the party ranks and manifested itself in such a manner that Novotný's clique hesitated to adopt the measures to which Gottwald had resorted in 1950. The national economy did not develop as successfully as the official statements prognosticated. The third five-year plan (1961–65) had to be scrapped, and for several years production was regulated by short-term plans. In January 1965 the Central Committee adopted a more flexible system of planning and relaxed the bureaucratic control over management. The project was debated by the 13th Party Congress (May 31–June 4, 1966) and was put into effect on January 1, 1967. Although reluctant to admit that grave errors had been committed, the political bureau sanctioned revision of the purge trials from Gottwald's era. A commission was set up and made public its findings in August 1963. Among those fully rehabilitated was Vladimír Clementis. Rudolf Slánský and some of his associates were rehabilitated legally, but their expulsion from the party was confirmed. In 1965 a slight improvement occurred in church–state relations. Archbishop Josef Beran, promoted to cardinal, was allowed to leave for Rome, where he died in 1969. Bishop František Tomášek was appointed apostolic administrator of the archdiocese of Prague. Other bishops were allowed to return to their sees, but no general agreement with the Vatican was reached. Novotný sacrificed some of the most unpopular figures to relieve the tension. In September 1963 a moderate Slovak, Jozef Lenárt, replaced Viliam Široký as prime minister.

Improvement of church–state relations

Systematic efforts to bring intellectual life into harmony with party policy did not meet with full success despite heavy concentration on that wide field. Coercive measures alternated with attempts to win good-will by material advantages, but scholarly and creative activities could not be so easily curbed as the press or education in government-controlled schools. The universities and other institutions of that rank lost their traditional autonomy in 1950. Both the teaching staff and student body were purged in Gottwald's era; new appointments and admissions were rigidly supervised. In 1952 the Czechoslovak Academy of Sciences (with a branch at Bratislava) was founded as a centre for research. Its institutes launched learned journals and series of publications and provided opportunities for a large number of scholars, many of whom earned international recognition. Writers, artists, composers, actors, and film producers were organized into unions, over which the party exercised control through pliable officers. Despite all precautions, news of the 20th Congress of the Communist Party of the Soviet Union and the gradual de-Stalinization reached Czechoslovakia and had lively echoes among intellectuals, young and old. The second congress of the Union of Czechoslovak Writers (1956) was pervaded by an invigorating atmosphere. Novotný's regime imposed restrictions of all kinds but did not achieve more than a temporary restitution of the former uniformity. Concessions, which the party leadership could not avoid in other spheres of national life, could not be put off indefinitely when intellectual activities came up for discussion. Liberalization was not sanctioned by official declaration, but it was pioneered by talented members of that generation that had joined the party prior to 1948, endorsed wholeheartedly the Communist program, but was increasingly resentful of stifling regimentation. The fourth congress of the Union of Czechoslovak Writers (June 1967) took a resolute stand against censorship and the party's efforts to interfere with creative processes. Retaliatory measures, aimed especially at the Union's weekly *Literární noviny*, supplied further evidence of Novotný's determination to remain at the helm. Although governed by the same basic laws as the Czech lands, Slovakia did not follow their implementation in exactly the same path. The constitution of 1960 attempted to restrict the power of the Board of Trustees and other autonomous bodies. The policy of centralization encountered more and more articulate opposition. In the end, Slovak resentment of Novotný's regime precipitated a crisis.

The "Prague Spring." The critical phase of the reform movement began early in January and closed late in August 1968—the so-called "Prague Spring." The movement became nearly synonymous with the name of Alexander Dubček, since April 1963 the first secretary of the Slovak Communist Party. Dubček was elected to succeed Novotný as the first secretary of the Czechoslovak Communist Party on January 5, 1968. Other changes both in the party and state administrations followed in rapid succession. Oldřich Černík became the prime minister and Josef Smrkovský, chairman of the National Assembly. Two vice premiers rose to prominence: Ota Šik, an advocate of decentralization in industrial enterprises; and Gustáv Husák, an ardent champion of Slovak nationalism. Under ever increasing pressure, Novotný retired from the presidency and was succeeded by General Svoboda. Early in April the party announced a program outlining reforms in political, economic, and cultural matters. Despite changes in the top level, the conservative wing of the party still held a strong position in the Central Committee and other party organs. In order to secure more support, Dubček called an extraordinary congress for September. Special committees were formed to prepare material for that session.

The rise of Dubček

Dubček's activities were supported by the press and other media. Students and writers enthusiastically joined the efforts to give "human face" to Socialism. Not only the non-Communist members of the National Front but also nonparty intellectuals felt encouraged to take active part in public affairs. The movement was not uniform. While the Slovaks put the need of federalization above other concerns, the Czechs insisted on a complete rehabilitation of persons afflicted by "violation of socialist legality" and on better guarantees of freedom of expression. The new leadership adopted a sympathetic attitude to the fresh currents in public life without, however,

giving up the idea that the Communist Party must retain its dominant position.

It was apparent, however, that the conservative wing had not lost backing among the rank and file and in the cumbersome apparatus created by Gottwald and his successors. From the countries of the Warsaw Pact, warnings were heard more often than applause. Walter Ulbricht and Władysław Gomułka were less cautious than Soviet statesmen in assessing the new trends. Several meetings and exchanges of visits occurred in the spring. Early in May, Dubček, Svoboda and other prominent officials journeyed to Moscow to dispel suspicions; they gave consent to army manoeuvres of the Warsaw Pact to be held in Czechoslovakia in summer. A group of senior officers arrived in Prague late in May to complete preparations for the war games.

As spring advanced, it became clear that the official decisions and the reforming zeal of the radicals were no longer in harmony. The Central Committee endorsed the April Action Program in principle but modified some of its concrete stipulations. These retrogressive steps were balanced by other measures, such as relaxation of censorship and a law providing for the rehabilitation of victims of political persecutions. Both groups reacted promptly. Diehards headed by Vasil Bilak widened their contacts with corresponding groups in the Warsaw Pact countries. The champions of liberalization on June 27 published a manifesto, the "Two Thousand Words," signed by more than 70 prominent persons. Although the party Presidium promptly disavowed the manifesto, the Soviet, Polish, and East German presses intensified their hostile campaign. The critical period opened.

Army manoeuvres began on June 20 and were officially closed on July 2. The movements of the Soviet units, ostensibly heading eastward, caused a good deal of anxiety. Symptoms of uneasiness multiplied rapidly. The Soviet, Bulgarian, East German, Hungarian, and Polish leaders met in Warsaw in mid-July to review the situation. Dubček replied promptly to their letter and reaffirmed his loyalty to the Warsaw Pact. But the exchange of messages did not dispel suspicions. On July 29 the Soviet Politburo met with the Czechoslovak delegation at a border railroad station, Čierna nad Tisou. The agreement reached there was vague, allowing for different interpretation of its articles. On August 3 the same men who attended the Warsaw conference met with Dubček and his group at Bratislava; the statement issued there contributed nothing new to the controversy. Seemingly, the differences were settled to mutual satisfaction. But events occurring on both sides increased the tension. In Czechoslovakia the liberalization continued; problems handled in other Communist countries behind closed doors were ventilated publicly and with unprecedented audacity. To fasten the ties with countries not present in either Warsaw or Bratislava, Dubček invited Tito and President Ceausescu to Prague; Tito arrived on August 11, Ceausescu, on August 15. They were received with great enthusiasm. Ulbricht requested a meeting with Dubček; Marshal Andrey Grechko went to East Germany for a conference with the defense minister, General Hoffman. Criticism of "a noticeable intensification in Prague of subversive activities by anti-Socialist forces," published in *Pravda* on August 18, portended nothing good.

Foreign invasion

In the night of August 20–21, the troops of the five Warsaw Pact countries, namely, the Soviet Union, Poland, Hungary, Bulgaria, and the German Democratic Republic, crossed the Czechoslovak frontier and took over the country with amazing efficiency. Late on August 21 Dubček and some of his aides were taken away to an unknown destination. The arrival of foreign troops aroused spontaneous reaction among the population. Attempts to form a pliable Cabinet came to nothing. On August 23, President Svoboda, accompanied by six notables, left for Moscow. Dubček and his aides were transferred there and participated in negotiations. An agreement on "normalization of the situation" was eventually reached; the Czechoslovak delegation returned to Prague on August 27. During the absence of the top leaders, the population had been guided by radio and television; other sources of information were underground newspapers, leaflets, and posters urging people to avoid clashes with foreign troops. The National Assembly held sessions during the critical period. On August 22 an extraordinary congress of the party, which had been scheduled for September 9, was hastily convened; 1,192 out of 1,543 elected delegates met in a Prague suburb and debated reports of the committees. In Slovakia, Vasil Bilak was, on August 28, replaced by Husák as the first secretary. In a public statement Husák expressed doubts concerning the validity of the extraordinary congress; his rise to prominence began.

Normalization and consolidation (1968–71). Forces of the smaller countries soon withdrew, but reduced Soviet units remained quartered in strategic positions. A treaty on their stationing was signed on October 16, 1968, and then endorsed by the National Assembly. Soviet leaders increased efforts to obtain an additional sanction of their military intervention. The August events were to be harmonized with the principles defined in a *Pravda* article (September 26) and called henceforth the Brezhnev Doctrine. Except for some replacements, Dubček's team retained the leading positions both in the party and state. Attempts were made to put the less objectionable points of the April program into effect and thus restore people's confidence. On October 28, the 50th anniversary of the republic, the law concerning federalization of the country was approved; it was signed in Bratislava on October 30. It brought about noticeable changes in every sector of public life. New posts both in the party and in the government were created and filled to give the Slovaks parity with the Czechs. Dubček's repeated efforts to recapture people's sympathy missed the mark. The population was guided more by instinct than by official proclamations. Its hero became Jan Palach, a student who on January 16, 1969, set himself on fire in protest against the erosion of freedom in Czechoslovakia. But his heroic deed did not bring about a significant change; the bulk of the population was rapidly losing interest in public affairs; material problems and everyday worries took precedence over big issues affecting the country.

The Soviet determination to obtain an additional approval of the August events was the strongest factor in shaping Czechoslovak policy in 1969 and thereafter. Since Dubček's statements were found evasive, he was gradually moved from the top level to total insignificance. In April 1969 the foremost champion of the "new realism," G. Husák, took over as the first secretary and found enough support in Moscow and at home to silence the opposition, which was entrenched in cultural organizations and educational institutions. Černík's fall in January 1970 marked the end of another chapter. Czechoslovak ties with the Soviet Union were tightened in May 1970 by a 20 years' treaty of friendship, cooperation, and mutual aid. The most important development at home was a thorough purge of the party, which went on until the end of 1970.

Demotion of Dubček

The final step toward "normalization" was made by Husák at the 24th Soviet Party Congress in Moscow (March–April 1971). Husák admitted that Socialism in Czechoslovakia had been in danger and that the armed intervention by members of the Warsaw Pact "saved his country from civil war and counterrevolution." The Czechoslovak Party Congress (late May 1971) reaffirmed Husák's declaration and annulled the extraordinary assembly of August 1968. The congress approved the list of new members of both the Presidium and the Central Committee of the party; Husák was acclaimed as secretary general. The large scale purge was sanctioned; the principle of permanent purge opened the door to continuous search for unreliable elements. In the presence of delegates from the five Warsaw Pact countries, Czechoslovakia was fully rehabilitated and readmitted into the Moscow-led bloc as a loyal and dependable member.

BIBLIOGRAPHY. Standard works in Czech, Slovak, English, German, and French are listed in PAUL L. HORECKY (ed.), *East Central Europe: A Guide to Basic Publications* (1969). The following books supply information on various

aspects of national history of the Czechs and Slovaks. Although outdated in some points, R.W. SETON-WATSON, *History of the Czechs and Slovaks* (1943), can be used for general orientation. S.H. THOMSON, *Czechoslovakia in European History*, 2nd ed. (1953), is arranged topically rather than chronologically and treats the basic issues. The perennial problem of Czech–German coexistence is presented by E. WISKEMANN in *Czechs and Germans*, 2nd ed. (1967). JOZEF LETTRICH, *History of Modern Slovakia* (1955), has as its main theme the Czech–Slovak relationship in the era of democracy. R.R. BETTS devoted most of his *Essays in Czech History* (1969) to the Hussite movement. Two books by F. DVORNIK, *The Making of Central and Eastern Europe* (1949) and *Byzantine Missions Among the Slavs* (1970), although dealing with general issues of medieval Slavic history, contain information on early Bohemia and its rise as an independent country. R.E. WELTSCH, *Archbishop John of Jenstein (1348–1400)* (1968), is a biography of the third archbishop of Prague and sheds light on state–church relations around 1390. From among several books of M. SPINKA, *John Hus: A Biography* (1968), should be mentioned as the best summary of the reformer's life and teachings. H. KAMINSKY, *A History of the Hussite Revolution* (1967); and F.G. HEYMANN, *John Žižka and the Hussite Revolution* (1955), have much in common; but while the former analyzes competently the ideas of the Hussite leaders in relation to political events, the latter describes the defense of the Hussite inheritance by Žižka and his followers. F.G. HEYMANN, *George of Bohemia, King of Heretics* (1965); and O. ODLOZILIK, *The Hussite King* (1965), cover much the same ground, the era of consolidation under the moderate Hussite leadership. There is no outline in English of Czech religious history in the 15th and 16th centuries. Two solidly documented monographs contribute to the knowledge of the "third party," the Unity: PETER BROCK writes on *Political and Social Doctrines of the Unity of Czech Brethren in the Fifteenth and Early Sixteenth Centuries* (1957); and J.K. ZEMAN traces the contacts of two groups, *The Anabaptists and the Czech Brethren in Moravia, 1526–1628* (1969). R.J. KERNER, *Bohemia in the Eighteenth Century* (1932, reprinted 1969), has a general survey of post-White Mountain developments, but its main emphasis is on social and economic conditions under the "Enlightened despots." The book of essays, edited by PETER BROCK and H. GORDON SKILLING, *The Czech Renascence of the Nineteenth Century* (1970), examines significant episodes in the development of modern Czech nationalism. Three of its contributors published monographs on specific features of Czech national life in the 19th century: S. PECH on *The Czech Revolution of 1848* (1969); J.F. ZACEK on *Palacký: The Historian As Scholar and Nationalist* (1970); and S.B. KIMBALL on *Czech Nationalism: A Study of the National Theater Movement, 1845–1883* (1964). The book of essays *Czechoslovakia: Twenty Years of Independence*, ed. by R.J. KERNER, appeared in 1940 as a record of the achievements of the Czechoslovak democracy. D.H. PERMAN, *The Shaping of the Czechoslovak State* (1962), is an excellent survey of peace negotiations pertaining to the creation of the republic. R. LUZA, *The Transfer of the Sudeten Germans: A Study of Czech-German Relations, 1933–1962* (1964), connects topically with Wiskemann's book and carries the story into the post-Munich era. V. MASTNY, *The Czechs Under Nazi Rule* (1971), is a succinct account of the early period of the German control of Bohemia and Moravia. Many books appeared in recent years on the Communist movement in Czechoslovakia, some of them hastily compiled and quickly antiquated. Three serious studies illustrate the earlier period: P.E. ZINNER, *Communist Strategy and Tactics in Czechoslovakia, 1918–1948* (1963); J. KORBEL, *The Communist Subversion of Czechoslovakia, 1938–1948* (1959); EDWARD TABORSKY, *Communism in Czechoslovakia, 1948–60* (1961). Three well-informed authors introduce the rather complicated story of liberalization and its gradual suppression after August 1968: T. SZULC, *Czechoslovakia Since World War II* (1971); W. SHAWCROSS, *Dubček* (1970); P. WINDSOR and A. ROBERTS, *Czechoslovakia, 1968* (1969). The debate of the merits of the movement symbolized by Alexander Dubček goes on.

(O.O.)

Bohemond I, Prince of Antioch

Norman adventurer, crusading leader, and prince of Antioch, Bohemond was one of the most important figures in the First Crusade. History records him as a handsome man, a warrior of genius, and a gifted diplomat.

Son of Robert Guiscard (the Wily) and his first wife, Alberada, Bohemond was christened Marc but nicknamed after a legendary giant named Bohemond. The nickname proved well taken because physically Bohemond was the ideally tall and strong knight—in the words of a contemporary, "a wonderful spectacle." His boyhood home was in southern Italy, where his Norman father Robert had gone as a mercenary and had risen to the rank of duke of Apulia and Calabria. Here Bohemond became involved in his father's wars and learned his trade as a fighter and leader. This early training must be inferred, however, as Bohemond's childhood is poorly recorded. His date of birth is placed between 1050 and 1058. In 1079 he was in command of a unit of his father's army. Meanwhile, his stepmother, Sigelgaita, bore his father's heir-to-be, Roger Borsa; thus, Bohemond no doubt felt early in life that he would have no patrimony because of his half-brother and so would have to seek lands and fortune in the weakened condition of the Greek empire.

In 1081 Bohemond, in command of his father's army, captured Avlona, a town south of Durazzo; but in this same year Alexius I Comnenus became ruler of the Byzantine Empire and challenged the Norman threat. For over three decades Alexius and Bohemond were rivals. In the opening struggle, 1081–85, Bohemond and his father came close to dismembering the Greek empire in the West. The Norman army won a few brilliant victories, but the persistence of Alexius drove Bohemond from Larissa in Thessaly in 1083, and the death of his father in 1085 left Bohemond without a patrimony and with little hope of success against Byzantium. In the next four years Roger Borsa allowed Bohemond to gain a foothold in Bari (about 65 miles northwest of Brindisi), where he awaited another chance to move decisively against Alexius.

Rivalry with Alexius I Comnenus

The chance came when Pope Urban II launched the First Crusade in November 1095 by offering rewards in both this world and the next for those who wrested the Holy Sepulchre from the Saracens. When the word reached Bohemond, he set off for the East. Bohemond and his small band of Normans crossed the Greek lands in the winter of 1096–97 with few incidents; and on passing through Constantinople, he made friendly, though cautious, terms with the emperor Alexius. The latter managed to extract oaths from most of the leaders, including Bohemond, and helped them cross the Bosporus, speeding them with promises of aid if they would return to the sovereignty of the Emperor the Byzantine lands recaptured from the Muslims. In the ensuing campaigns against the Turks, Bohemond distinguished himself at Nicaea, Dorylaeum, and Antioch, which was besieged from October 1097 until June 3, 1098. The city of Antioch fell to the crusaders through his cunning and his negotiations with a traitor, Fīrūz. After a brief, unsuccessful countersiege by the Turks, during which Bohemond more or less assumed command, the crusaders dawdled away the summer and fall.

When the crusading army marched southward to Jerusalem in January 1099, Bohemond was left the de facto possessor of Antioch, although his claim was not openly supported for fear of violating the oath of Alexius. The Norman leader did not participate in the capture of Jerusalem but did, for the sake of appearances, journey later to the Holy Sepulchre. With the departure of many crusaders for their homelands, Bohemond was left with his city, hemmed in by Alexius to the north and the Muslim world to the east. Following sorties against Aleppo, Bohemond made the mistake of moving against the emir of Sebastea (Sivas), to the north of Antioch. The "little God of the Christians," as he was termed by the Turks, fell into an ambush and was captured and held in chains for months.

Bohemond assumes the rule of Antioch

Released in 1103, he returned to Antioch and its problems. In 1105 Bohemond was in Bari to enlist reinforcements for his struggle with the Greeks. In September 1105 he went to Rome to interview the Pope and then journeyed, early in 1106, through France. There, babies were named for him, crowds heard him denounce the perfidious Alexius, and shrines received sacred relics from his hands. In the spring of 1106 Bohemond married Constance, the daughter of Philip I of France, in a brilliant ceremony at Chartres.

Bohemond, who 30 years before had been a landless young man, now stood at the pinnacle of his career. By September 1107 he was ready to launch his crusade against the Greeks and within a month had landed a large army at Avlona. In the months that followed, Durazzo held firm against the Normans, and Bohemond met with misfortune in Albania. In this impasse Alexius, anxious to end the war, offered Bohemond Antioch and other Greek cities in return for vassalage. In accepting these terms, Bohemond suffered humiliation even though he retained control of Antioch. He had used the crusade against Alexius, however, to further his ambition for an empire that stretched from Apulia to Antioch and had thereby cheapened the crusading idea.

The years following this peace of discord are poorly recorded. Constance bore Bohemond two sons, one of whom later became Prince of Antioch. Bohemond probably sought to raise another army, but he died on March 7, 1111. His combat with the Greeks was ended, and his rival Alexius followed him in death in 1118. Nicknamed for a giant, Bohemond had fought against gigantic odds and at death bequeathed to his heirs one of the important crusader states, the Principality of Antioch.

BIBLIOGRAPHY. R.B. YEWDALE, *Bohemond I: Prince of Antioch* (1917), offers the best biography of Bohemond; R.L. NICHOLSON, *Tancred: A Study of His Career and Work in Their Relation to the First Crusade and the Establishment of the Latin States in Syria and Palestine* (1940), supplements the work of Yewdale; K.M. SETTON and M.W. BALDWIN (eds.), *A History of the Crusades*, vol. 1 (1955), is the best general work on Bohemond.

(J.H.Hi./L.L.H.)

Böhme, Jakob

Jakob Böhme, a German philosophical mystic who summed up various Renaissance and Reformation mystical views in his writings, became one of the most influential leaders in later intellectual movements. Through his religious and mystical writings he exerted a profound influence on movements such as idealism and romanticism, and on the views of such modern thinkers as Nikolay Berdyayev and Paul Tillich. His influence was also important in the development and expansion of the Quakers, German Pietism, and theosophy.

By courtesy of the Staatsbibliothek Berlin



Böhme, woodcut by Hugo Bürkner (1818–97).

Early mystical experiences

Böhme was born in 1575, at the end of the Reformation period, in the Upper Lusatian (German) village of Altseidenberg. After receiving a rudimentary education, he went, in 1594 or 1595, to nearby Görlitz, a town where controversies over Reformation issues seethed. Here crypto-Calvinists (Lutherans charged with maintaining Calvinist views), Anabaptists (radical Protestants), Schwenkfeldians (followers of the Reformer Schwenkfeld), Paracelsian physicians (followers of the occultic physician Paracelus), and humanists vied with orthodox Lutherans. Martin Möller, the Lutheran pastor of Görlitz, was "awakening" many in the conventicles that he had established. At these religious meetings traditional materials, both early Christian and medieval German, were studied by a varied group of persons.

In 1600, newly married and just established with a shoemaker's bench of his own, Böhme, probably stimulated by Möller, had a religious experience within the period of a quarter hour wherein he gained an empirical and speculative insight that helped him to resolve the tensions of his age. The strain between medieval and Renaissance cosmologies (dealing with the order of the universe), the perennial problem of evil, the collapse of feudal hierarchies, and the political and religious bifurcation of the time found resolution in Böhme's rediscovery, as he said, of the dialectical principle that "in Yes and No all things consist." Basically Lutheran ("we shall fear and love God," as Luther's Small Catechism states), this principle became with Böhme a *Realdialektik* ("real dialectic"), a wide-ranging polarization of empirical or natural reality.

The Aurora his first writing

Germinating for several years, the insight led him to commit his thoughts to paper, at first for his own use. The manuscript was entitled *The Aurora* (1612) and was written in stages. Called by Böhme a "childlike beginning" it was a conglomeration of theology, philosophy, and what then passed for astrology, all bound together by a common devotional theme. Circulating among Böhme's friends, a copy of *The Aurora* fell into the hands of Gregory Richter, successor to Martin Möller as pastor, who condemned the shoemaker's pretensions to theology. Richter brought the matter up with the Görlitz town council, which forbade further writing on Böhme's part.

A period of silence ensued during which Böhme's ideas matured and his outer affairs prospered. He read the "high masters" as well as other unnamed books that were lent to him by the circle of neighbours and friends who were awed by the book-writing intellectual cobbler. These friends—some physicians, and others of the nobility—introduced Böhme to speculative alchemy, especially to the writings of the Swiss physician Paracelsus (*c.* 1490–1541) that were then quite popular. The alchemical and mystical views of Paracelsus further inspired Böhme's interest in nature mysticism and gave him the terminology which, in a partly integrated way, dominated his next period.

Although he never worked in a laboratory himself, Böhme did use its alchemical terms to describe both his nature mysticism and his subjective experiences, which he sought to integrate into a common framework. During this period Böhme wrote at least six tracts that were circulated guardedly among his friends, creating an influential and respected reputation for him. This second period of writing activity began in 1619, the year when the Thirty Years' War (1618–48) was beginning to gain momentum; in fact, Böhme himself was in Prague (Czechoslovakia) when the Winter King, Frederick V of the Rhine Palatinate, entered. The various strident controversies of the age forced Böhme into a period of religious apologetics wherein he had to protest his orthodoxy against accusations, more implied than actual, of Calvinism (Reformed views), chiliasm (belief in the 1,000-year reign of God's people at the end of history), and rabid sectarianism. Reconstructing his theological views, he wrote a series of devotional tracts dealing with penitence, resignation, regeneration—traditional themes of German mysticism. In 1622 his friends had several of these devotional tracts printed in Görlitz under the title *The Way to Christ*, a small work joining nature mysticism with devotional fervour. Publication of this tract brought about the intense displeasure of Richter, who incited the populace against Böhme.

In 1623, the year of his maturity, he wrote two major works: *The Great Mystery* and *On the Election of Grace*. The former explained the creation of the universe as told in Genesis in terms of the Paracelsian three principles (including the mystical elements "salt," "sulfur,"

and "mercury"), thus joining Renaissance nature mysticism with biblical religion. The latter, more philosophical, gave exposition in terms of dialectical insight to the problem of freedom that Calvinist predestination (the view that man's destiny is foreknown by God) was then making acute. This theme later was taken up by the idealist philosopher Friedrich Schelling and by a German theologian, Franz von Baader, whose commentary *On the Election of Grace* is still held in high regard by scholars.

Later career

Böhme continued his writing at hectic pace, perhaps freed from business obligations by financial help from his friends. Between 1619, when he defiantly renewed his writing, and 1624, when he died, he produced at least 30 works. His defiance of the town council of Görlitz brought him further difficulty and he was banished, being cited to the Elector's court in Dresden, where, to all appearances, he found vindication because he returned to his home. Although vindicated by the theologians who had examined his views, he was not free from the rancorous moods of his neighbours who were instigated in their attacks by Richter. Esteemed by his friends among the nobility, physicians, and intellectuals, he fled to one of the neighbouring castles where he clearly was the central figure in some kind of secretive group. There he fell sick and, sensing that his end was near, he was taken back home to Görlitz where, attended by his wife and sons, he began to weaken. He was examined by ecclesiastical authorities, found orthodox enough to be given the sacrament, and on November 21, 1624, in a mood of charismatic expectancy, he died.

BIBLIOGRAPHY. The 17th-century translation of Böhme into English by J. ELLISTONE and J. SPARROW, *The Works of Jacob Behmen*, 4 vol., reprinted in the 20th century, is considered to have more grace and elegance than modern translations of some tracts by J.R. EARLE and J.J. STOUDT. The 1730 German edition of the collected works, edited by J.W. UEBERFELD and reprinted in facsimile during the 1950s by W.E. PEUCKERT, remains the best text, although in modernized German; valuable biographical matter may be found in vol. 10. The most serviceable English biography utilizing modern materials is J.J. STOUDT, *Jacob Boehme: His Life and Thought* (1968), based on the standard German biographical works by W.E. PEUCKERT, *Das Leben Jakob Böhmes* (1924); and R. JECHT, *Jakob Böhme, Gedenkgabe der Stadt Görlitz* (1924). The best modern interpretation is A. KOYRE, *La Philosophie de Jacob Boehme* (1929); the most profound is E. BENZ, *Der vollkommene Mensch nach Jakob Böhme* (1937).

(J.J.S.)

Bohr, Niels

One of the foremost scientists of the 20th century, Niels Henrik David Bohr was the first to apply the quantum theory (that the energy of a system at the atomic or molecular scale can take on only certain discrete values) to the problem of atomic structure. He was a guiding spirit and major contributor to the development of quantum physics for almost 50 years.

Early life in Denmark and Britain

Bohr was born in Copenhagen on October 7, 1885, the son of Christian Bohr, professor of physiology at the University of Copenhagen, and Ellen Adler Bohr, whose family was prominent in Danish banking. His scientific interests and abilities were evident early and were encouraged and fostered at home. As a 22-year-old student at the University of Copenhagen he won a gold medal from the Royal Danish Academy of Sciences and Letters for his precise experimental study of the surface tension of water. In 1911 Bohr received his doctorate for a thesis on the electron theory of metals. He then went to England intending to continue this work with J.J. Thomson at Cambridge, but soon moved to Manchester to join Ernest Rutherford's group studying the structure of the atom.

His first great work began with a study of the theoretical implications of the nuclear model of the atom proposed by Rutherford. Bohr was among the first to see the importance of the atomic number (a number assigned to each element and equal to the electric charge on the nuclei of its atoms) in characterizing the physical and chemical properties of the elements. In 1913 he combined the concept of the nuclear atom with the quantum theory of Max Planck and Albert Einstein in a theory that accounted quantitatively for the lines observed in the spectrum of light emitted by atomic hydrogen. To do so, he departed radically from classical physics, postulating that any atom could exist only in a discrete set of stable energy states, that radiation would be emitted only when the atom made a transition between two of these states, and that the frequency of this radiation depended only on the difference in energy between initial and final states. This meant that the atom could neither absorb nor emit energy continuously but only in finite steps or quantum jumps.



Bohr.
By courtesy of the Nobelstiftelsen ,Stockholm

Bohr was married to Margrethe Nørlund in 1912, at just the time when he was working out his bold new ideas. He stayed on at Manchester, with some interruptions, until 1916 when he returned to Copenhagen as professor. In 1920 he became director of the newly created Institute for Theoretical Physics, a position he held until his death. Bohr's work on atomic structure was recognized by the Nobel Prize for Physics in 1922.

Bohr elaborated his ideas in the early 1920s to explain the atomic basis of the periodic table of the elements and many atomic properties. More than any of his contemporaries, however, Bohr stressed the tentative and symbolic nature of the atomic models that were used, since he was convinced that an even more radical change in physics was still to come.

By the time that change arrived, with the quantum mechanics of 1925–26 by Werner Heisenberg, Erwin Schrödinger, Max Born, Wolfgang Pauli, Paul Dirac, and others, Bohr's institute had become the world capital for atomic physics. A steady stream of people flowed into Copenhagen to work on quantum theory with Bohr, work that was carried on by intense discussions as well as lengthy calculations. Out of these discussions came a physical interpretation of the new mathematical description of nature, linking it with the procedures and results of the experimental physicists. Bohr expressed the characteristic feature of quantum physics in his principle of complementarity:

Principle of complementarity

> Evidence obtained under different experimental conditions cannot be comprehended within a single picture, but must be regarded as complementary in the sense that only the totality of the phenomena exhausts the possible information about the objects.

This Copenhagen interpretation of the meaning of quantum physics, with its altered view of the meaning of physical explanation, was generally accepted by scientists, but Albert Einstein, among others, held out against it until the end of his life. He and Bohr often discussed the fundamental questions of physics over the years, and Einstein's challenging objections played an important part in the evolution of Bohr's ideas.

During the 1930s Bohr continued to work on the epistemological problems raised by the quantum theory, and also contributed to the new field of nuclear physics. His concept of the atomic nucleus, which he likened to a liquid droplet, played an essential part in the understanding of nuclear fission (the splitting of a heavy nucleus into two parts, almost equal in mass, with the release of a tremendous amount of energy).

Role under German occupation

When Denmark was overrun and occupied by the Germans in 1940, Bohr did what he could to maintain the work of his institute and to preserve the integrity of Danish culture against Nazi influences. In 1943, under threat of immediate arrest (Bohr's mother had been Jewish and his anti-Nazi views were no secret), he escaped to Sweden with his family by fishing boat with the help of the Danish resistance movement. He and his son, Aage, also a theoretical physicist and Bohr's eventual successor as director of the institute, took part in the projects for making a nuclear fission bomb, first in England and then in Los Alamos, New Mexico. Bohr's concern about the terrifying prospects for humanity produced by such atomic weapons was evident as early as 1944, when he tried to persuade Churchill and Roosevelt of the need for international cooperation in dealing with these problems. Although this did not succeed, Bohr continued to argue for rational, peaceful policies, advocating an "open world" in a public letter to the United Nations in 1950. Bohr was convinced that free exchange of people and ideas among countries was necessary to achieve control of nuclear weapons. He led in promoting such international efforts as the First International Conference on the Peaceful Uses of Atomic Energy, Geneva (1955), and in creating the European nuclear physics laboratory (CERN). Among his many honours, Bohr received the first U.S. Atoms for Peace Award in 1957.

In his later years Bohr tried to point out ways in which the idea of complementarity could throw light on many aspects of human life and thought from biology and anthropology to the age-old problems arising from the fact that man both acts and is the spectator of his actions. Bohr had a major influence on several generations of physicists, deepening their approach to their science and to their lives. He himself was always ready to learn from his younger collaborators. He drew strength from his close family ties with his sons, his wife, and his younger brother, Harald, a famous mathematician in his own right and Bohr's closest friend.

Niels Bohr died peacefully on November 18, 1962. Bohr's works include *The Theory of Spectra and Atomic Constitution* (1922), *Atomic Theory and the Description of Nature* (1934), *Atomic Physics and Human Knowledge* (1958), and *Essays, 1958–1962, on Atomic Physics and Human Knowledge* (1963). The last two include, respectively, his wonderful accounts of his relationships with Einstein and Rutherford.

BIBLIOGRAPHY. There is no definitive biography of Bohr; however, STEFAN ROZENTAL (ed.), *Niels Bohr: His Life and Work As Seen by His Friends and Colleagues* (1967), contains much biographical material. See also the article by LEON ROSENFELD in the *Dictionary of Scientific Biography*, vol. 2, pp. 239–254 (1970), an authoritative account of Bohr's life and work with bibliographical references; RUTH MOORE, *Niels Bohr: The Man, His Science and the World They Changed* (1966), a popular biography; WERNER HEISENBERG, *Der Teil und das Ganze* (1969; Eng. trans., *Physics and Beyond: Encounters and Conversations*, 1971), contains much material on Bohr and his work, by one of his leading collaborators; MAX JAMMER, *The Conceptual Development of Quantum Mechanics* (1966), a history of quantum physics with technical details on Bohr's work through 1927; and BRYCE S. DEWITT and R. NEILL GRAHAM, "Resource Letter IQM-1 on the Interpretation of Quantum Mechanics," *Am. J. Phys.*, 39:724–738 (1971), an annotated bibliography placing some of Bohr's contributions in perspective.

(M.J.K.)

Bolingbroke, Henry St. John, 1st Viscount

A prominent English politician in the reign of Queen Anne (1702–14) and, later, a major political propagandist in opposition to the Whig Party then led by Sir Robert Walpole, Bolingbroke was also a historian of some talent and a would-be philosopher in touch with many of the trends of the European Enlightenment. Intelligent and widely read, he was also noted for his handsome appearance, graceful manners, and brilliant conversation. Clear and forceful in speech and in print and imperious in temperament, he captivated some of the finest minds of his age. On the other hand, he was a notorious libertine and a poor manager of men who tended to lose his nerve in a crisis, and his unscrupulous ambition betrayed him into serious political errors and gained him a reputation for treachery.

By courtesy of the National Portrait Gallery, London



Bolingbroke, oil painting attributed to Alexis-Simon Belle, probably 1712. In the National Portrait Gallery, London.

The eldest son of Sir Henry St. John (afterward 1st Viscount St. John; pronounced sin' jun), he was born on September 16, 1678, probably in Wiltshire; he was possibly educated at a Dissenting academy rather than at Eton and Oxford, as has been claimed. In 1698–99 he travelled in Europe, and in 1700 he married Frances Winchcombe. In 1701, he entered Parliament, where he soon won a reputation by his superb oratory and his support of partisan Tory measures, including attacks on the previous Whig ministry and on the Protestant Dissenters, who were the Whigs' staunchest allies. His conduct soon brought him to the notice of the government, and, after he was made secretary at war (1704), he was converted, temporarily, to the moderate policies of Robert Harley, one of Queen Anne's principal ministers. For four years, he worked hard to provide the Duke of Marlborough with troops and equipment for the War of the Spanish Succession against France and then resigned with Harley (February 1708) when they failed to prevent the Whigs from dictating government policy. Failing to gain a seat in the 1708–10 Parliament, he urged Harley to ally with the Tory Party as the best means to defeat the Whigs.

Alliance with Harley

In 1710 St. John became northern secretary of state in Harley's new ministry, but he soon emerged as an opponent of Harley's moderation and a rival to his authority. His efforts to control the government's policies and to supplant Harley (after 1711 the earl of Oxford) were largely unsuccessful. Oxford had initiated secret peace negotiations with France, but, even after he had learned of these and had forced his way into the discussions, St. John (after 1712 Viscount Bolingbroke) was not able to dictate the terms that were finally settled at the Treaty of Utrecht (1713). In Parliament, Bolingbroke was no more successful in leading a Tory rebellion against Oxford. He won over some Tories by such partisan measures as the Schism Act (1714), which aimed at depriving the Dissenters of their schools, but he failed to persuade the majority to support his leadership and was unable to give the Tories a clear lead on the disputed succession to Queen

Anne. Oxford was eventually dismissed on July 27, 1714, but the Queen's death, on August 1, ruined Bolingbroke's hopes of replacing him.

Exile in France

Dismissed from office by George I and fearing impeachment because of his role in the peace negotiations with France and his intrigues with the Jacobites (the supporters of James III, the Old Pretender), Bolingbroke fled to France (March 1715) and became the Old Pretender's secretary of state in July. This enabled the British government to pass an act of attainder against him by which his property and civil liberties were taken away. As a result, Bolingbroke's political future depended upon a successful Jacobite rebellion. Despite Bolingbroke's hard work, the attempted Jacobite rising in 1715 was a dismal failure. Amidst bitter recriminations, Bolingbroke was dismissed by the Old Pretender and at once sought to ingratiate himself with the Whig government in England. In 1717 he wrote a *Letter to Sir William Wyndham* (not published until 1753) to defend his actions since 1710 and to persuade the Tories to abandon the Jacobite cause. Not surprisingly, he found it difficult to persuade men to forget his recent conduct.

Forced to remain in exile, Bolingbroke sought other outlets for his talents. Mixing with aristocrats and scholars, including Voltaire, he embarked on biblical, historical, and philosophical studies and wrote several works, including *Reflections upon Exile* and *Reflections Concerning Innate Moral Principles*. Shortly after the death of his first wife, he married a French widow, the Marquise de Villette (1719).

After years of petitioning the British government and of trying to assist it with his limited influence at the French court, Bolingbroke was pardoned in 1723. He did not, however, resettle in England until 1725, when an act allowed him to buy a small estate at Dawley, near London; his attainder was never fully reversed, and he was unable to regain his peerage or reclaim his seat in the Lords. He imputed this exclusion from parliamentary life to the animosity of Sir Robert Walpole. Though his own frustrated ambition clearly motivated his long campaign against Walpole's political ascendancy, he was also concerned by the way Walpole appeared to monopolize power by the excessive use of bribery and corruption. While charges of such behaviour were exaggerated, there was enough truth in them to build up a formidable opposition to Walpole. At the centre of a literary circle that included Jonathan Swift, Alexander Pope, and John Gay, Bolingbroke waged an influential propaganda campaign. His major contributions to *The Craftsman*, an opposition journal, were the "Remarks on the History of England" (1730–31) and "A Dissertation upon Parties" (1733–34), both of which sought to end the old Whig–Tory disputes and to weld the disparate elements of the opposition to Walpole into a new Country Party, which would protect the independence of Parliament against the encroachments of a corrupt government.

Despite occasional successes, Bolingbroke was unable to bring down Walpole or create a united opposition party. In 1735 he retreated to France, where he continued his studies in philosophy and history, lamenting his countrymen's lack of patriotism in the struggle against Walpole. After a short visit to England in 1738, his hopes revived of a new opposition party that was gathering at Leicester House around George II's son Frederick, prince of Wales. For this group, he wrote *The Idea of a Patriot King*. It was his most famous work, but it offered no real solution to the problems of defeating Walpole or of creating a "patriot" party. In any event, Prince Frederick did not live to become king, and Walpole's final defeat, in 1742, was not engineered by Bolingbroke.

In his last years, Bolingbroke lacked any real political influence, though he still made vain efforts to create a patriot ministry. He was further embittered by his discovery, in 1744, that Alexander Pope had secretly printed 1,500 copies of *The Idea of a Patriot King* for publication. When, in 1749, Bolingbroke published a corrected version of this work, he was bitterly attacked for taking the opportunity to reveal Pope's earlier breach of faith. Bolingbroke's failing health was further undermined by his distress at his wife's death (March 1750). He himself died at Battersea of a facial cancer on December 12, 1751. Though he died a neglected figure, the posthumous publication of his works in 1754 stirred considerable controversy. His unorthodox religious views were at last made public and were denounced on all sides. Modern scholars have paid much less attention to his philosophical works, but he is widely regarded as one of the best contemporary analysts of the politics of the Whig supremacy.

BIBLIOGRAPHY. H.T. DICKINSON, *Bolingbroke* (1970), now the standard life, based on extensive research and seeking to integrate and interpret Bolingbroke's political career and intellectual development; W.S. SICHEL, *Bolingbroke and His Times*, 2 vol. (1901–02, reprinted 1968), an old-fashioned biography, which is not very good on interpretation but which quotes much useful source material; G. HOLMES, *British Politics in the Age of Anne* (1967), a brilliant analysis, based on exhaustive research; I. KRAMNICK, *Bolingbroke and His Circle* (1968), a stimulating study of the political ideology of the age of Walpole.

(H.T.D.)

Bolívar, Simón

Simón Bolívar, soldier-statesman to whom six Latin-American republics owe their freedom from Spanish rule, is regarded by many as the greatest genius the Hispanic-American world has produced. A man of international renown in his own day, his reputation has steadily increased since his death. There are few figures in European history and none in the history of the United States that display the rare combination of strength and weakness, character and temperament, prophetic vision and poetic power that distinguish Simón Bolívar. As a consequence, his life and his work have grown to mythical dimensions among the people of his continent.

By courtesy of the Library of Congress, Washington, D. C.



Bolívar, engraving by C.G. Childs (1793–1871).

Bolívar, the son of a Venezuelan aristocrat of Spanish descent, was born to wealth and position on July 24, 1783, in Caracas. His father died when the boy was three, and his mother died six years later. His uncle administered his inheritance and provided him with tutors. He was an unruly child, and only one of his teachers had a lasting influence on him. This was Simón Rodríguez, a disciple of Jean-Jacques Rousseau, who introduced Bolívar to the world of 18th-century liberal thought, which produced on him a deep and lasting effect.

Education in Europe; marriage

At the age of 16, Bolívar was sent to Europe to complete his education. For three years he lived in Spain and in 1801 married the daughter of a Spanish nobleman, with whom he returned to Caracas. The young bride died of yellow fever less than a year after her marriage. Bolívar

often pondered on the effect her tragic death may have had on him and believed that it was the reason for his taking up a political career while still a very young man. In 1804, when Napoleon was approaching the pinnacle of his career, Bolívar returned to Europe. In Paris he met his friend and tutor, and under Simón Rodríguez' renewed guidance steeped himself in the writings of such European rationalist thinkers as Locke, Hobbes, Buffon, d'Alembert, and Helvetius. In addition, he read Voltaire, Montesquieu, and Rousseau. The latter two had the deepest influence on his political thinking, but Voltaire coloured his philosophy of life. In Paris he also met the German scientist Alexander von Humboldt, who had just returned from his voyage through Hispanic America, and who told Bolívar that he believed the Spanish colonies were ripe for independence. This idea took root in Bolívar's imagination and on a trip to Rome, standing on the heights of the Monte Sacro, he made a vow to liberate his country.

One other experience enriched his intellect at this time: he watched the extraordinary performance that culminated in Napoleon's coronation in 1804 as emperor of the French. Bolívar's reaction to the coronation wavered between admiration of the accomplishments of a single man and revulsion at Napoleon's betrayal of the ideals of the Revolution. The desire for glory was one of the permanent traits in Bolívar's character, and there can be little doubt that it was stimulated by Napoleon. The example of Napoleon was, nevertheless, a warning that Bolívar heeded. In his later days he always insisted that the title of "liberator" was higher than any other and that he would not exchange it for that of king or emperor. In 1807 he returned to Venezuela by way of the United States, visiting the eastern cities.

Independence movement. The Latin-American independence movement was launched a year after Bolívar's return. It was touched off by Napoleon's invasion of the Iberian peninsula. The Spanish people resisted the intruder, and a junta assumed governmental powers during the absence of the legitimate king. Napoleon also failed completely in his attempt to gain the support of the Spanish colonies, who claimed the right to nominate their own officials. Following the example of the mother country, they wished to establish juntas to rule in the name of the deposed Spanish king. Many of the Spanish settlers, however, saw in these events an opportunity to sever their ties with Spain. Bolívar participated in many conspiratorial meetings, and when the governor of Venezuela, Vicente Emparán, reprimanded him, he answered boldly that he had declared war on Spain and could not withdraw. On April 19, 1810, Emparán was officially deprived of his powers and expelled from Venezuela, and a junta took over. To obtain help, Bolívar was sent on a mission to London, where he arrived in July. His assignment was to explain to England the plight of the revolutionary colony, to gain recognition for it, and to obtain arms and support. Although he failed in his negotiations on all counts, his English sojourn was in other respects a fruitful one. It gave him an opportunity to study the institutions of the United Kingdom, which remained for him models of political wisdom and stability. More important, he fostered the cause of the revolution by persuading the exiled Francisco de Miranda, who in 1806 had attempted to liberate Venezuela singlehandedly, to return to Caracas and to assume command of the independence movement.

The mission to London

Venezuela was in ferment. In March 1811 a national congress met in Caracas to draft a constitution. Bolívar, though not a delegate, threw himself into the debate that aroused the country. In the first public speech of his career he declared "Let us lay the cornerstone of American freedom without fear. To hesitate is to perish." After long deliberation the national assembly declared Venezuela's independence on July 5, 1811. Bolívar now entered the army of the young republic, whose commander in chief was Miranda. In the short time since their London meeting, however, the two men had drifted apart. Miranda called Bolívar a "dangerous youth," and Bolívar had misgivings about the ability of the aging general. When the Spaniards rallied to retain control of the colony, Bolívar was placed in charge of Puerto Cabello, a port vital to Venezuela. Treasonable action by one of Bolívar's officers opened the fortress to the Spanish forces, and Miranda, believing the loss of Puerto Cabello fatal to any future military action, entered into negotiations with the Spanish commander in chief. An armistice was signed (July 1812) which left the entire country to the mercy of Spain. Miranda was turned over to the Spaniards—some authorities say at Bolívar's instigation—and spent the rest of his life in Spanish dungeons.

Determined to continue the struggle, Bolívar obtained a passport to leave the country and went to Cartagena in New Granada (present-day Colombia). There he published the first of his great political statements: *El Manifiesto de Cartagena.*

Cartagena manifesto

> I am a son of unhappy Caracas, miraculously escaped from its political and material ruins, who . . . has come here to follow the banner of freedom.

His reference to miraculous escape refers to the catastrophic earthquake of March 26, 1812, which took a tremendous toll of lives in Caracas and other towns. He did not hesitate to point out the underlying causes for Venezuela's downfall: "Not the Spaniards, but our own disunity had led us back into slavery. The most important error committed by Venezuela upon entering the political theatre, was, without doubt, the fatal adoption of a system of tolerance, a weak and inefficient system, . . ." Bolívar had now emerged as the champion of strong government for the nascent republics of Hispanic America.

His plea was answered by the people of New Granada, and he was named commander of an expeditionary force whose task was to liberate Venezuela. In a sweeping hard-fought campaign he vanquished the Spaniards in six pitched battles and regained control of the capital. On August 6, 1813, he entered Caracas, was given the title of liberator, and assumed political dictatorship.

But the war of independence was just beginning. The majority of the people of Venezuela were hostile to the forces of independence and weary of the sacrifices imposed. A cruel, anarchic, murderous civil war broke out. Bolívar himself resorted to extreme measures, such as the "war to the Death" and the shooting of prisoners. But his severity failed in its object. In 1814 he was once more defeated by the Spanish, who had drawn the *llaneros,* the cowboys of the Orinoco Valley, into the war. The Spanish had converted the *llaneros,* led by José Tomás Boves, into an undisciplined but savagely effective cavalry that Bolívar was unable to repulse. Boves captured Caracas in 1814 and subjected the city to terrible atrocities. Thus ended the second Venezuelan republic. Bolívar narrowly escaped Miranda's fate. He managed to reach Cartagena, where he was commissioned to oust a separatist faction from Bogotá and succeeded in doing so. He then laid siege to the New Granadan port city of Cartagena, but failed to unite the revolutionary forces, and fled to Jamaica.

In exile he turned his energies toward gaining support from Great Britain, and in an effort to convince the British people of their stake in the freedom of the Spanish colonies, he wrote the greatest document of his career: *La Carta de Jamaica* ("The Letter from Jamaica"). "The bonds," wrote Bolívar, "that united us to Spain have been severed." He was not dismayed that the Spaniards had in certain instances won the upper hand. "A people that love freedom will in the end be free." Bolívar outlined a grandiose panorama from Chile and Argentina to Mexico. "We are," he said proudly, "a microcosm of the human race. We are a world apart, confined within two oceans, young in arts and sciences, but old as a human society. We are neither Indians nor Europeans, yet we are a part of each." What kind of government should emerge from the cataclysm of the independence movement? "The American states need the efforts of paternal governments to heal the wounds and scars made by despotism and war." He proposed constitutional republics, modelled on the government of Great Britain, with a hereditary upper house, an elected lower house,

"The Letter from Jamaica"

and a president chosen for life. The last provision, to which Bolívar clung throughout his career, constituted the most dubious feature of his political thinking.

In "The Letter from Jamaica" Bolívar showed himself as a great internationalist. Looking forward to the day when the representatives of all Hispanic-American nations would gather in a central location such as Panama, he wrote,

> How ineffable it would be if the Isthmus of Panama should become for America what the Straits of Corinth were for the Greeks. May God grant that we can some day enjoy the good fortune of opening a congress of representatives of the republics, kingdoms, and empires that would discuss peace and war with the rest of the nations of the world.

By 1815, however, Spain had sent to its seditious colonies the strongest expeditionary force that had ever crossed the Atlantic. Its commander was Pablo Morillo. Since neither Great Britain nor the United States would promise aid, Bolívar turned to Haiti, a small republic that had freed itself from French rule, where he was given a friendly reception, as well as money and weapons.

Liberation of New Granada. Three years of indecisive defeats and victories followed. In 1817 Bolívar decided to set up headquarters in the Orinoco region, which had not been devastated by war and from which the Spaniards could not easily oust him. He engaged the services of several thousand foreign soldiers and officers, mostly British and Irish, established his capital at Angostura (now Ciudad Bolívar), began to publish a newspaper, and established liaison with the revolutionary forces of the plains. Their leader, José Antonio Páez, recognized Bolívar's authority, and by 1819 the latter was firmly in command of the entire region. In the spring of that year he conceived his master plan of attacking the Spanish position on its western flank. He had previously centred his hopes on the liberation of Caracas, but he now concentrated on the bolder project of an attack on the Spanish viceroyalty of New Granada. In the plains of Casanare a force of patriots had withstood all Spanish attempts to destroy them. Their leader, Francisco de Paula Santander, made contact with Bolívar, who operated in the eastern part of the plains, and on June 12, 1819, the armies of Bolívar and Santander met and merged.

Páez and Santander

Bolívar's attack on New Granada will always be considered one of the most daring in military history. The route of the small army (about 2,500 men, including the British legion) led through the plains, but it was the rainy season and the rivers had become lakes. For seven days, said one of Bolívar's aides, they marched in water up to their waists. Ten navigable rivers were crossed, most of them in cowhide boats. But the journey through the plains seemed child's play in comparison with their ascent of the Andean mountains that stood between Bolívar and the city of Bogotá. Bolívar had chosen to cross the pass of Pisba, which the Spanish considered an inconceivable approach. An icy wind blew across the heights of the pass, and many of the scantily clad troops died of cold and exposure. But fatigue and loss were more than outweighed by the advantage gained in descending unopposed into New Granada. When the Spaniards recovered from their surprise, it was too late to throw Bolívar back. A series of skirmishes followed, culminating in the crucial Battle of Boyacá on August 7, 1819, when the bulk of the royalist army surrendered to Bolívar. Three days later he entered Bogotá. It was the turning point in the history of northern South America.

Indefatigably Bolívar set out to complete his task. He appointed Santander vice president in charge of the administration and in December 1819 made his appearance before the congress that had assembled in Angostura. Bolívar was made president and military dictator. As he said, the union of New Granada and Venezuela had been his goal since his earliest fighting days. He urged the legislators to proclaim the creation of a new state: the republic of Great Colombia, and three days later La República de Colombia was established. It was a federation and, since two of its three departments, Venezuela and Quito (Ecuador), were still under royalist control, it was only a paper achievement. Bolívar knew, however, that victory was finally within his reach. A revolution in Spain had forced the Spanish king to recognize the ideals of liberalism on the home front, and his action quite naturally discouraged the Spanish forces in South America. Bolívar persuaded Morillo to open armistice negotiations, and the two warriors met in a memorable encounter at Santa Ana, signing, in November 1820, a treaty that ended hostilities for a six-month period. When fighting was resumed, Bolívar found it easy, with his superior manpower, to defeat the Spanish forces in Venezuela. The Battle of Carabobo, June 1821, opened the gates of Caracas, and Bolívar's homeland was at last free. In the autumn of the same year a congress convened in Cúcuta to draft a constitution for Colombia. Its provisions disappointed Bolívar. Although he had been elected president, he thought the constitution too liberal in character to guarantee the survival of his creation. As long as more urgent assignments claimed his attention, however, he was willing to put up with its weak structure. Leaving the administration to Santander, he asked permission to continue his military campaign.

Creation of Great Colombia

At the end of a year, Ecuador was liberated. In this campaign Bolívar was assisted by the most brilliant of his officers, Antonio José de Sucre. While Bolívar engaged the Spaniards in the mountains that defended the northern access to Quito, capital of modern Ecuador, Sucre marched from the Pacific coast to the interior. At Pichincha on May 24, 1822, he won a victory that freed Ecuador from the Spanish yoke. On the following day the capital fell and Bolívar joined forces with Sucre on June 16.

It was in Quito that the Liberator met the great passion of his life, the vivacious and beautiful Manuela Sáenz. She was the ideal woman for a gallant soldier like Bolívar, the perfect mixture of Amazon and courtesan. An ardent revolutionary, she freely admitted her love for the Liberator and accompanied Bolívar from the battlefields to the presidential palace.

Liberation of Peru. The territory of Colombia had now been completely recovered from Spain and its new government recognized by the United States. Only Peru remained in the hands of the Spaniards. It was the Peruvian problem that brought Bolívar and the Argentinian revolutionary José de San Martín together. San Martín had done for the southern part of the continent what Bolívar had accomplished for the north. In addition, he had already entered Lima and proclaimed Peru's independence. But the Spanish forces retreated into the highlands, and San Martín, unable to follow them, decided to consult with Bolívar. On July 26, 1822, the two men met in the port city of Guayaquil, Ecuador, and their conference has been a source of controversy ever since. What took place has never become known, except in the form of either biased or fragmentary information. Apparently San Martín came to request military aid from Bolívar, and in addition he wanted to reach an understanding on problems of boundaries and the political future of Latin America. There was scant sympathy between the two. Bolívar, brilliant, ambitious, self-centred, was convinced that he was the "chosen son," singled out by providence to complete the independence of his people. San Martín was stoic, taciturn, self-effacing, and moderate. His failure to influence Bolívar was almost a foregone conclusion, and on his return from Guayaquil he resigned his office in Lima and went into exile. Whether he took this action to give Bolívar a free hand or out of a sense of personal frustration, is still not clear.

Meeting with José de San Martín

The avenue leading to Bolívar's ultimate ambition was now open. In September 1823 he arrived in Lima. The Spanish Army occupied the Sierra east of Lima and its position was considered unassailable. But for Bolívar, after the trials he had successfully passed, this was no deterrent. Men, horses, mules, and ammunition were systematically assembled to form an army, and in the summer of 1824 he ascended into the high mountain country. Once more, as chief of staff, Sucre was his able assistant. The first major battle, at Junín, was easily won by Bolívar, who then left the successful termination of

the campaign to Sucre. On December 9, 1824, the Spanish viceroy lost an important battle and surrendered with his entire army.

Bolivia. Bolívar was now president of Colombia and Peru. Only a small section of the continent—Upper Peru—was still defended by royalist forces. The liberation of this region fell to Sucre, and in April of 1825 he reported that the task had been terminated. The new nation chose to be called Bolivia after the name of the Liberator. For this child of his genius, Bolívar drafted a constitution that showed once more his authoritarian inclinations: a lifetime president, a legislative body without power, and a highly restricted suffrage. Bolívar was devoted to his own creation, but as an instrument of social reform it was a failure.

Treaties of alliance

Bolívar had now reached the high point of his career. His power extended from the Caribbean to the Argentine-Bolivian border. He had conquered severe illness, which during his sojourn in Peru had made him practically an invalid for months at a time. Another of his favourite projects, a league of Hispanic-American states, came to fruition in 1826. He had long advocated treaties of alliance between the American republics, whose weakness he correctly apprehended. By 1824 such treaties had been signed and ratified by the republics of Colombia, Peru, Mexico, Central America, and the united Provinces of Río de la Plata. In 1826 a general American congress convened in Panama. Compared with Bolívar's original proposals, it was a fragmentary affair, since only Colombia, Peru, Central America, and Mexico sent representatives. The delegation from the United States never reached Panama. The four nations who attended signed a treaty of alliance and invited all other nations to adhere to it. A common army and navy were planned, and a biannual assembly representing the federated states was projected. All controversies among the states were to be solved by arbitration. Despite its meagre results, the congress of Panama laid the cornerstone for future hemispheric solidarity and understanding. The Organization of American States and the United Nations can look to Bolívar as one of the first statesmen in the world sincerely interested in advocating and implementing international cooperation.

But Bolívar was aware that his plans for hemispheric organization had met with only limited acceptance. His contemporaries thought in terms of individual nation-states, Bolívar in continents. In the field of domestic policy he continued to be an authoritarian republican. Many of his followers offered him the crown, but the Liberator preferred his old title to that of "Simón the First." He thought of himself as a rallying point, and anticipated civil war as soon as his words should no longer be heeded. Such a prophecy, made in 1824, was fulfilled in 1826.

Civil war. Venezuela and New Granada began to chafe at the bonds of their union. The protagonists in each country, Páez in Venezuela and Santander in New Granada, opposed each other, and at length civil war broke out. Bolívar left Lima in haste, and most authorities agree that Peru was glad to see the end of his three-year reign and its liberation from Colombian influence. In Bogotá, Bolívar found Santander upholding the constitution of Cúcuta and arguing that Páez be punished as a rebel. But Bolívar was determined to preserve the unity of Colombia and was therefore willing to appease Páez, with whom he became reconciled early in 1827. Páez bowed to the supreme authority of the Liberator, and in turn Bolívar promised a new constitution that would do justice to Venezuela's desire for regional independence. He took over the presidency and called for a national convention that met in April 1828. Bolívar refused to influence the elections, with the result that the liberals under the leadership of Santander gained the majority. Bolívar had hoped that the constitution of Cúcuta would be revised and presidential authority strengthened, but the liberals blocked any such attempts. A stalemate developed. Arguing that the old constitution was no longer valid and that no new one had taken its place, Bolívar assumed dictatorial powers. He flattered himself when he said that "the whole nation recognizes my authority," but he soon learned the bitter truth. A group of liberal conspirators invaded the presidential palace on the night of September 25, and Bolívar was saved from the daggers of the assassins only by the quick-wittedness of Manuela Sáenz. But though this attempt on his life had failed, the storm signals increased. Bolívar's precarious health began to fail. Peru invaded Colombia with the intention of annexing Guayaquil. Once more Sucre saved Colombia and defeated the Peruvians at Tarqui. A few months later, one of Bolívar's most honoured generals, José María Córdoba, staged a revolt. It was crushed, but Bolívar was disheartened by the continued ingratitude of his former adherents. France, England, and the United States tried to intervene in the domestic affairs of the country. In the fall of 1829, Venezuela seceded from Colombia.

Assassination plot

Reluctantly, Bolívar realized that his very existence presented a danger to the internal and external peace of the nations that owed their independence to him, and on May 8, 1830, he left Bogotá planning to take refuge in Europe. Reaching the Atlantic coast, he learned that Sucre, whom he had trained as his successor, had been assassinated. Bolívar's grief was boundless. In vain a military uprising in Bogotá called him back. The projected trip to Europe was cancelled, and at the invitation of a Spanish admirer, Bolívar journeyed to his estate near Santa Marta. Ironically his life ended in the house of a Spaniard, where on December 17, 1830, he died of tuberculosis.

BIBLIOGRAPHY. GERHARD MASUR, *Simón Bolívar*, 2nd ed. (1969), recognized as the best biography in English; V.A. BELAUNDE, *Bolívar and the Political Thought of the Spanish American Revolution* (1938), a penetrating study of Bolívar's ideas; D.F. O'LEARY, *Bolívar and the War of Independence* (1970), an abridged translation of the recollections of one of Bolívar's closest political and military aides (ends at 1826); S. DE MADARIAGA, *Bolívar* (1951; Eng. trans., 1952), an unfavourable assessment of Bolívar by a great Spanish writer; V.W. VON HAGEN, *The Four Seasons of Manuela, a Biography: The Love Story of Manuela Sáenz and Simón Bolívar* (1952), the best biography of Manuela Sáenz; SIMON BOLIVAR, *Cartas del Libertador*, ed. by VICENTE LECUNA, 12 vol. (1929–59), the most important collection of source material for the personality of Bolívar—a selection from these letters may be found in SIMON BOLIVAR, *Selected Writings*, comp. by VICENTE LECUNA and H.A. BIERCK, 2 vol. (1951); D. BUSHNELL, *The Liberator: Simón Bolívar* (1970), a useful introduction to the man and his image; PAN AMERICAN UNION, *Bibliography of the Libertador, Simón Bolívar* (1930), incomplete but still useful; R.A. HUMPHREYS, *Latin American History: A Guide to the Literature in English* (1958), an indispensable tool for those who do not read Spanish.

(G.S.M.)